INHOUD

Literatuur

Adema, J.J., & van der Linden, W.J. (1989). Algorithms for computerized test construction of parallel tests using classical item parameters. *Journal of Educational Statistics, 15,* 129-145.

Aitchison, J., & Silvey, S.D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics, 29,* 813-828.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Andersen, E.B. (1970). Asymptotic properties of conditional maximum likelihood estimation. *Journal of the Royal Statistical Society, Series B, 32,* 283-301.

Andersen, E.B. (1973a). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123-140.

Andersen, E.B. (1973b). *Conditional inference and models for measuring.* (Unpublished Ph.D. Thesis). Copenhagen: Mentalhygiejnisk Forlag.

Andersen, E.B. (1973c). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology, 26,* 31-44.

Andersen, E.B., & Madsen, M. (1977). Estimating the parameters of the latent population distribution. *Psychometrika, 42,* 357-374.

Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42,* 69-81.

Andersen, E.B. (1980). *Discrete statistical models with social science applications.* Amsterdam: North Holland.

Andersen, E.B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 46,* 443-459.

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Andrich, D. (1978b). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement, 38,* 665-680.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In: R.L. Thorndike (red.). *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Armstrong, R.D., Jones, D.H., & Wu, I. (1992). An automated test development of parallel tests from a seed test. *Psychometrika, 57,* 271-288.

Bartko, J.J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports, 19,* 3-11.

Bartko, J.J., & Carpenter, W.T. (1976). On the methods and theory of reliability. *The Journal of Nervous and Mental Disease, 163,* 307-317.

Bejar, I.I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7,* 303-310.

Bentler, P. M. (1985). *Theory and implementation of EQS: A structural equations program*. Los Angeles: BMDP Statistical Software.

Berger, J.O. (1980). *Statistical decision theory: Foundations, concepts and methods.* New York: Springer.

Berk, R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56,* 137-172.

Beuk, C.H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement, 21,* 147-152.

Bezembinder, Thom. G. G. (1970). *Van rangorde naar continuum*. Deventer: Van Loghum Slaterus.

Birnbaum, A. (1968). Some latent trait models. In: F.M. Lord, & M.R. Novick. *Statistical theories of mental test scores* (pp. 397-424). Reading: Addison-Wesley.

Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge: The MIT Press.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Bock, R.D. (1976). Basic issues in the measurement of change. In: D.N.M. de Gruijter, & L.J.Th. van der Kamp (red.). *Advances in psychological and educational measurement* (pp. 75-96). London: Wiley.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM-algorithm. *Psychometrika, 46,* 443-459.

Bock, R.D., Gibbons, R.D., & Muraki, E. (1988). Full-information factor analysis. *Psychological Measurement, 13,* 261-280.

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks. *Journal of Educational Statistics, 15,* 129-145.

Bol, E., & Verhelst, N.D. (1985). Inhoudelijke en statistische analyse van een leertoets. *Tijdschrift voor Onderwijsresearch, 10,* 49-68.

Bollen, K.A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bosch, L. van den, Gillijns, P., Krom, R., & Moelands, F. (1991). *Handleiding schaal vorderingen in spellingvaardigheid 1.* Arnhem: Cito.

Bradley, T.B. (1983). Remediation of cognitive deficits: A critical appraisal of the Feuerstein model. *Journal of Mental Deficiency Research, 27,* 79-92.

Braun, W.I., & Holland, P.W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In: P.W. Holland, & D.B. Rubin (red.). *Test equating* (pp. 9-49). New York: Academic Press.

Brennan, R.L. (1992). Elements of generalizability theory. Iowa City: ACT.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296-322.

Bügel, K. (1991). Sexeverschillen in onderwijsprestaties in Nederland: Een overzicht van de literatuur en enkele nieuwe gegevens. *Pedagogische Studiën, 68,* 350-370.

Bügel, K. (1993). Tekstbegrip moderne vreemde talen: De invloed van sekse en tekstonderwerp op de scores van centrale examens. *Tijdschrift voor Onderwijs-wetenschappen, 23,* 162-176.

Bügel, K., & Glas, C.A.W. (1991). Item specifieke verschillen in prestaties tussen jongens en meisjes bij tekstbegrip examens moderne vreemde talen. *Tijdschrift voor Onderwijsresearch, 16,* 337-351.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the mulititrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Campbell, D.T., & Stanley, J.C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Coombs, C.H. (1964). *A theory of data.* New York: Wiley.

Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement, 18,* 183-204; *19,* 331-332.

Cicchetti, D.V. (1972). A new measure of agreement between rank ordered variables. *In Proceedings of the 80th Annual Convention of the American Psychological Association 7,* 17-18.

Cicchetti, D.V. (1976). Assessing inter-rater reliability for rating scales: Resolving some basic issues. *British Journal of Psychiatry, 129,* 452-456.

Cochran, W. G. (1977). *Sampling techniques.* New York: Wiley.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psycholo-*
gical Measurement, 20, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provisions for scales disagreement of partial credit. *Psychological Bulletin, 70,* 213-220.

Cornfield, J., & J.W. Tukey (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics, 27,* 907-949.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Cronbach, L.J. (1971). Test validation. In: R.L. Thorndike (red.). *Educational Measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Cronbach, L.J., & Furby, L. (1970). How we should measure "change" - or should we? *Psychological Bulletin, 74,* 68-80.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Dirickx, Y.M.I., Baas, S.M., & Dorhout, B. (1987). *Operationele research*. Schoonhoven: Academic Service.

Divgi, D.R. (1981). *Two direct procedures for scaling and equating tests with item response theory*. Paper presented at the annual meeting of the National Council on Measurement in Education.

Dixon, W.J. (red.) (1992). *BMDP statistical software manual: Vol. 1 and 2.* Berkeley: University of California Press.

Dousma, T., & Horsten, A. (1989). *Tentamineren.* Groningen: Wolters-Noordhoff.

Drenth, P.J.D., & Sijtsma, K. (1990). *Testtheorie: Inleiding in de theorie van de psycholo- gische test en zijn toepassingen.* Houten: Bohn Stafleu Van Loghum.

Dunn, G. (1989). *Design and analysis of reliability studies: The statistical evaluation of measurement errors.* New York: Oxford University Press.

Ebel, R.L. (1967). The relation of item discrimination to test reliability. *Journal of Educational Measurement, 4,* 125-128.

Ebel, R.L. (1972). *Essentials of educational measurement.* Englewood Cliffs: Prentice-Hall.

Ebel, R.L. (1983). The practical validation of tests of ability. *Educational Measurement: Issues and Practice, 2,* 7-10.

Ebel, R.L., & Frisbie, D.A. (1986). *Essentials of educational measurement.* Englewood Cliffs: Prentice Hall.

Eggen, T.J.H.M. (1990). Innovative procedures in the calibration of measurement scales. In: W.H. Schreiber, & K. Ingenkamp (red.). *International developments in large scale*

assessment (pp.199-212). Windsor, Berkshire: NFER-NELSON.

Eggen, T.J.H.M., & Verhelst, N.D. (1992). *Item calibration in incomplete testing designs.* (Measurement and Research Department Reports 92-3). Arnhem: Cito.

Elliott, C.D., Murray, D.J., & Saunders, R. (1977). *Goodness of fit to the Rasch model as a criterion of test unidimensionality.* Manchester: University of Manchester.

Evers, A., Vliet-Mulder, J.C. van, & Laak, J. ter. (1992). *Documentatie van tests en testresearch in Nederland.* Amsterdam: Nederlands Instituut van Psychologen.

Fagot, R.F. (1991). Reliability of ratings for multiple judges: Intraclass correlation and metric scales. *Applied Psychological Measurement, 15,* 1-11.

Fagot, R.F. (1993). A generalized family of coefficients of relational agreement for numeri-
cal scales. *Psychometrika, 58,* 357-370.

Feldt, L.S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika, 30,* 357-370.

Feldt, L.S. (1993). The relationship between the distribution of item difficulties and test reliability. *Applied Measurement in Education 6,* 37-49.

Feldt, L.S., Steffen, M., & Gupta, N.C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement, 9,* 351-361.

Feldt, L.S, & Brennan, R.L. (1989). Reliability. In: R.L. Linn (red.). *Educational Measure-*
ment (3rd ed., pp. 105-146). Washington, DC: American Council on Education.

Ferguson, G.A., & Takane, Y. (1989). *Statistical analysis in psychology and education.* New York: McGraw-Hill.

Feuerstein, R. (1980). *Instrumental enrichment: An intervention program for cognitive modifiability.* Baltimore: University Park Press.

Fischer, G.H. (1972). *A step towards dynamic test-theory.* (Research Bulletin Nr. 10/72). Universität Wien: Psychologisches Institut.

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-373.

Fischer, G.H. (1974). *Einführung in die theorie psychologischer tests.* Bern: Huber.

Fischer, G.H. (1981). On the existence and uniqueness of maximum likelihood estimates in the Rasch model. *Psychometrika, 46,* 59-77.

Fischer, G.H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48,* 3-26.

Fischer, G.H. (in voorbereiding). Derivations of the Rasch model. In: G.H. Fischer, & I.W. Molenaar (red.). *Rasch models: Their foundations, recent developments and applica-*

tions.

Fischer, G.H., & Scheiblechner, H. (1970). Algorithmen und programme für das probabi- listische testmodell von Rasch. *Psychologische Beiträge, 12,* 23-51.

Flanagan, J.C. (1951). Units, scores and norms. In: E.F. Lindquist (red.). *Educational measurement* (pp. 695-763). Washington, DC: American Council on Education.

Fleiss, J.L. (1986). *The design and analysis of clinical experiments.* New York: Wiley.

Fleiss, J.L., Cohen, J., & Everitt, B.S. (1969) Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72,* 5, 323-327.

Fleiss, J.L., & Shrout, P.E. (1978). Approximate interval estimation for a certain intraclass correlation coefficient. *Psychometrika, 43,* 259-262.

Follman, D. (1988). Consistent estimation in the Rasch model based on nonparametric margins. *Psychometrika, 53,* 553-562.

Freeman, M.F., & Tukey, J.W. (1950). Transformations related to the angular and square root. *The Annals of Mathematical Statistics, 21,* 607-611.

Frisbie, D.A. (1988). Reliability of scores from teacher-made tests. *Educational Measure- ment: Issues and practice, 7,* 53-63.

Glas, C.A.W. (1981). *Het Raschmodel bij data in een onvolledig design*. (PSM-Progress reports, 81-1). Utrecht: Vakgroep PSM van de subfaculteit Psychologie.

Glas, C.A.W. (1989). *Contributions to estimating and testing Rasch models.* Arnhem: Cito.

Glas, C.A.W. (1992). A Rasch model with a multivariate distribution of ability. In: M. Wilson (red.). *Objective measurement: Theory into practice: Vol. 1* (pp. 236-258). Norwood: Ablex.

Glas, C.A.W., & Verhelst, N.D. (1989). Extensions of the partial credit model. *Psychometrika, 54,* 635-659.

Glas, C.A.W., & Verhelst, N.D. (in voorbereiding). Testing the Rasch model. In: G.H.Fischer, & I.W.Molenaar (red.). *Rasch models: Their foundations, recent developments and applications.*

Green, S.B., & Lissitz, R.W. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37,* 827-838.

Groot, A.D. de (1966). *Vijven en zessen.* Groningen: Wolters.

Groot, A.D. de, & Naerssen, R.F. (1973). *Studietoetsen, construeren, afnemen, analyseren: Deel I en II.* Den Haag: Mouton.

Gruijter, D.N.M. de (1985). Compromise models for establishing examination standards. *Journal of Educational Measurement, 22,* 263-269.

Guilford, J.P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education.* Tokyo: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Gustafsson, J.E. (1979). *PML: A computer program for conditional estimation and testing in the Rasch model for dichotomous items.* (Reports from the Institute of Education, nr. 63). Göteborg: University of Göteborg.

Guttman, L. A. (1950). The Basis of Scalogram Analysis. In: S.A. Stouffer, L.A. Gutmann, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, & J.A. Clausen (red.). *Measurement and prediction: Studies in social psychology in World War II: Vol. 4.* Princeton: Princeton University Press.

Guttman, L. A. (1954). A new approach to factor analysis: The radex. In: P.F. Lazersfeld (red.). *Mathematical thinking in the social sciences* (pp. 258-348). New York: Colombia University Press.

Haggard, E.A. (1958). *Intraclass correlation and the analysis of variance.* New York: The Dryden Press.

Hambleton, R.K., & Novick, M.R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10,* 159-170.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer Academic Publishers.

Hambleton, R.K., & Rogers, H.J. (1989). Detecting potentially biased test items: Compa- rison of IRT area and Mantel-Haenszel methods. *Applied Psychological Measurement, 2,* 313-334.

Harris, D.H., & Crouse, J.D. (1992). *A study of criteria used in equating.* Paper presented at the annual meeting of the National Council on Measurement in Education.

Heinen, T. (1993). *Discrete latent variable models.* Proefschrift, Katholieke Universiteit Brabant.

Henrysson, S. (1963). Correction of item-total correlations in item analysis. *Psychometrika, 28,* 211-218.

Hofstee, W.K.B. (1977). Cesuurprobleem opgelost. *Onderzoek van Onderwijs, 6/2,* 6-7.

Hofstee, W.K.B. (1981). *Psychologische uitspraken over personen.* Deventer: Van Loghum Slaterus.

Hofstee, W.K.B. (1983). The case for compromise in educational selection and grading. In Anderson, S.B., & Helmick, J.S. (red.). *On educational testing.* San Francisco: Jossey-Bass.

Hoijtink, H., & Boomsma, A. (1991). *Statistical inference with latent ability estimates.* (Prepublication Department of Statistics and Measurement Theory). Groningen: University of Groningen.

Hoijtink, H. (red.). (1993). *Kwantitatieve Methoden nr. 42.*

Holland, P.W., & Rubin, D.B. (1982). *Test equating.* New York: Academic Press.

Holland, P.W., & Thayer, D.T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In: H. Wainer, & H.I. Braun (red.). *Test validity* (pp.129-145). Hillsdale: Lawrence Erlbaum.

Hommel, G. (1983). Tests of the overall hypothesis for arbitrary dependence structures. *Biometrical Journal, 25,* 423-430.

Houston, W.M., Raymond, M.R., & Svec, J.C. (1991). Adjustments for rater effects in performance assessment. *Applied Psychological Measurement, 15,* 409-421.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). *Item response theory: Applications to psychological measurement.* Homewood: Dow-Jones Irwin.

Iker, H.P., & Perry, N.C.A. (1960). A further note concerning the reliability of the point-biserial correlation. *Educational and Psychological Measurement, 20,* 505-507.

Imbos, Tj. (1989). *Het gebruik van einddoel toetsen bij aanvang van de studie.* Proefschrift, Rijksuniversiteit Limburg.

Inspectierapport. (1992). *Examens op punten getoetst: Onderzoek naar de ontwikkeling van de normen bij de centrale examens in het voortgezet onderwijs.*

James, L.R., Demaree, R.G., & Wolf, G. (1984). Estimating within-group interrater relia- bility with and without response bias. *Journal of Applied Psychology, 69,* 85-98.

Jannarone, R.J. (1986). Conjunctive item response theory kernels. *Psychometrika, 51,* 357-373.

Jansen, G.G.H. (1979). *Het meten van veranderingen in de klassieke testtheorie.* (Bulletinreeks nr. 2). Arnhem: Cito.

Jarjoura, D. (1983). Best linear prediction of composite universe scores. *Psychometrika, 48,* 525-539.

Jazwinsky, A.H. (1970). *Stochastic processes and filtering theory.* New York: Academic Press.

Johnson, H.M. (1935). Some neglected principles in aptitude testing. *American Journal of Psychology, 47* 159-165.

Jonge, H. de (1963). *Inleiding tot de medische statistiek: Deel I.* Groningen: Wolters-Noordhoff.

Jöreskog, K.G. (1970). Estimation and testing of simplex models. *The British Journal of Mathematical and Statistical Psychology, 23,* 121-145.

Jöreskog, K.G., & Sörbom, D. (1989). *LISREL 7, user's reference guide.* Mooresville: Scientific Software.

Kamphuis, F.H., & Engelen, R.J.H. (in voorbereiding). Estimation and testing of structured latent ability covariance matrices in IRT models.

Kane, M.T. (1992). An argument-based approach to validation. *Psychological Bulletin, 112,* 527-535.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49,* 223-245.

Kelderman, H. (1988). *Loglinear multidimensional IRT model for polytomously scored items.* (Research Report 88-17). Enschede: Universiteit Twente.

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika, 54,* 681-697.

Kelderman, H., & Steen, R. (1988). *LOGIMO I: Loglinear item response theory modeling.* (Computer Program). Enschede: University of Twente, Department of Educational Technology.

Kelderman, H., & Macready, G.B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement, 27,* 307-327.

Kelley, T.L. (1947). *Fundamentals of statistics.* Cambridge: Harvard University Press.

Kendall, M., & Stuart, A. (1973). *The advanced theory of statistics: Vol. 2.* Londen: Griffin.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27,* 887-903.

Klauer, K.C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika, 56,* 213-228.

Kolen, M.J. (1988). Defining score scales in relation to measurement error. *Journal of Educational Measurement, 25,* 97-110.

Koppen, M.G.M. (1987). On finding the bidimension of a relation. *Journal of Mathematical Psychology, 3*1, 155-178.

Knol, D.L. (1986). *Een overzicht van meerdimensionale itemresponsmodellen*. (Rapport R-86-5). Enschede: Univeriteit Twente, Faculteit TO, vakgroep OMD.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement, 30,* 61-70.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Beverly Hills: Sage Publications.

Kuder, G.F., & Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151-160.

Lahey, M.A., Downey, R.G., & Saal, F.E. (1983). Intraclass correlations: There's more than meets the eye. *Psychological Bulletin, 93,* 586-595.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174.

Laros, J.A., & Tellegen, P.J. (1991). *Construction and validation of the SON-R 5½-17, the Snijders-Oomen non-verbal intelligence test.* Groningen: Wolters-Noordhoff.

Lazarsfeld, P.F. (1950). Logical and mathematical foundations of latent structure analysis. In: S.A. Stouffer. *Studies in social psychology in World War II, IV*. Princeton, NJ: Princeton University Press.

LBR (1988). *Psychologische tests en allochtonen.* Symposiumverslag 1987, LBR-Reeks nr. 6.

LBR (1990). *Toepasbaarheid van psychologische tests bij allochtonen.* Rapport van de testscreeningscommissie ingesteld door het LBR in overleg met het NIP, LBR-Reeks nr. 11.

Leeuw, J. de, & Verhelst, N.D. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics, 11*, 183-196.

Leeuwe, J.F.J. van (1990). *Probabilistic conjunctive models.* Proefschrift. Nijmegen: NICI.

Linden, W.J. van der (red.). (1982). Aspects of criterion-referenced measurement. *Evalua- tion in Education: An International Review Series, 5.*

Linden, W.J. van der (1983). *Van standaardtest naar itembank*. Universiteit Twente (oratie).

Linden, W.J. van der (1984). Some thoughts on the use of decision theory to set cutoff scores: Comment on De Gruijter and Hambleton. *Applied Psychological Measurement, 8,* 9-17.

Linden, W.J. van der (1985). Decision theory in educational research and testing. In: T. Husén, & T.N. Postlethwaite (red.). *International encyclopedia of education: Research and studies.* Oxford: Pergamon Press.

Linden, W.J. van der, & Boekkooi-Timminga, E. (1988). A zero-one programming approach to Gulliksen's matched random subtests method. *Applied Psychological Measurement, 12,* 201-209.

Linden, W.J. van der, & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika, 54,* 237-247.

Lindsay, B., Clifford, C.C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association, 86,* 96-107.

Linn, R.L. (red.). (1989). *Intelligence: Measurement, theory, and public policy*. Chicago: University of Illinois Press.

Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Livingston, S.A., & Zieky, M.J. (1982). *Passing scores: A manual for setting standards of performance on educational and performance tests.* Princeton, NJ: Educational Testing Service.

Lord, F.M. (1950). *Notes on comparable scales for test scores* (Research Bulletin 50-48). Princeton, NJ: Educational Testing Service.

Lord, F.M. (1952). The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika, 17*, 181-194.

Lord, F.M. (1953). On the statistical treatment of football numbers. *The American Psycholo- gist, 8,* 750-751.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale: Lawrence Erlbaum.

Lord, F.M. (1983a). Unbiased estimators of ability parameters, their variance and of their parallel-forms reliability. *Psychometrika, 48,* 233-245.

Lord, F.M. (1983b). *Estimating the imputed social cost of errors of measurement.* (Report RR-83-33-ONR). Princeton, NJ: Educational Testing Service.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading: Addison-Wesley.

Lord, F.M. & Wingerskey, M.S. (1983). Comparison of IRT true-score and equipercentile observed-score 'equatings'. *Applied Psychological Measurement, 8, 453-461.*

MacCann, R.G. (1990). Derivations of observed score equating methods that cater to populations differing in ability. *Journal of Educational Statistics, 15*, 146-170.

Maris, E. (1992). *Psychometric models for psychological processes and structures.* Proefschrift, Universiteit Leuven.

Martin-Löf, P. (1973). *Statistika Modeller: Anteckningar från seminarier Lasåret 1969-1970, utarbetade av Rolf Sunberg. Obetydligt ändrat nytryck, oktober 1973.* Stockholm: Institutet för Försäkringsmatematik och Matematisk Statistik vid Stockholms Universitet.

Martin-Löf, P. (1974). The notion of redundancy and its use as a quantitative measure if the discrepancy between a statistical hypothesis and a set of observational data. *Scandinavian Journal of Statistics, 1*, 3-18.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Masters, G.N., & Wright, B.D. (1984). The essential process in a family of measurement models. Psychometrika, 49, 529-544.

Maxwell, A.E., & Pilliner, A.E.G. (1968). Deriving coefficients of reliability and agreement. *The British Journal of Mathematical and Statistical Psychology, 21,* 105-116.

McKinley, R.L., & Reckase, M.D. (1983). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods and Instrumentation, 15,* 389-390.

Meerling (1981). *Methoden en technieken van psychologisch onderzoek: Deel 1.* Meppel: Boom.

Mellenbergh, G.J. (1977). The replicability of measures. *Psychological Bulletin, 84,* 378-384.

Mellenbergh, G.J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105-118.

Mellenbergh, G.J. (1983). Conditional item bias methods. In: S.H. Irvine, & W.J. Berry (red.). *Human assessment and cultural factors* (pp. 293-302). New York: Plenum Press.

Mellenbergh, G.J. (1985). Vraag-onzuiverheid: definitie, detectie en onderzoek. *Nederlands Tijdschrift voor Psychologie, 40,* 425-435.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In: H. Wainer, & H.I. Braun (red.). *Test validity* (pp.33-45). Hillsdale: Lawrence Erlbaum.

Messick, S. (1989). Validity. In: R.L. Linn (red.). *Educational Measurement* (3rd ed., pp. 13-103). Washington, DC: American Council on Education.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In: R.L. Linn (red.). *Educational Measurement* (3rd ed., pp. 335-366). Washington, DC: American Council on Education.

Mills, C.N., & Melican, G.J. (1987). *A preliminary investigation of three compromise methods for establishing cut-off scores.* (Report RR-87-14). Princeton, NJ: Educational Testing Service.

Mislevy, R.J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359-381.

Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177-195.

Mislevy, R.J., & Bock, R.D. (1986). *PC-BILOG: Maximum likelihood item analysis and test scoring with logistic models for binary items.* Mooresville: Scientific Software.

Mislevy, R.J., & Wu, P.K. (1988). *Inferring examinee ability when some item responses are missing.* (Research Report RR-88-48-ONR). Princeton, NJ: Educational Testing Service.

Mislevy, R.J., & Sheenan, K.M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54,* 661-680.

Moelands, A.H.J. (1988). *Entreetoets: Basisvaardigheden taal, rekenen en informatieverwerking (Verantwoording)*. Arnhem: Cito.

Mokken, R.J. (1971). *A theory and procedure of scale analysis.* Den Haag: Mouton.

Molenaar, I.W. (1981). *Programmabeschrijving van PML (versie 3.1) voor het Raschmodel.* (Heymans Bulletins Psychologische Instituten R.U.Groningen, nr. HB-81-538-RP). Groningen: Rijksuniversiteit Groningen.

Molenaar, I.W. (1983). *Item steps.* (Heymans Bulletins Psychologische Instituten R.U Groningen, nr. HB-83-630-EX). Groningen: Rijksuniversiteit Groningen.

Molenaar I.W., & Hoijtink, H (1990). The many null-distributions of person fit indices. *Psychometrika, 55,* 75-106.

Muskens, G.J. (1980). *Frames of meaning - are they measurable?* Proefschrift, Katholieke Universiteit Nijmegen.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika, 49,* 115-132.

Muthén, B. (1989). *LISCOMP: Analysis of linear structural equations with a comprehensive measurement model.* Mooresville: Scientific Software.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Nederlands Instituut van Psychologen. (1988). *Richtlijnen voor ontwikkeling en gebruik van psychologische tests en studietoetsen.* Amsterdam: Nederlands Instituut van Psychologen.

Novick, M.R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3,* 1-18.

Oud, J.H.L., & Mommers (1988). Longitudinale computerondersteunende ondersteuning van lees- en spellingsmoeilijkheden: Een toepassing van het Kalmanfilter in de onderwijs- praktijk. *Tijdschrift voor Onderwijsresearch, 13,* 31-50.

Pennings, A.H. (1988). The development of strategies in embedded figure tasks. *International Journal of Psychology, 23,* 65-78.

Pennings, A.H. (1991). *Individual differences in the development of the restructuring ability in children.* Proefschrift, Rijksuniversiteit Utrecht.

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, norming, and equating. In R.L. Linn (red.). *Eductional Measurement* (3rd ed., pp. 221-262). Washington, DC: American Council on Education.

Popping, R. (1983). *Overeenstemmingsmaten voor nominale data.* Proefschrift, Rijksuniversi- teit Groningen.

Popping, R. (1989). *AGREE: Computing agreement on nominal data, version 5.* (User's manual) Groningen: IEC ProGamma.

Popping, R. (1992). *Taxonomy on nominal scale agreement 1945 - 1990.* Groningen: IEC ProGamma.

Rao, C.R. (1948). Large sample tests of statistical hypothesis concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society, 44,* 50-57.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1961). On the general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* 321-333. Berkeley: University of California Press.

Rasch, G. (1977). *On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements.* Berkeley: University of California Press.

Read, T.R.C., & Cressie, N.A.C. (1988). *Goodness-of-fit statistics for discrete multivariate data.* New York: Springer.

Reckase, M.D., & Mckinley, R.L. (1985). Some latent trait theory in a multidimensional latent space. In: D.I. Weiss (red.). *Proceedings of the 1982 computerized adaptive testing conference* (pp. 151-177). Minneapolis: University of Minnesota.

Rigdon S.E., & Tsutakawa, R.K. (1983). Parameter estimation in latent trait models. *Psychometrika, 48,* 567-574.

Rigdon S.E., & Tsutakawa, R.K. (1986). Estimation for the Rasch model when both ability and difficulty parameters are random. *Journal of Educational Statistics, 12,* 76-86.

Roskam, E.E. (1982). Hypotheses non fingo, een methodologische gevalstudie over onderzoek van intelligentietests. *Nederlands Tijdschrift voor de Psychologie, 37,* 331-359.

Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63,* 581-592.

Rubin, D.B. (1980). Using empirical Bayes techniques in law school validity studies. *Journal of the American Statistical Association, 75,* 801-816.

Saal, F.E., Downey, R.G., & Lahey, M. (1980). Rating the ratings: Assessing the psycho- metric quality of rating data. *Psychological Bulletin, 88,* 413-428.

Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. *(Psycho- metric Monograph No. 17).* Psychometric Society.

Samejima, F. (1972). A general model for free response data. *(Psychometric Monograph No. 18)*. Psychometric Society.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38,* 203-219.

Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika, 42,* 193-198.

Sanders, P.F., Hendrix, A.C., & Luijten, A.J.M. (1984). De beoordeling van de samenvatting Nederlands. *Tijdschrift voor Taalbeheersing, 6,* 241-251.

Sanders, P.F., Theunissen, T.J.J.M., & Baas, S.M. (1989). Minimizing the number of observations: A generalization of the Spearman-Brown formula. *Psychometrika, 54,* 587-598.

Schouten, H.J.A. (1985). *Statistical measurement of interobserver agreement: Analysis of agreement and disagreement between observers.* Proefschrift, Rijksuniversiteit Utrecht.

Shavelson, R.J., & Webb, N.M. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology, 34,* 133-166.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer.* Newbury Park: Sage Publications.

Shepard, L.A. (1993). Evaluating test validity. In: L. Darling-Hammond (red.). *Review of research in education: Vol. 19* (pp.405-450). Washington, DC: American Educational Research Association.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliabi-
lity. *Psychological Bulletin, 86,* 420-428.

Shumway, R.H., & Stoffer, D.S. (1982). An approach to time series smoothing and forcasting using EM algorithm. *Journal of Time Series Analysis,* 3, 253-264.

Siegel, S., & Castellan, N.J.Jr. (1988). *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill.

Sijtsma, K., & Molenaar, I.W. (1987). Reliability of test scores in non-parametric item response theory. *Psychometrika, 52,* 79-97.

Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, series B, 13,* 238-241.

Sirotnik, K. (1970). An analysis of variance framework for matrix sampling. *Educational and Psychological Measurement, 30,* 891-908.

Sluijter, C., Boertien, H., de Klijn, W., & van Roosmalen, W. (1991). *De constructie van plaatsingstoetsen.* (Onderzoeksrapporten beginfase voortgezet onderwijs nr. 6). Arnhem: Cito.

Smith, P.L. (1978). Sampling errors of variance components in small sample multifacet generalizability studies. *Journal of Educational Statistics, 3,* 319-346.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271-295.

Staphorsius, G. (1992a). Welk boek is gemakkelijk, mijnheer ? *RAIN informatiebulletin, 2,* 7-10.

Staphorsius, G. (1992b). *Clib-toetsen.* Arnhem: Cito.

Staphorsius, G., & Krom, R.S.H. (1985a). *Leesbaarheidsindex voor het basisonderwijs.* (Bulletin nr. 36). Arnhem: Cito.

Staphorsius, G., & Krom, R.S.H. (1985b). Predictie van leesbaarheid. *Tijdschrift voor Taal- beheersing, 7,* 192-211.

Stine, W.W. (1989). Interobserver relational agreement. *Psychological Bulletin, 106,* 341 -347.

Suen, H.K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data.* Hillsdale: Lawrence Erlbaum.

Tatsuoka, K.K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 95-110.

Theunissen, T.J.J.M. (1985). Binary programming and test design. *Psychometrika, 50,* 411-420.

Theunissen, T.J.J.M. (1986). Some applications of optimization algorithms in test design and adaptive testing. *Applied Psychological Measurement, 10,* 381-389.

Theunissen, T.J.J.M. (1987). Text banking and test design. *Language Testing, 4,* 1-8.

Thissen, D. (1988). *MULTILOG: Multiple categorical item analysis and test scoring using item response theory*. Mooresville: Scientific Software.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567-577.

Thorndike, R.L. (1951). Reliability. In: E.F. Lindquist (red.). *Educational Measurement* (pp. 560-620). Washington, DC: American Council on Education.

Thorndike, R.L. (1982). *Applied psychometrics.* Boston: Houghton Mifflin Company.

Tinsley, H.E.A., & Weiss, D.J. (1975). Interrater reliability and agreement of subjective judgements. *Journal of Counseling Psychology, 23,* 358-376.

Uebersax, J.S. (1984). *Reliablility, validity and the kappa coefficient.* (Technical Report No. 12). Austin: University of Texas.

Uebersax, J.S. (1991). *Quantitative methods for the analysis of observer agreement: Towards a unifying model.* Santa Monica: RAND Corporation.

Uiterwijk, J.H. (1990). Verschillen tussen autochtonen en allochtonen bij de overgang van basisonderwijs naar voortgezet onderwijs. In: C.A.C. Klaassen, & P.L.M.

Jungbluth (red.). *Onderwijs researchdagen 1990, onderwijs en samenleving.* Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.

Uiterwijk, J.H., & Engelen, R.J.H. (1993). *Verantwoording eindtoets basisonderwijs 1990.* Arnhem: Cito.

Umesh, U.N., Peterson, R.A., & Sauber, M.H. (1989). Interjudge agreement and the maximum value of kappa. *Educational and Psychological Measurement, 49,* 835-850.

Vale, C.D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10,* 333-344.

Verhelst, N.D. (1989). Informatiewinst bij vertakt toetsen. In: W.J. van der Linden, & L.J.Th. van der Kamp (red.). *Meetmethoden en data-analyse* (pp. 89-96). Lisse: Swets en Zeitlinger.

Verhelst, N.D. (1993). *On the standard errors of parameter estimates in the Rasch model.* (Measurement and Research Department Reports 93-1). Arnhem: Cito.

Verhelst, N.D., Glas, C.A.W., & van der Sluis, A. (1984). Estimation problems in the Rasch model: The basic symmetric functions. *Computational Statistics Quarterly, 1,* 245-262.

Verhelst, N.D., & Eggen, T.J.H.M. (1989). *Psychometrische en statistische aspecten van peilingsonderzoek*. (PPON-rapport, nr. 4). Arnhem: Cito.

Verhelst, N.D., & Kamphuis, F.H. (1989). *Statistiek met* $\hat{\theta}$.(Bulletinreeks nr. 77). Arnhem: Cito.

Verhelst, N.D., Verstralen, H.H.F.M., & Eggen, T.J.H.M. (1991). *Finding starting values for the item parameters and suitable discrimination indices in the one-parameter logistic model.* (Measurement and Research Department Reports 91-10). Arnhem: Cito.

Verhelst, N.D., & Veldhuijzen, N.H. (1991). *A new algorithm for computing elementary symmetric functions and their first and second derivatives.* (Measurement and Research Department Reports 91-1). Arnhem: Cito.

Verhelst, N.D., & Verstralen, H.H.F.M. (1991). *The partial credit model with non-sequential solution strategies.* (Measurement and Research Department Reports 91-5). Arnhem: Cito.

Verhelst, N.D., & Glas, C.A.W. (in druk). A dynamic generalization of the Rasch model. *Psychometrika, 58.*

Verhelst, N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1993). *OPLM: One parameter logistic model.* Computer program and manual. Arnhem: Cito.

Verhelst, N.D., Verstralen.H.H.F.M., & Jansen, M.G.H. (1993) *A logistic model for time limit tests.* (Measurement and Research Department Reports 92-1). Arnhem: Cito.

Verschoor, A.J. (1991). *Optimal test design.* (Computer progamm and manual). Arnhem: Cito.

Verschoor. A.J., & Sanders, P.F. (1993). *Parallel test construction using the framework of classical test theory*. (Measurement and Research Department Reports 93-2). Arnhem: Cito.

Verstralen, H.H.F.M., & Verhelst, N.D. (1992). *The sample strategy of a test information function in computerized test design.* (Measurement and Research Department Reports 91-6). Arnhem: Cito.

Vogel, M., & Washburne, C. (1928). An objective method of determining grade placement of children's reading material. *Elementary School Journal, 28,* 373-381.

Wainer, H., & Mislevy, R.J. (1990). Item response theory, item calibration and proficiency estimation. In: H. Wainer (red.). *Computerized adaptive testing: A primer* (pp. 65-101). Hillsdale: Lawrence Erlbaum.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society, 54,* 426-482.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427-450.

Weiss, D.J. (red.). (1983). *New horizons in testing.* New York: Academic Press.

Wijnstra, J.M. (1988). *Balans van het rekenonderwijs in de basisschool.* Arnhem: Cito.

Wilson, D.T., Wood, R., & Gibbons, R.T. (1991). *TESTFACT.* Chicago: Scientific Software.

Wilson, M., & G.N. Masters, (1993). The partial credit model and null categories. *Psycho- metrika, 58,* 87-99.

Witkin, H.A. (1950). Individual differences in ease of perception of embedded figures. *Jour- nal of Personality, 19,* 1-15.

Witkin, H.A., & Goodenough, D.R. (1981). Cognitive styles: Essence and origins. *Psychological Issues* (Monograph 51). New York: International Universities Press.

Wollenberg, A.L. van den (1979). *The Rasch model and time limit tests.* Nijmegen: Studentenpers.

Wollenberg, A.L. van den (1982). Two new test statistics for the Rasch model. *Psychometrika, 47,* 123-140.

Wright, B.D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23-48.

Wright, B.D., & Mead, R.J. (1977). *BICAL: Calibrating items and scales with the Rasch model.* (Research Memorandum 23). Chicago: University of Chicago, Department of Education, Statistical Laboratory.

Wright, B.D., & Stone, M.H. (1979). *Best test design.* Chicago: MESA Press.

Yen, W.M. (1981). Using simultaneous results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245-262.

Yen, W.M. (1984). Tau-equivalence and equipercentile equating. *Psychometrika, 48,* 353- 369.

Zegers, F.E. (1989). Het meten van overeenstemming. *Nederlands Tijdschrift voor de Psychologie, 44,* 145-156.

Zegers, F.E. (1991). Coefficients for interrater agreement. *Applied Psychological Measurement, 15,* 321-333.

Zieky, M.J. (1987). *Methods of setting standards of performance on criterion referenced tests.* Paper presented at the 13th International Conference of the IAEA, Bangkok.

Zwinderman, A.H. (1991). *Studies of estimating and testing Rasch models.* (NICI Technical Report 91-02). Nijmegen: NICI.

1

Inleiding

Aan het construeren van studietoetsen, psychologische tests en andere sociaalwetenschappelijke meetinstrumenten kan een kwalitatieve en een kwantitatieve component onderscheiden worden. Het belangrijkste aspect van de kwalitatieve component betreft het ontwikkelen van de vragen of opdrachten waaruit het meetinstrument bestaat. De kwantitatieve component betreft het analyseren van antwoorden van personen op vragen of opdrachten. De kwantitatieve component van het toetsconstructieproces vormt het aandachtsgebied van de psychometrie. In dit boek wordt beschreven hoe door toepassing van psychometrische theorieën en statistische technieken de kwaliteit van meetinstrumenten beschreven, onderzocht en verbeterd kan worden.

Dit hoofdstuk bestaat uit twee verschillende onderdelen. Het doel van het eerste onderdeel is de bijdrage van de psychometrie voor de testpraktijk aan te geven. Daartoe wordt eerst in paragraaf 1.1 aan de hand van testindelingen een overzicht gegeven van de meetinstrumenten die mede met behulp van de psychometrie ontwikkeld zijn. Vervolgens wordt in paragraaf 1.2 beschreven wat de psychometrie bijdraagt aan de verschillende fasen van het toetsconstructieproces. In het tweede onderdeel van dit hoofdstuk worden de belangrijkste psychometrische aspecten van meetinstrumenten besproken. In paragraaf 1.3 wordt het valideren van meetinstrumenten besproken. In paragraaf 1.4 worden verschillende psychometrische theorieën besproken die bij het construeren van meetinstrumenten worden toegepast.

1.1 Testindelingen

De 'Documentatie van tests en testresearch in Nederland' (Evers, Van Vliet-Mulder, & Ter Laak, 1992) bevat een overzicht van bijna vierhonderd Nederlandstalige psychologische en andere meetinstrumenten en van het onderzoek dat ermee is verricht. Met meetpretentie als indelingsprincipe worden in dat overzicht drie klassen of soorten meetinstrumenten onderscheiden:

De eerste klasse bevat meetinstrumenten die als meetpretentie hebben stabiele persoonlijkheidskenmerken van personen te meten. Het gaat hierbij om kenmerken die zoveel mogelijk onafhankelijk zijn van bijvoorbeeld een arbeids- of opleidingssituatie. Voorbeelden van meetinstrumenten uit deze klasse zijn intelligentietests en persoonlijkheidsvragenlijsten. Ook de verborgen-figurentest die in hoofdstuk 7 besproken wordt, is een meetinstrument uit deze klasse.

De tweede klasse betreft meetinstrumenten die als meetpretentie hebben kenmerken te meten van personen in interactie met een (klasse van) situatie(s). Tot deze klasse behoren meetinstrumenten zoals beroepeninteressevragenlijsten en studietoetsen. In de navolgende hoofdstukken worden met name studietoetsen besproken. Een algemeen bekende Nederlandse studietoets is de Eindtoets Basisonderwijs (Uiterwijk & Engelen, 1993).

Bij de derde klasse gaat het om meetinstrumenten waarmee personen (beoordelaars) een oordeel over bepaalde situaties geven, bijvoorbeeld het oordeel van chefs over taakinhouden en taakkenmerken in de arbeidssituatie. In hoofdstuk 12 worden verschillende beoordelingssituaties besproken waarbij beoordelaars de meetinstrumenten zijn.

In de 'Richtlijnen voor ontwikkeling en gebruik van psychologische tests en studietoetsen' (1988) wordt een onderscheid gemaakt tussen 'test' en 'studietoets'. De term test wordt gebruikt voor meetinstrumenten die geschiktheid of aanleg meten. In voorgaande indeling behoren deze tests tot de meetinstrumenten uit de eerste klasse. De term studietoets wordt gebruikt voor meetinstrumenten die vaardigheden meten, bijvoorbeeld reken- of leesvaardigheid, die het resultaat zijn van onderwijs, training of instructie.

De indeling op basis van meetpretentie is een van de vele mogelijke indelingsprincipes voor het indelen van meetinstrumenten. Drenth en Sijtsma (1990, p. 36-63) bespreken drie testindelingen. De eerste indeling is gebaseerd op het gedrag van de onderzochte persoon. Hierbij is het belangrijkste onderscheid dat tussen tests voor prestatieniveau, bijvoorbeeld intelligentietests, en tests voor gedragswijze, bijvoorbeeld zelfbeoordelingen. Een tweede testindeling is die op basis van verschillende wijzen van instructie en afname. Twee belangrijke onderscheidingen hierbij zijn die tussen de individuele test en de groepstest en die tussen de snelheidstest ('speed test') en de niveautest ('power test'). De derde testindeling is een indeling die gebaseerd is op de aard van de testvragen, bijvoorbeeld tussen toetsen met gesloten vragen ('multiple choice') en toetsen met open vragen.

In dit boek worden meetinstrumenten onderscheiden op basis van het doel dat met het meetinstrument beoogd wordt. Aangezien dit indelingsprincipe geen inhoudelijke

onderscheidingen tussen meetinstrumenten maakt, heeft daarmee ook een terminologisch onderscheid zoals dat tussen test en toets geen betekenis meer. De termen test en toets worden in dit boek dan ook als synoniem gebruikt. Aangezien het toepassingsgebied van dit boek met name studietoetsen betreft, zal in de meeste gevallen de term toets gebezigd worden.

Met het doel van de toets als indelingsprincipe kunnen drie categorieën toetsen onderscheiden worden. De eerste categorie betreft toetsen waarvan het doel is het leren en onderwijzen in de klas te ondersteunen en te sturen. Deze toetsen geven de docent informatie over de vorderingen van elke leerling waarop de docent zijn onderwijs aan zijn leerlingen kan baseren. In hoofdstuk 10 worden voorbeelden van toetsen uit deze categorie besproken.

De tweede categorie betreft toetsen waarvan het doel is uitspraken te doen over hoe onderwijsprogramma's of onderwijssystemen functioneren. Deze toetsen zijn in de eerste plaats bedoeld om informatie aan bijvoorbeeld leerplanontwikkelaars of beleidmakers te geven. Tot deze categorie behoren de toetsen die onderdeel uitmaken van het peilingsonderzoek dat in hoofdstuk 7 besproken wordt.

De derde categorie betreft toetsen die selectie, plaatsing of certificering van leerlingen tot doel hebben. We spreken van selectie als de toets tot doel heeft leerlingen toe te laten of af te wijzen voor een opleiding. Deze toetsen worden met name gebruikt door opleidingen met een beperkt aantal opleidingsplaatsen. De selectie zal strenger zijn naarmate het aantal opleidingsplaatsen beperkter en de opleiding duurder is, bijvoorbeeld de toelating voor de opleiding tot piloot. Wanneer het doel van de toets is leerlingen naar een bepaald onderwijsprogramma te verwijzen, spreken we van plaatsing of classificatie. Voorbeelden zijn toetsen die gebruikt worden om een leerling naar een school voor speciaal onderwijs te verwijzen, of toetsen die gebruikt worden om te beslissen of een leerling na afsluiting van de brugperiode naar mavo, havo of vwo moet gaan. We spreken van certificering als het doel van de toets is te beslissen of leerlingen de leerinhouden van het onderwijsprogramma waaraan zij hebben deelgenomen wel of niet beheersen. De bekendste voorbeelden zijn de zeer vele examens en tentamens die in alle vormen van onderwijs afgenomen worden. Voor bepaalde opleidingen geldt dat de leerlingen na het behalen van een aantal certificaten in het bezit kunnen komen van een diploma. Beheersingsbeslissingen veronderstellen een zogenaamde drempel of cesuur die aangeeft welke toetsscore als laagste voldoende prestatie aangemerkt kan worden. In hoofdstuk 13 worden methoden voor cesuurbepaling besproken.

1.2 Toetsconstructie

Het constructieproces van een toets kan in een aantal fasen uiteen worden gelegd. Het proces begint met het operationaliseren van de vaardigheid die gemeten wordt en het vaststellen van het gebruiksdoel van de toets en eindigt met het schrijven van de handleiding en de verantwoording van de toets. Tussen de eerste en laatste fase moeten talrijke beslissingen genomen en activiteiten ondernomen worden. In onderstaande beschrijving van het toetsconstructieproces worden acht fasen onderscheiden en toegelicht. Bij deze beschrijving zijn de volgende twee opmerkingen van belang. De eerste opmerking is dat de beschrijving niet geïnterpreteerd moet worden als dat het toetsconstructieproces altijd uit acht fasen zou bestaan. De beschrijving is met name van toepassing op studietoetsen maar zelfs daar kan afhankelijk van de toets het proces uit meer of minder fasen bestaan. De tweede opmerking is dat in de beschrijving het toetsconstructieproces lineair verloopt, terwijl het proces in werkelijkheid eerder iteratief zal zijn. De output van de ene fase is weliswaar de input voor de volgende fase, maar dit betekent niet dat men op beslissingen die in een bepaalde fase genomen zijn niet kan of moet terugkomen.

Fase 1: Doelspecificatie

De eerste fase van het toetsconstructieproces bestaat uit het operationaliseren van de vaardigheid die de toets moet meten en het vaststellen van het gebruiksdoel van de toets. De plaatsingstoetsen Engels voor de brugklas operationaliseren het meten van de vaardigheid Engels als reproduktieve en produktieve aspecten van leesvaardigheid (Sluijter, Boertien, De Klijn, & Van Roosmalen, 1991). Als gebruiksdoel van de plaatsingstoetsen wordt het bepalen van de meest geschikte categorale onderwijsvorm voor leerlingen na afsluiting van de brugperiode genoemd.

Fase 2: Toetsspecificatie

Op basis van de operationalisatie van de te meten vaardigheid en het gebruiksdoel van de toets, worden in deze fase de kenmerken van de toets vastgesteld. Hieronder wordt een niet uitputtende opsomming van vragen gegeven waarmee de toetsconstructeur bij de constructie van een toets te maken kan krijgen (Millman & Greene, 1989, p. 339). De eerste drie vragen betreffen externe randvoorwaarden waarmee de toetsconstructeur rekening moet houden. De vragen daarna hebben betrekking op de kenmerken van de toets waarbij de eerste vraag naar de inhoud van de toets de belangrijkste vraag is.

Bij wie wordt de toets afgenomen?

- Voor het vaststellen van de toetsspecificaties is het noodzakelijk te weten bij welke personen de toets met welk doel wordt afgenomen. Het toetsconstructieproces zal anders verlopen wanneer het een toets betreft voor een heterogene groep personen voor een certificaat, dan wanneer het een toets betreft voor een homogene groep personen met het doel om de meest vaardige personen te selecteren.

Hoeveel toetstijd is er beschikbaar?

- Hoewel door praktische omstandigheden de beschikbare toetstijd vaak beperkt is, moeten leerlingen ruim de tijd krijgen voor het beantwoorden van de toets. Wanneer leerlingen te weinig toetstijd krijgen, dan wordt niet alleen het niveau van de uitvoering maar ook de snelheid van uitvoering beoordeeld. In het laatste geval wordt een andere vaardigheid gemeten dan wanneer alleen het niveau van de uitvoering gemeten wordt. Wanneer de toetstijd te beperkt is, kan dat ook betekenen dat te weinig vragen afgenomen kunnen worden om de vaardigheid van de leerlingen verantwoord te kunnen meten.

Hoe wordt de toets afgenomen?

- Wanneer gekozen kan worden tussen een individuele of groepsgewijze toetsafname, zal om praktische redenen groepsgewijze afname altijd de voorkeur verdienen. Groepsgewijze afname gaat meestal gepaard met schriftelijke toetsen. Hiermee worden toetsen bedoeld waarbij de antwoorden op papier gezet moeten worden. Merk op dat dit laatste ook kan gelden voor toetsen die niet in schriftelijke vorm aangeboden kunnen worden, bijvoorbeeld luistertoetsen. Het is ook mogelijk om de vragen via een beeldscherm te presenteren, de antwoorden in de computer in te voeren en te laten scoren. Door deze mogelijkheid wordt individuele toetsafname niet alleen minder bezwaarlijk maar kan voor bepaalde toepassingen zelfs grote voordelen hebben.

Wat is de inhoud van de toets?

- Het vaststellen van de inhoud van de toets is de belangrijkste toetsspecificatie. Voor deze specificatie wordt bij studietoetsen gebruik gemaakt van een toetsmatrijs die meestal twee-dimensionaal is. Bij de eerder genoemde plaatsingstoetsen Engels bestaat de ene dimensie uit zes inhoudscategorieën die aangeven wat een vraag meet (de betekenis van enkele zinnen, relaties tussen alinea's e.d.). De andere dimensie bestaat uit zes gedragscategorieën die aangeven wat een leerling moet kunnen om het goede antwoord op een vraag te kunnen geven (gegevens combineren en vergelijken, conclusies trekken e.d.). Aan de hand van de toetsmatrijs wordt vastgesteld hoe de vragen uit de toets verdeeld zullen worden over de inhouds- en

gedragscategorieën. De toetsen die op basis van de toetsmatrijs geconstrueerd worden, zijn doorgaans een afspiegeling van hetgeen onderwezen is. Dit laatste kan op verschillende manieren (bijv. curriculum- en functieanalyse) onderzocht worden. In het geval van de plaatsingstoetsen Engels werd aan docenten gevraagd of de vakonderdelen waarop de opgaven betrekking hadden door de docent behandeld waren.

In welke vorm wordt de toets afgenomen?

- Wanneer de vaardigheid met een schriftelijke toets gemeten kan worden, zullen meestal gesloten vragen of open vragen gebruikt worden. Een gesloten vraag is een vraagtype waarbij een persoon uit twee of meer alternatieven of antwoordmogelijkheden het goede antwoord moet kiezen. Vanwege het laatste zou het trouwens juister zijn om de term 'gesloten-antwoord vraag' te gebruiken. De open vraag, ofwel de 'open-antwoord vraag', is een vraagtype waarbij een leerling het antwoord zelf moet formuleren. Studietoetsen, bijvoorbeeld schriftelijke examens, bestaan veelal uit subtoetsen of clusters van vragen die structureel bij elkaar horen. Zo bestaan de schriftelijke examens voor de moderne vreemde talen gewoonlijk uit vijf subtoetsen: vijf teksten waarover tien vragen gesteld worden. In de Engelstalige psychometrische literatuur wordt een subtoets aangeduid met de term 'testlet'. Over de voor- en nadelen van beide vraagtypen is veel gepubliceerd. Als voordelen van gesloten vragen worden genoemd dat men in relatief korte tijd veel vragen kan afnemen en dat die vragen machinaal scoorbaar zijn. Nadelen zouden zijn dat het goede antwoord geraden kan worden en dat de hogere cognitieve vaardigheden niet met gesloten vragen gemeten zouden kunnen worden. Dit laatste zou wel mogelijk zijn met open vragen. Nadelen van open vragen zouden zijn dat er vaak maar weinig vragen voorgelegd kunnen worden en dat de antwoorden beoordeeld moeten worden door beoordelaars die het vaak niet met elkaar eens zijn. Dit laatste komt in hoofdstuk 12 aan de orde bij de bespreking van een toets die slechts uit één open vraag bestaat, namelijk de samenvattingsopdracht. Voor het meten van psychomotorische vaardigheden zoals autorijden, typen en timmeren, kan de motorische component niet met een schriftelijke toets gemeten worden. Bij deze zogenaamde 'performance tests' zal de opdracht of toetsvorm veelal gelijk zijn aan de situatie waarin het geleerde moet worden toegepast.

Hoe worden de vragen of opdrachten gescoord?

- We kunnen bij het scoren van vragen een onderscheid maken tussen dichotome en polytome scoring. Bij dichotome scoring wordt uitsluitend aan het goede antwoord een puntenaantal, meestal één scorepunt, toegekend. Bij polytome scoring wordt ook aan een antwoord dat gedeeltelijk goed is een puntenaantal toegekend. Bij de

beoordeling van de antwoorden op open vragen en opdrachten wordt veelal gebruik gemaakt van een antwoordmodel dat de antwoorden en de bij de verschillende antwoorden behorende aantallen scorepunten bevat. Een antwoordmodel is bedoeld om tot een objectieve beoordeling te komen, dat wil zeggen een beoordeling waarbij het aantal toegekende scorepunten onafhankelijk is van de persoon die beoordeelt. In hoofdstuk 12 wordt beschreven hoe de objectiviteit van een antwoordmodel onderzocht kan worden.

Hoeveel items moeten geconstrueerd worden?

- Ook het antwoord op deze vraag is van een groot aantal factoren afhankelijk. In welke mate wil men dat de onderscheiden categorieën uit de toetsmatrijs bevraagd worden? Hoeveel vragen blijven bij een bepaald vak gewoonlijk over na een proeftoets? Hoeveel toetsversies moeten er geconstrueerd worden?

Wat zijn de gewenste psychometrische kenmerken van de items en de toets?

- Afhankelijk van het doel van de toets zullen de items en de bijbehorende toets andere kenmerken dienen te hebben. Aan toetsen die bedoeld zijn om de docent te informeren over de voortgang van de leerlingen zullen andere eisen gesteld worden dan aan toetsen die bedoeld zijn om beleidmakers te informeren over stand van zaken in het basisonderwijs. Wanneer de toets bedoeld is voor het selecteren van goede leerlingen, zal de toets moeilijker items moeten bevatten dan wanneer de toets bedoeld is voor het selecteren van zwakke leerlingen. In verschillende hoofdstukken wordt uitgebreid ingegaan op de relatie tussen toetsdoel en kenmerken van items en toetsen.

Fase 3: Itemconstructie

Vragen en opdrachten worden ontwikkeld door teams van vakinhoudelijke deskundigen. Daarbij kan het zo zijn dat er één persoon is die de itemspecificaties formuleert, terwijl anderen de items feitelijk schrijven. Recepten voor hoe itemschrijvers goede items kunnen maken bestaan er niet. De verwachting is dat als gevolg van de toegenomen mogelijkheden op automatiseringsgebied het ambachtelijke karakter van dit aspect van het constructieproces in de toekomst zal veranderen.

Fase 4: Toetsafname

We moeten bij toetsafname een onderscheid maken tussen een try-out of proefname en de definitieve toetsafname. Een proefafname is bedoeld om een indruk te krijgen van hoe de items inhoudelijk en psychometrisch functioneren bij de leerlingen waarvoor de definitieve toets bedoeld is. Op basis van de resultaten van de proefafname zullen sommige items verwijderd of gereviseerd worden. Na revisie zal er opnieuw een proefafname moeten plaatsvinden. Het aantal leerlingen waaraan de toets voorgelegd wordt, is bij een proefafname kleiner dan bij een definitieve toetsafname. Voor toetsen die voor onderzoeksdoeleinden gebruikt worden, bijvoorbeeld peilingsonderzoek, laat men om praktische redenen de proefafname soms achterwege en vindt er alleen een definitieve toetsafname plaats. Dit laatste betekent wel dat de toetsafname zeer goed voorbereid dient te worden.

Het is essentieel belang dat de toets onder gestandaardiseerde condities afgenomen wordt. Standaardisatie houdt in dat de toets door alle leerlingen onder gelijke omstandigheden uitgevoerd wordt. Alleen dan is het mogelijk de toetsprestaties van leerlingen met elkaar te vergelijken. Wanneer in dit boek over toetsen gesproken wordt, worden altijd gestandaardiseerde toetsen of meetinstrumenten bedoeld.

Fase 5: Itemevaluatie

Methoden voor het evalueren van items kunnen in twee categorieën verdeeld worden. De eerste categorie bestaat uit kwalitatieve methoden voor het evalueren van de inhoud van items. De Groot en van Naerssen (1973, p. 69) bespreken zes eisen waaraan gesloten vragen moeten voldoen. Gesloten vragen moeten objectief zijn, wat inhoudt dat verschillende vakdeskundigen hetzelfde alternatief als het juiste aanwijzen. Een andere eis is die van specificiteit. Een vraag is specifiek voor een bepaalde leerstof wanneer alleen leerlingen die de leerstof bestudeerd hebben de vraag kunnen oplossen. Kwantitatieve methoden voor het analyseren van antwoorden op items, bijvoorbeeld voor het bepalen van hoe moeilijk een item is, worden met name in de hoofdstukken 3, 4 en 5 behandeld.

Fase 6: Toetssamenstelling

Voor het kunnen selecteren van vragen is het nodig dat zowel kwalitatieve kenmerken, bijvoorbeeld leerstofcategorieën, als kwantitatieve kenmerken, bijvoorbeeld moeilijkheidsgraad, van de items bekend zijn. De mogelijkheden voor selectie worden uiteraard

bepaald door de omvang van de verzameling items. Wanneer de verzameling uit een groot aantal items bestaat die van kwalitatieve en kwantitatieve kenmerken voorzien zijn, spreekt men van een itembank. Itembanken zijn vaak onderdeel van een zogenaamd toetsservicesysteem, een geautomatiseerd stelsel van voorzieningen voor het opslaan, terugzoeken en selecteren van items, het samenstellen van toetsen en het analyseren van toetsresultaten. Methoden voor het selecteren van items gegeven de kenmerken waaraan de toets moet voldoen, worden in hoofdstuk 11 besproken.

Fase 7: Referentiekader

In deze fase wordt de wijze van rapporteren van de scores vastgesteld. De scores die op een toets behaald worden, hebben op zichzelf geen betekenis. De score die een leerling behaalt, krijgt pas betekenis wanneer die score vergeleken wordt met een bepaalde standaard of met de scores die andere leerlingen behaald hebben. De rapportage van scores wordt in hoofdstuk 13 behandeld.

Fase 8: Handleiding en verantwoording

Deze laatste fase bestaat uit het maken van handleiding en instructies voor de diverse categorieën personen die bij de toetsing betrokken zijn. Ten behoeve van de opdrachtgever en het wetenschappelijk forum dient een verantwoording geschreven te worden. In de eerder genoemde Richtlijnen en de Documentatie staan de eisen beschreven waarop toetsmateriaal, handleiding en verantwoording beoordeeld worden.

1.3 Het valideren van meetinstrumenten

Het hoofdstuk over validiteit in de Richtlijnen (1988), een vertaling van de Amerikaanse 'Standards for educational and psychological testing' (1985), nemen we als uitgangspunt voor onze bespreking van validiteit. Het hoofdstuk opent met "Bij de beoordeling van een test verdient de validiteit de meeste aandacht. Validiteit heeft te maken met de betekenis ('meaningfulness'), de bruikbaarheid ('usefulness') en de juistheid ('appropriateness') van de conclusies ('inferences') die uit testskores worden getrokken. Het valideren van een test is het verzamelen van gegevens met de bedoeling

na te gaan of deze conclusies juist zijn. Uit de testskores kunnen verschillende soorten conclusies worden getrokken en er bestaan veel manieren om informatie te verzamelen ter ondersteuning van elke gevolgtrekking. Validiteit is een overkoepelend begrip (’unitary concept’) dat in deze grote verscheidenheid struktuur aanbrengt. De gevolgtrekkingen (’consequences’) bij een specifieke toepassing worden gevalideerd, niet de test" (p. 11). Merk op dat we om de rest van deze paragraaf beter te kunnen begrijpen, bij een aantal begrippen de oorspronkelijke Engelse termen achter de Nederlandse vertaling vermeld hebben.

Over het inzicht dat in de laatste zin van het citaat staat en dat we te danken hebben aan Cronbach (1971, p. 447) bestaat algemeen consensus. Drenth en Sijtsma (1990) bijvoorbeeld omschrijven de validiteit van een test als "...de mate waarin de test aan zijn doel beantwoordt" (p. 173). Om het belang van dit inzicht nog eens te benadrukken geven we de omschrijving van De Groot en van Naerssen (1973): "De validiteitsvraag heeft altijd -bij definitie - betrekking op de mate waarin dat instrument beantwoordt aan het doel waarvoor het wordt gebruikt. Bij studietoetsen is dat doel in het algemeen: bepalen, ’meten’, van de stand van zaken van kennis en inzicht van leerlingen, op een bepaald gebied" (p. 30). Uit het voorgaande en de rest van het citaat uit de Richtlijnen kunnen we twee conclusies trekken.

De eerste conclusie is dat we niet kunnen spreken van de validiteit van een test, maar dat afhankelijk van het doel van de toets, de toets meer of minder valide kan zijn. De tweede conclusie is dat we voor het onderbouwen van de validiteit gegevens dienen te verzamelen. In de Richtlijnen worden drie manieren voor de onderbouwing van de validiteit van een toets onderscheiden: inhoudsvaliditeit, criteriumvaliditeit en begripsvaliditeit. In de Standards worden deze begrippen respectievelijk aangeduid met ’content-related’, ’criterion-related’ en ’construct-related evidence of validity’.

De belangrijkste theoretici op het gebied van validiteit, Cronbach (1971) en Messick (1989), zijn evenals de Richtlijnen van mening dat "Validiteit is een overkoepelend begrip dat in deze grote verscheidenheid struktuur aanbrengt", maar hebben kritiek op de wijze waarop de Richtlijnen daar vervolgens invulling aan geeft door drie soorten validiteit te onderscheiden. Aanleiding voor de kritiek was de toelichting bij de eerste richtlijn. Deze toelichting (Richtlijnen, 1988) luidt: "Het hangt van de aard van de vraagstelling, de kontext en de omvang van eerder verkregen bewijsmateriaal af of één of meer soorten validiteitsgegevens vereist zijn " (p. 19). De bezwaren van onder andere Messick (1988) vloeien voort uit zijn opvatting van validiteit die hij aldus verwoord heeft: "The heart of the unified view of validity is that appropriateness, meaningfulness, and usefulness of score-based inferences are inseparable and that the unifying force is empirically grounded construct interpretation. Thus from the

perspective of validity as a unified concept, all educational and psychological measurement should be construct-referenced because construct interpretation undergirds all score-based inferences - not just those related to interpretive meaningfulness but also the content- and criterion-related inferences specific to applied decisions and actions based on test scores. As a consequence, although construct-related evidence may not be the whole of validity, there can be no validity without it. That is, there is no way to judge responsibly the appropriateness, meaningfulness, and usefulness of score inferences in the absence of evidence to what the scores mean" (p. 35). Als gevolg van de toelichting bij de eerste richtlijn vreest Messick (1988) dat: "But the comment also leaves the door open for an interpretation that there exist circumstances under which one kind of validity evidence - be it content-related, for example, or criterion-related - may be adequate and fitting for an applied purpose" (p. 35).

Wat de Richtlijnen onder inhoudsvaliditeit en criteriumvaliditeit verstaan en waarom deze onvoldoende zijn voor het valideren van meetinstrumenten lichten we nu toe. Voor het onderbouwen van de inhoudsvaliditeit van een toets zijn volgens de Richtlijnen gegevens nodig die aantonen dat de steekproef van vragen waaruit de toets bestaat representatief is voor wat men wil toetsen. Zoals we eerder zagen was die onderbouwing bij de plaatsingstoetsen Engels gebaseerd op het oordeel van docenten. Een analyse van de inhoud alleen is volgens Shepard (1993, p. 414) echter onvoldoende om daarmee de validiteit van een toets te verdedigen, omdat er altijd onverwachte effecten zijn die de bedoelde relatie tussen testscore en het begrip of construct kunnen verstoren. Zij geeft een voorbeeld dat ontleend is aan onderzoek met betrekking tot plaatsingstoetsen. De inhoud van deze toetsen was gebaseerd op zorgvuldige curriculum specificaties. Empirisch onderzoek liet echter zien dat er aanzienlijke sexe-verschillen waren. De subtoetsen die uit meerkeuzevragen bestonden waren relatief gemakkelijker voor de mannen terwijl de subtoetsen die uit open vragen bestonden relatief gemakkelijker waren voor de vrouwen. Dit betekent dat onderdelen van de toetsen bij mannen een andere vaardigheid meten dan bij vrouwen en men moet zich dan ook de vraag stellen of de validiteit van die toets nog wel verdedigbaar is. Voornoemde opvatting van inhoudsvaliditeit wijkt nogal af van die van Ebel (1983) die van mening is dat inhoudsvaliditeit de enige validiteit is voor toetsen die na afloop van onderwijs of training afgenomen worden.

Voor het onderbouwen van de criteriumvaliditeit van een toets zijn volgens de Richtlijnen gegevens nodig die de samenhang aantonen tussen de testscores met een criterium. Criteriumvaliditeit is vooral belangrijk voor toetsen bedoeld voor selectie- en plaatsingsbeslissingen, omdat die beslissingen expliciet gebaseerd zijn op de relatie tussen de prestatie op de toets en de prestatie op het criterium. De criteriumvaliditeit

van bijvoorbeeld een plaatsingstoets moet dan ook onderbouwd worden door het aantonen van een empirische relatie tussen de scores op de plaatsingstoets en het succes van de plaatsingsbeslissingen. Afgezien van het feit dat het grootste probleem bij het onderzoek naar de criteriumvaliditeit van toetsen paradoxaal genoeg het ontbreken van valide criteria is, zijn empirische relaties met externe criteria noodzakelijk maar niet voldoende voor het onderbouwen van de validiteit van een toets (Shepard, 1993, p. 411). De hedendaagse opvatting van validiteit (= begripsvaliditeit), vereist dat niet alleen de relevantie en de integriteit van de criteriummaten geëvalueerd wordt, maar dat de voorspellingen zelf ook verdedigd worden. Toetsconstructeurs moeten kunnen verklaren waarom de toets voorspelt en waarom we op die relatie kunnen vertrouwen bij het nemen van beslissingen.

Voor het onderbouwen van de begripsvaliditeit zijn volgens de Richtlijnen gegevens nodig die de betekenis van de testscore duidelijk maken. Voor een toets tekstbegrip kan die onderbouwing bijvoorbeeld bestaan uit empirisch vastgestelde relaties met andere relevante meetinstrumenten, een zogenaamd nomologisch netwerk (Cronbach & Meehl, 1955), dat de betekenis of begripsvaliditeit van de toets duidelijk maakt. Dit is het geval wanneer de toets hoog correleert met soortgelijke toetsen (soortgenootvaliditeit) maar laag correleert met andere toetsen. Bij hoge correlaties spreken we van confirmerende validiteit en bij lage correlaties van discriminante validiteit.

Begripsvaliditeit kan op vele manieren (bijv. logische en empirische analyse, correlationeel en experimenteel onderzoek) en met vele analysetechnieken (bijv. multivariate analyse) onderzocht worden. Voor een overzicht van die manieren en technieken verwijzen we naar Messick (1989, p. 49 e.v.). Hier volstaan we met het noemen van twee analysetechnieken. De eerste is de multitrek-multimethode-benadering van Campbell en Fiske (1959). De tweede analysetechniek betreft psychometrische modellen waarmee de interne structuur of dimensionaliteit van toetsen onderzocht kan worden. In hoofdstuk 5 worden een aantal mogelijke modellen besproken.

Hoewel enerzijds iedereen de opvatting deelt dat bij een beoordeling van een test de validiteit de meeste aandacht verdient, moet anderzijds ook geconstateerd worden dat begripsvalidatie van toetsen op de manier zoals hiervoor en bij Shepard (1993, p. 432 e.v.) beschreven is, in de praktijk niet of nauwelijks voorkomt. Shepard (1993, p. 407) spreekt zelfs van een kloof tussen validiteitstheorie en toetspraktijk. Deze kloof is volgens Kane (1992) te wijten aan het ontbreken van praktische richtlijnen voor het valideren van toetsscores. Hij stelt de ’argument-based approach to validity’ voor en licht deze benadering toe met een plaatsingstoets wiskunde. Op deze benadering gaan we hier verder niet in.

Aan het eind van deze paragraaf willen we toelichten waarom in dit boek geen afzonderlijk hoofdstuk aan validiteit gewijd is. Zoals de bespreking van validiteit heeft laten zien, wordt onderzoek naar validiteit in het algemeen uitgevoerd met in de sociale wetenschappen algemeen bekende onderzoeksmethoden en analysetechnieken. Die methoden en technieken worden in vele uitstekende boeken meer uitgebreid behandeld dan in het kader van dit boek mogelijk geweest zou zijn. Van een behandeling van die methoden en technieken is dan ook afgezien. In dit boek beperkt validiteitsonderzoek zich tot onderzoek waarbij psychometrische technieken een rol spelen. Met name in de hoofdstukken 5 en 9 komen psychometrische modellen en technieken voor validiteitsonderzoek aan de orde.

1.4 Psychometrie in de praktijk

Het meest essentiële kenmerk van een toets als meetinstrument is dat het resultaat van de meting feilbaar is. De resultaten op toetsen zijn, zoals iedereen wel eens ervaren zal hebben, onderhevig aan allerlei toevalsfactoren. Een agglomeraat van toevalsfactoren in de condities waaronder getoetst wordt, in de persoon die getoetst wordt en ook in het meetinstrument zelf, maakt dat de metingen met toetsen nooit exact zullen kunnen zijn. Het zal ook duidelijk zijn dat de waarde van de informatie, die gebaseerd is op resultaten gemeten met deze instrumenten, en de rol die deze informatie kan spelen in het eerder beschreven toetsconstructieproces staat of valt met de nauwkeurigheid hiervan. Het aandachtsgebied van de psychometrie als toegepaste wetenschap is altijd geweest aan de gebruiker van meetinstrumenten de nauwkeurigheid van metingen zichtbaar te maken en die gebruiker methoden aan te bieden om de kwaliteit van meetinstrumenten te beoordelen. Vaardigheden die niet nauwkeurig gemeten worden, kunnen ook niet valide zijn. Dat wil niet zeggen dat nauwkeurige metingen ook valide metingen zijn. Meetnauwkeurigheid is een noodzakelijke maar geen voldoende voorwaarde voor validiteit.

Zoals we reeds eerder opmerkten richt de psychometrie zich op die aspecten van het toetsconstructieproces waarbij gebruik gemaakt wordt van empirische gegevens. In hoofdstuk 2 wordt een aantal algemene begrippen besproken die bij het verzamelen van deze gegevens een rol speelt. In de psychometrie bestaan die empirische gegevens in ieder geval uit kwantificeringen van kenmerken van personen die op zijn minst de aan- of afwezigheid van het kenmerk indiceren. Doorgaans zijn de te analyseren gegevens echter veel rijker. Bij toetsscores duidt de hoogte van de score op zijn minst ook de mate van aanwezigheid van het kenmerk van de persoon aan. De kenmerken die we

willen bestuderen, zijn doorgaans niet direct waarneembaar. De variabelen waarin we feitelijk geïnteresseerd zijn noemen we latent. De theorieën in de psychometrie leggen relaties tussen latente variabelen en geobserveerde variabelen. De rekenvaardigheid van een leerling kunnen we slechts proberen vast te stellen door de antwoorden op waarneembare indicatoren van dit kenmerk, bijvoorbeeld rekenopgaven, te beschouwen. De notie dat de observaties nooit een exacte weergave zullen zijn van de werkelijke aanwezigheid van een kenmerk, maakt dat psychometrische theorieën zich bedienen van formele beschrijvingssystemen die rekening houden met toevalsfactoren. De gebruikte modellen zijn dan ook probabilistische of stochastische modellen. De methoden en technieken die bij de ontwikkeling van modellen en bij het analyseren van gegevens worden gebruikt en die we in dit boek zullen beschrijven, maken deel uit van wat in de wiskunde bekend staat als de toegepaste statistiek.

De psychometrie bestond tot halverwege deze eeuw alleen uit de klassieke testtheorie. Een eerste volledige behandeling is te vinden in Gulliksen (1950). Een formeel volledige beschrijving en een aantal uitbreidingen vinden we in het boek van Lord en Novick (1968) dat nu nog steeds het standaardwerk van deze theorie is. Het uitgangspunt van de theorie is dat de geobserveerde score van een persoon op een toets de som is van een ware score, de waarde van een niet waarneembare variabele waarin we geïnteresseerd zijn, en een niet systematische, niet controleerbare meetfout. In de theorie worden deze begrippen preciezer gedefinieerd en veronderstellingen gedaan omtrent het stochastische karakter van de meetfout. In het werken met het klassieke testmodel hebben we uiteraard altijd te maken met toetsscores van meerdere personen, waarvan dan aangenomen wordt dat deze aselect getrokken zijn uit een of andere populatie. De statistiek die we in deze theorie gebruiken, generaliseert dan naar deze populatie van personen. Het primaire doel van de klassieke testtheorie is een beschrijving te geven van de nauwkeurigheid van de metingen. In de klassieke testtheorie staan daarvoor de begrippen betrouwbaarheid en standaardmeetfout centraal. Na Lord en Novick (1968) is de formele klassieke testtheorie nog nauwelijks uitgebreid. Ingegeven door de theoretisch enigszins magere fundering van het klassieke testmodel, maar ook door zijn inherente beperkingen en praktische problemen, kwam de moderne testtheorie, genaamd itemresponstheorie of latente trek theorie, tot ontwikkeling. Dat wil echter niet zeggen dat de klassieke testtheorie inmiddels volledig vervangen is door deze moderne theorie. De klassieke testtheorie heeft zoveel bruikbare methoden en technieken opgeleverd die kunnen bijdragen aan de kwaliteitsbeheersing van toetsen, dat met name in de tegenwoordige psychometrische praktijk nog veelvuldig gebruik gemaakt wordt van de klassieke testtheorie. Deze

theorie zal daarom in hoofdstuk 3 worden behandeld en ook in verschillende andere hoofdstukken ruime aandacht krijgen.

Alvorens in te gaan op de moderne testtheorie staan we even stil bij theorieën die we kunnen beschouwen als belangrijke uitbreidingen van de klassieke testtheorie. Op de eerste plaats is dat de generaliseerbaarheidstheorie (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). In tegenstelling tot de klassieke testtheorie kunnen in de generaliseerbaarheidstheorie verschillende foutenbronnen onderscheiden worden. De generaliseerbaarheidstheorie biedt dan ook de mogelijkheid verschillende 'betrouwbaarheden' te schatten. De theorie wordt in hoofdstuk 3 behandeld en in hoofdstuk 11 toegepast.

Andere uitbreidingen van de klassieke testtheorie zijn modellen waarbij er sterkere aannames over de meetfouten worden gedaan dan in het klassieke testmodel. Bekende modellen die met een gespecificeerde verdeling van de meetfouten werken zijn het binomiale-foutenmodel en het poisson-foutenmodel. Deze modellen die onder andere in Lord en Novick (1968) beschreven worden, zullen we in dit boek niet behandelen omdat de toepassing in de huidige psychometrische praktijk slechts incidenteel is.

In de moderne testtheorie met als startpunten Lord (1952) en Rasch (1960) wordt niet de score op een toets, samengesteld uit de scores op de items, gemodelleerd, maar wordt een expliciet model aangenomen voor de respons op elk afzonderlijk item. De kans dat een persoon een bepaalde respons op een item geeft, is een gespecificeerde functie van de te meten latente variabele van de persoon, de vaardigheidsparameter, en één of meerdere itemparameters. De itemresponstheorie heeft veel van de bezwaren van de klassieke testtheorie weggenomen. In de itemresponstheorie bestaat, in tegenstelling tot de klassieke testtheorie, de mogelijkheid de geldigheid van het aangenomen model expliciet te toetsen. Daarnaast zijn de itemkarakteristieken onafhankelijk van de specifieke toets waarin de items zitten. Bovendien levert de theorie methoden en technieken die nieuwe toepassingen van de psychometrie mogelijk maken. Was de klassieke testtheorie volledig geconcentreerd op het resultaat van de meting, in de itemresponstheorie zijn er veel meer mogelijkheden om te onderzoeken hoe dit resultaat tot stand is gekomen.

De toepassingsmogelijkheden van de eerste itemresponsmodellen zijn beperkt. Het zijn modellen die uitgaan van dichotoom gescoorde items en die zulke strenge eisen aan de responsen opleggen, dat in veel praktijkgevallen het model als ongeldig moest worden verklaard. Heden ten dage echter zijn de modellen op allerlei manieren uitgebreid. Er zijn modellen met meer itemparameters en de beperking tot dichotoom gescoorde items is vervallen. Daar komt bij dat de analyses in de itemresponstheorie hogere statistische en rekentechnische eisen stellen dan de analyses in de klassieke

testtheorie. Pas na enkele decennia werk van een groot aantal psychometrici en door de enorme ontwikkelingen op computergebied, heeft de itemresponstheorie ook een zeer belangrijke plaats in de psychometrische praktijk gekregen. Een verschuiving van wat Van der Linden (1983) noemt het klassieke complex, het werken met gestandaardiseerde toetsen en de klassieke testtheorie, naar het moderne complex, het werken met itembanken en itemresponstheorie, is waar te nemen.

In hoofdstuk 4 zal een uitvoerige inleiding worden gegeven in de basisconcepten en de schattings- en toetsingsmethoden in de itemresponstheorie. Dit zal worden besproken aan de hand van het model van Rasch (1960). In hoofdstuk 5 wordt een overzicht gegeven van uitbreidingen van het Raschmodel en andere itemresponsmodellen. Aparte aandacht krijgt, met name vanwege het grote belang voor de praktijk, de itemresponstheorie in zogenaamde onvolledige gegevensverzamelingen. Enkele concrete toepassingen van itemresponstheorie worden in hoofdstuk 7 behandeld.

Omdat toetsen vaak gebruikt worden om beslissingen te nemen over personen kan een besliskundige benadering van de psychometrie ook zeer vruchtbaar zijn. Wij zullen om praktische redenen deze benadering niet expliciet behandelen. Voor een overzicht van de besliskundige testtheorie verwijzen wij naar Van der Linden (1985).

In hoofdstuk 8 tot en met 10 worden problemen uit de praktijk besproken die met behulp van de itemresponstheorie worden opgelost. Achtereenvolgens komen daarbij de volgende onderwerpen aan de orde: het equivaleren van toetsen, vraagonzuiverheid en het meten van veranderingen. Hierbij worden, evenals in het volgende hoofdstuk, zowel oplossingen met behulp van de klassieke testtheorie als de itemresponstheorie besproken. Hoofdstuk gaat over het samenstellen van optimale toetsen met behulp van mathematische programmering. De beoordeling van niet zonder meer objectief scoorbare toetsen of opdrachten is het onderwerp van hoofdstuk 12. Zoals elk toetsconstructieproces, en trouwens ook elke toets, wordt dit boek afgesloten met een behandeling van de rapportage van de toetsresultaten.

2

Dataverzameling

We verzamelen gegevens omdat we iets te weten willen komen. We willen bijvoorbeeld weten of kinderen kunnen optellen en welke begrippen ze beheersen. Soms willen we iets weten van een individu, soms van een bepaalde groep individuen, bijvoorbeeld van een etnische minder- heid. We kunnen individuen onderling vergelijken of hen stuk voor stuk vergelijken met een norm. Dikwijls zijn we niet in de eerste plaats geïnteresseerd in een vergelijking van individuen, maar in een vergelijking van vragen en opgaven. Dan kunnen we ons afvragen of de ene opgave moeilijker is dan de andere, maar ook of vragen bepaalde gewenste eigenschappen hebben. Om dergelijke vragen te beantwoorden, is het meestal nodig op systematische wijze gegevens te verzamelen en data te analyseren.

In dit hoofdstuk komen begrippen ter sprake die in de volgende hoofdstukken worden gebruikt. In paragraaf 2.1 wordt beschreven op welke wijze men van waarnemingen tot data komt. De nadruk ligt er op dat waarnemingen op zichzelf beschouwd niets zeggen, maar dat zij geïnterpreteerd moeten worden. Aansluitend hierop worden er in paragraaf 2.2 diverse schaalniveaus behandeld. We gaan er van uit dat waarnemingen worden gecodeerd in getallen; men noemt dit wel het scoren van de waarnemingen. Schaalniveaus hebben te maken met de eigenschappen die men aan de gebruikte scores kan toekennen. Dat men zich in de praktijk vaak gemakkelijk schikt in assumpties over schaalniveaus, en dat men dit vaak zonder bezwaar kan doen, wordt uiteengezet in paragraaf 2.3. In paragraaf 2.4 komen enige algemene procedures voor het verzamelen van data aan de orde. Twee belangrijke begrippen die bij zulke procedures behoren, zijn betrouwbaarheid en validiteit; zij worden kort behandeld in paragraaf 2.5. In paragraaf 2.6 bespreken we het gebruik van steekproeven van personen. In paragraaf 2.7 gaan we in op het gebruik van proefopzetten; dat zijn procedures om stimuli over personen te verdelen. In paragraaf 2.8 bespreken we de soorten stimuli die voorkomen in de psychometrie, en in paragraaf 2.9 het gebruik van meetmodellen.

2.1 Van waarnemingen tot data

We observeren in het algemeen het gedrag van personen. We beperken ons hier tot het gedrag dat personen vertonen op vragen en opgaven: het gaat om de antwoorden die de personen geven en om de wijze waarop zij een taak volbrengen. Het is van groot belang, vast te stellen dat we observaties nog geen data noemen. Pas als we een interpretatie aan de observaties geven, spreken we van data. Zoals Bezembinder (1970, p. 41) het uitdrukt: "Data zijn relaties tussen objecten, en deze relaties zijn interpretaties van observaties. Kale, niet-geïnterpreteerde observaties, bestaan niet. Maagdelijke data evenmin. De onschuldige observatie is een fictie." Een goed voorbeeld hiervan is te vinden in een artikel van Lord (1953). Een professor geniet het voorrecht, de rugnummers te mogen uitdelen aan de spelers in het rugbyteam. De eerstejaars-studenten beklagen zich: zij zouden wel erg veel lage nummers hebben gekregen. De professor verweert zich tegen de aanklacht door er op te wijzen dat rugnummers slechts etiketten zijn: zij houden slechts de spelers uit elkaar, en de nummers hadden ook letters en plaatjes mogen zijn. Als getuige à charge treedt de statisticus van de universiteit op. Deze voert blijmoedig een t-toets uit voor twee groepen, en stelt vast dat de klagers gelijk hebben. Aan de mededeling dat de rugnummers slechts etiketten zijn, heeft hij geen boodschap: "Die nummers weten immers niet waar zij vandaan komen". We zien dat de studenten de rug- nummers interpreteren als kwalificaties: die rugnummers zouden een ordening in de spelers aanbrengen. De professor ziet de rugnummers als naamkaartjes en hecht geen betekenis aan de numerieke eigenschappen van de rugnummers. De crux van het verhaal is natuurlijk de rol van de statisticus: kan hij wel rugnummers van spelers middelen en hun spreiding bepalen? "Natuurlijk kan ik dat; ik heb het toch zojuist gedaan?" antwoordt de statisticus in het verhaal.

2.2 Schaalniveaus

Het probleem dat is verwoord in het zojuist geparafraseerde artikel van Lord, betreft de toelaatbaarheid van rekenkundige operaties op in getallen weergegeven observaties. Men spreekt wel van het probleem van het schaalniveau. We gaan er van uit dat alle observaties op de een of andere manier zijn omgezet in getallen. Een schaal is een verzameling getallen en tussen die getallen gedefinieerde relaties die een empirische interpretatie hebben. De aan waarnemingen toe- gekende scores zijn getallen die tot

een schaal behoren. Door de met de schaal gegeven empirische interpretatie kan men op grond van de scores empirische uitspraken over de waarnemingen doen. Scores worden geacht van een bepaald schaalniveau te zijn als zij bepaalde transformaties kunnen ondergaan zonder dat de interpretatie van de getallen verandert. Men kan met scores rekenen; het gaat er om vast te stellen welke rekenkundige bewerkingen tot resultaten leiden die geïnterpreteerd kunnen worden in termen van de oorspronkelijke waarnemingen. Hoewel het aantal te onderscheiden schaalniveaus in beginsel heel erg groot is, maakt men doorgaans alleen maar onderscheid in de volgende vijf schaalniveaus: nominaal, ordinaal, interval-, ratio- en absoluut schaalniveau. Deze schaalniveaus zijn opgesomd in volgorde van afnemende vrijheid. Elk volgend schaalniveau in de opsomming laat minder manipulaties met scores toe, maar verschaft meer informatie.

Het nominale schaalniveau biedt de onderzoeker grote vrijheid in het manipuleren van scores. De aan observaties toegekende getallen mogen worden vervangen door willekeurige andere getallen mits men zich aan de volgende beperking houdt: aan observaties waaraan gelijke respectievelijk verschillende getallen zijn toegekend, worden na de transformatie wederom gelijke respectievelijk verschillende getallen toegekend. De getallen dienen er slechts toe, als gelijk beschouwde observaties dezelfde scores te geven en als verschillend beschouwde observaties verschillende scores te geven. Daaruit blijkt dat de scores weinig informatie verschaffen. Zij geven slechts aan welke observaties men als gelijk respectievelijk verschillend beschouwt. Het is niet mogelijk te spreken over de mate waarin observaties verschillen. De toegekende getallen fungeren slechts als etiketten of namen; hieraan ontleent het besproken schaalniveau zijn naam. Het is van belang er op te wijzen dat de onderzoeker uiteindelijk bepaalt van welk schaalniveau hij zijn observaties acht. De professor uit het artikel van Lord beschouwt de rugnummers van de studenten als observaties van nominaal niveau: de rugnummers dienen er slechts toe de studenten uit elkaar te houden. In zijn ogen heeft het dan ook geen zin het gemiddelde rugnummer te berekenen: dat getal betekent even weinig als de gemiddelde naam. De studenten in het artikel van Lord zijn een duidelijk andere mening toegedaan. Zij beschouwen de rugnummers als een aanduiding van een ordening onder de studenten. Aan de klagers zouden wel erg veel lage nummers zijn toebedeeld. Die klagers vatten de rugnummers op als van, op zijn minst, ordinaal schaalniveau.

Aan observaties toegekende getallen of scores worden geacht van ordinaal schaalniveau te zijn als zij de een of andere ordening in de observaties weerspiegelen. Zulke getallen mogen worden vervangen door willekeurige andere getallen mits de ordening intact blijft. Dit wordt wiskundig uitgedrukt met de zegswijze dat men op

getallen van ordinaal schaalniveau willekeurige monotone transformaties mag uitvoeren. Voor observaties die geacht worden gemeten te zijn op ordinaal niveau heeft alleen de ordening betekenis. Men kan de observaties bijvoorbeeld onderling vergelijken in termen van groter of mooier; het is echter niet mogelijk te zeggen hoeveel groter of hoeveel mooier de ene observatie is dan de andere.

Men noemt aan observaties toegekende getallen van intervalschaalniveau als men betekenis kan hechten aan verschillen tussen dergelijke getallen. Een bekend voorbeeld van getallen die van intervalniveau zijn, is gegeven door de gangbare schalen voor temperatuur. Een voorwerp heeft een bepaalde temperatuur. Deze temperatuur kan men uitdrukken in graden Celsius maar ook in graden Fahrenheit. Voor dezelfde waarneming heeft men dus twee getallen: dezelfde waarneming is op twee manieren gescoord. De twee getallen kan men tot elkaar herleiden door er een lineaire transformatie op toe te passen. Een lineaire transformatie van x naar y schrijft men als: $y = ax+b$, waarin de getallen a en b willekeurige getallen zijn en a niet gelijk is aan nul. Doordat men zowel a als b vrij kan kiezen, zegt men wel dat men de oorsprong en de eenheid van de schaal vrij kan kiezen.

We illustreren het intervalschaalniveau aan het gebruik van de schalen voor het meten van temperatuur. Als men een bepaalde temperatuur kan beschrijven als x graden Celsius en ook als y graden Fahrenheit, dan bestaat er tussen de getallen x en y de volgende betrekking: $y = 1.8x+32$. Het is van belang er op te wijzen dat bij een lineaire transformatie de verhouding van twee verschillen constant blijft. Zij x het verschil tussen twee op de Celsius- schaal gemeten temperaturen x_1 en x_2, en x' het verschil tussen twee temperaturen x_3 en x_4. Zij de verhouding van de twee verschillen in temperatuur x en x' op de Celsiusschaal gelijk aan r: $r = x/x'$. Als men nu zowel x als x' transformeert naar de Fahrenheitschaal, krijgt men twee getallen y en y'. Daarvoor geldt dat $y = (1.8x_1+32)-(1.8x_2+32) = 1.8(x_1-x_2) = 1.8x$, en $y' = 1.8x'$. De verhouding r' van y en y' is dan gelijk aan x/x', en dus gelijk aan r. Voor getallen die geacht worden van intervalschaalniveau te zijn en dus alleen aan een lineaire transformatie onderworpen mogen worden, blijkt dat verhoudingen van verschillen onder dergelijke transformaties niet veranderen.

Men acht getallen die aan observaties worden toegekend van ratioschaalniveau, als men die getallen aan transformaties kan onderwerpen die de verhoudingen van getallen onverlet laten. De enige transformaties met deze eigenschap zijn de multiplicatieve transformaties: $y= ax$ voor een willekeurig getal a dat niet gelijk is aan nul. Een voorbeeld van meten op ratioschaalniveau is het meten van lengte. Men kan de lengte van een voorwerp uitdrukken in centimeters en in inches; maar ongeacht de keuze van de eenheid kent men het getal 0 toe aan een voorwerp dat 'geen lengte heeft'. De

meting 0 verandert niet door een multiplicatieve transformatie. Aangezien men alleen de schaalfactor *a* vrij kan kiezen, zegt men wel dat bij een ratioschaal alleen de eenheid vrij gekozen kan worden. Merk op dat verschillen tussen getallen die van intervalschaalniveau zijn, zelf van ratioschaalniveau zijn.

Men acht getallen van absoluut schaalniveau te zijn als er geen transformatie is toegestaan. Wiskundigen zeggen in zo'n geval dat alleen de identiteitstransformatie is toegestaan: elk getal kan alleen maar 'in zichzelf worden getransformeerd'. Van absoluut schaalniveau acht men bijvoorbeeld getallen die een aantal aanduiden. Zoals Bezembinder (1970, p. 73) het uitdrukt: "Een even robuust als rustiek voorbeeld van het gebruik van een absolute schaal levert ons de herder die zijn schaapjes telt".

2.3 Meten per fiat

Het is van belang er op te wijzen dat het toekennen van een schaalniveau aan getallen een activiteit is van de onderzoeker; getallen hebben niet van zichzelf enig schaalniveau. Het onderbrengen van getallen in een bepaald soort schaal is een kwestie van interpretatie. Het is vaak niet eenvoudig, vast te stellen van welk schaalniveau scores zijn. Als de herder dat zou willen, kan hij schapepoten tellen in plaats van schapen: voor hem zijn aantallen kennelijk van ratioschaalniveau. Maar dan moet hij natuurlijk geen schaap met vijf poten in zijn kudde hebben.

In de praktijk houdt men zich niet altijd intensief bezig met de vraag, van welk schaalniveau de verkregen observaties zijn. Dikwijls analyseert men data met methoden die eigenlijk getallen van intervalschaalniveau vereisen zonder dat men heeft onderzocht of zo'n assumptie gerechtvaardigd is. Uit de zinvolheid van de verkregen resultaten leidt men dan alsnog af dat de assumptie gerechtvaardigd is. Veel meetprocedures berusten op vaste af- spraken: men is het er over eens bepaalde zaken op een bepaalde manier te onderzoeken en te analyseren. Daarom spreekt men wel van meten per 'fiat'.

2.4 Procedures voor dataverzameling

De wijze waarop men gegevens verzamelt, en ook de beslissing welke gegevens te verzamelen, hangen af van een groot aantal factoren. Voor een deel zijn deze factoren bepaald door de theorie die men aanhangt, en voor een ander deel door statistische en economische overwegingen. Voor elk onderzoek is nu eenmaal een beperkt budget

beschik- baar en dat moet zo goed mogelijk worden gebruikt. Uit deze overwegingen vloeit voort dat men in elk geval op systematische wijze gegevens moet verzamelen: men zal een welomschreven procedure moeten volgen. Er zijn vele procedures om observaties te verzamelen. Deze procedures kunnen op een aantal manieren worden ingedeeld. De volgende classificaties van procedures voor het verzamelen van gegevens zijn ontleend aan Meerling (1981).

Men kan in de eerste plaats het onderscheid maken tussen directe observatie enerzijds en observatie door middel van een instrument anderzijds. Bij directe observatie nemen we het gedrag van een persoon waar en interpreteren dit gedrag direct bij waarneming. Denk bijvoorbeeld aan het observeren van het gedrag van spelende kinderen. Was die klap nu een goedmoedige por of een echte klap? Bij observatie door een instrument wordt het gedrag van een persoon geobserveerd op een stimulus die door de onderzoeker wordt aangeboden. Het gaat nu om uitgelokt gedrag. Denk aan het antwoord van leerlingen op items in een toets die optelvaardigheid meet of aan een enquête waarin gevraagd wordt naar stemgedrag.

In de tweede plaats kan men procedures onderscheiden naar de bron die de gegevens verschaft. Soms is het de onderzoeker zelf die waarneemt en dan selecteert en interpreteert, zoals de ontdekkingsreiziger in het oerwoud. Maar ook kan het de onderzochte persoon zijn, zoals de bekende Nederlander die de interviewer niet het achterste van zijn tong laat zien. Ook kan het zijn dat de observatie komt van een derde persoon, bijvoorbeeld een on- afhankelijke beoordelaar. Andere bronnen van gegevens zijn dossiers en archieven. Men maakt dan gebruik van gegevens die door anderen op een eerder tijdstip zijn vastgelegd.

In de derde plaats kan men procedures voor het verzamelen van gegevens onderscheiden naar de tegenstelling reactief en niet-reactief. Reactief noemt men de observatieprocedure die het normale gedragspatroon van de proefpersoon verstoort. Men kan hierbij denken aan experimentele behandelingen en in het algemeen aan uitgelokt gedrag. Niet-reactief noemt men procedures waarbij er geen gedrag wordt uitgelokt maar er louter wordt gekeken.

2.5 Betrouwbaarheid en validiteit

Als we het in dit boek hebben over data, hebben we het meestal over antwoorden van personen op items of uitvoeringen van opdrachten. Door deze items of opdrachten, al dan niet gebundeld in een toets, aan personen voor te leggen, hopen we iets te weten te komen over de personen en dikwijls ook over de items en de opdrachten. We

veronderstellen dat de items en de opdrachten operationalisaties zijn van het te onderzoeken gedrag. Het zijn concrete, duidelijk afgebakende stimuli die te zamen alle uitingsvormen bevatten van het te bestuderen gedrag. In hoofdstuk 3 wordt, in het deel over de generaliseerbaarheidstheorie, ingegaan op het idee van alle uitingsvormen van het te bestuderen gedrag. We interpreteren het geobser- veerde gedrag: als we optelitems voorleggen aan een leerling gaan we er van uit dat de antwoorden die de leerling geeft, ons iets zeggen over de optelvaardigheid van die leerling.

We beperken ons tot observaties door een instrument. We willen een interpretatie kunnen geven aan de observaties die verkregen worden door het voorleggen van een stimulus aan een persoon. Het gaat daarbij meestal om gedrag dat we niet direct kunnen observeren; we nemen uitingen van gedrag waar die we interpreteren als manifestaties van niet direct waar te nemen eigenschappen en vaardigheden. Zulke eigenschappen en vaardigheden noemt men wel latente variabelen. Zij zijn begrippen die in een theorie worden gepostuleerd en gedefinieerd.

Bij elke procedure voor het vergaren van data zijn twee begrippen van belang. In de eerste plaats is het belangrijk te weten wat we meten; dit is de vraag naar de validiteit van de procedure en van het instrument. Het afnemen van een instrument moet leiden tot een interpreteerbare observatie van het gedrag van de leerling op de vragen en de opdrachten. De geïnterpreteerde reactie geeft binnen het kader van de theorie aan, welke conclusies we kunnen trekken. Als we een leerling een optelopgave geven, interpreteren we een goed antwoord als: de leerling beschikt over voldoende optelvaardigheid om het in de opgave weergegeven probleem op te lossen.

In de tweede plaats is het belangrijk dat we een zo nauwkeurig mogelijke observatie hebben; dit is de vraag naar de betrouwbaarheid van de procedure en het instrument. Indien we een meting zouden kunnen herhalen onder identieke omstandigheden zouden we dezelfde meting moeten krijgen. Er zullen in praktijk echter altijd verstorende invloeden gelden. Zo is de eis van identieke omstandigheden meestal niet te vervullen: het aanbieden van een item zou al een leereffect kunnen hebben.

In de psychometrie besteden we aandacht aan personen, aan stimuli en aan de reacties van personen op stimuli. Analyse van de data moet antwoord geven op de gestelde onderzoeksvragen. Het moet dan mogelijk zijn individuen en groepen individuen met elkaar te vergelijken, en ook stimuli en groepen stimuli. We kunnen vaststellen dat de ene leerling beter kan optellen dan een andere, en dat de ene groep beter kan optellen dan een andere. Stimuli, bijvoorbeeld items, kunnen met elkaar worden vergeleken: het ene item is moeilijker dan het andere.

Dikwijls wil men het gedrag van een enkel persoon bestuderen. Voorbeelden daarvan zijn te vinden in de psychodiagnostiek en in het gebruik van toetsen voor het meten van

vorderingen op school. Maar even zo vaak stelt men geen belang in het individu. Zo tracht de psychonomie algemeen geldende wetten te vinden die psychologische functies beschrijven: hoe ziet een oog, hoe grijpt een hand. En in het onderwijs wil men vaak groepen personen op hun prestaties in een vak onderscheiden. Een belangrijk gebied waar groepen personen een rol spelen, is dat van het ontwikkelen van meetinstrumenten. Als een psycholoog de van een persoon verkregen responsen op een meetinstrument wil kunnen interpreteren, moet hij er staat op kunnen maken dat het instrument de tussen personen bestaande verschillen kan blootleggen. En als een leraar de vorderingen van een bepaalde leerling in de tijd wil kunnen volgen, moet hij er op kunnen rekenen dat het gebruikte instrument in staat is, werkelijk opgetreden veranderingen vast te stellen. Hier is de betrouwbaarheid van het instrument in het geding. De klassieke testtheorie, die in hoofdstuk 3 wordt behandeld, is een meettheorie waarin een kwantitatief begrip betrouwbaarheid is gedefinieerd. Om deze maat te schatten, heeft men waarnemingen nodig van groepen personen. Veel psychometrie houdt zich dan ook bezig met groepen personen. Daarbij komt men voor het probleem te staan dat men in een onderzoek veelal niet alle personen kan betrekken waar men iets over te weten wil komen. Men zal dan zijn toevlucht moeten nemen tot het trekken van steekproeven van personen. Een vergelijkbaar probleem, zeker bij het ontwikkelen van meetinstrumenten, is dat men vaak beschikt over veel kandidaatstimuli waarvan men de eigenschappen wil leren kennen; men kan echter niet alle stimuli aan elk der personen voorleggen. Men zal dan zijn toevlucht moeten nemen tot procedures om stimuli aan personen toe te wijzen. De combinatie van het trekken van steekproeven van personen en het verdelen van stimuli over de personen heet een proefopzet.

2.6 Steekproeven

Een steekproef van personen is een selectie van personen uit een duidelijk omschreven groep personen waar men belang in stelt. Deze laatste groep heet populatie, en dient zo gedefinieerd te zijn dat men van elke persoon kan vaststellen of hij tot de populatie behoort. Voorbeelden van populaties zijn: alle mensen met een leeftijd tussen vijftien en vijfenzestig jaar, en alle leerlingen uit groep acht van de basisschool in Nederland. Uit de voorbeelden blijkt dat het niet eenvoudig is een populatie te definiëren. Het zal immers vaak voorkomen dat een persoon slechts gedurende een beperkte tijd deel uitmaakt van een populatie. Wie de basisschool verlaat, verlaat tevens de zojuist als voorbeeld gegeven populatie. Men maakt daarom wel onderscheid tussen twee soorten

populaties: de doelpopulatie en de bemonsterde populatie. De bemonsterde populatie wordt ook wel aangeduid als het steekproefkader. De doelpopulatie is niet de groep maar de soort personen waar men belang in stelt. De bemonsterde populatie is de groep personen waar men een steekproef uit trekt. Bij de gegeven voorbeelden van doelpopulaties kan men de volgende bemonsterde populaties definiëren: alle mensen in Nederland die op 1 januari 1980 een leeftijd hebben tussen vijftien en vijfenzestig jaar, en alle leerlingen in Nederland die op 15 september 1990 in groep acht van de basisschool zitten. De statistiek verschaft de middelen om uit gegevens van een steekproef kansuitspraken te doen over eigenschappen van de bemonsterde populatie. In hoeverre men uit deze uitspraken iets kan concluderen over de doelpopulatie, is niet louter een kwestie van statistiek. Daarbij zijn kennis, ervaring en theoretische inzichten onontbeerlijk (Cornfield & Tukey, 1956). Voor het maken van generalisaties zijn twee statistische begrippen van belang: de representativiteit van een steekproef en de nauwkeurigheid van op steekproeven gebaseerde schattingen van kenmerken van de populatie. In het vervolg beperken wij ons tot het trekken van steekproeven uit de bemonsterde populatie, die we kortheidshalve populatie zullen noemen.

2.6.1 Representativiteit van steekproeven

Een noodzakelijke voorwaarde voor het op valide wijze kunnen generaliseren van de waarnemingen in een steekproef naar eigenschappen van een populatie, is dat de steekproef representatief is voor de populatie. De steekproef dient een goede weergave te zijn van de populatie. In beginsel kan men zich het begrip representativiteit als volgt voorstellen. De personen die deel uitmaken van de populatie kunnen op een veelheid van kenmerken worden onderscheiden. Deze kenmerken hebben een gezamenlijke verdeling in de populatie. Dezelfde verdeling van de kenmerken wil men graag terugzien in de steekproef. Als men, bijvoorbeeld, een algemene schets wil geven van de praktijk van een huisarts in Nederland, kan men niet volstaan met een steekproef van huisartsen uit Amsterdam. Daarmee kan men ten hoogste een beschrijving maken van de praktijk van een huisarts in een grote stad.

In de praktijk is het niet goed mogelijk, alle kenmerken van een populatie in beschouwing te nemen. In de eerste plaats kent men niet alle mogelijke kenmerken van een populatie. En in de tweede plaats acht men bepaalde eigenschappen niet van belang voor het onderzoek. Zo kan men zich voorstellen dat het er niet toe doet welke

kleur de auto van een huisarts heeft. Evenzo kan men zich voorstellen dat de omvang van een praktijk wel een belangrijk kenmerk is. Als men een kenmerk van een populatie in een onderzoek betrekt, kan blijken dat het kenmerk niet van belang is voor de onderzoeksvraag. In dat geval kan men vaak het bij de analyse van de gegevens gehanteerde model vereenvoudigen. Ernstiger is het buiten beschouwing laten van een kenmerk dat wel van belang is. In dit geval spreekt men van een specificatiefout. Specificatiefouten kunnen leiden tot verkeerde conclusies. Men zal zich bij het kiezen van de in een onderzoek te betrekken kenmerken van een populatie moeten laten leiden door een theorie. Men beperkt zich bij het vaststellen van de representativiteit van een steekproef tot de eigenschappen van een populatie die op grond van theoretische kennis van belang worden geacht voor het onderzoek.

2.6.2 Nauwkeurigheid

Veelal zal men op grond van een steekproef een schatting maken van een kwantitatief kenmerk van een populatie. Zo'n kenmerk noemt men een parameter van de populatie. De uit de steekproef berekende grootheid wordt een schatting van de parameter genoemd. Het voorschrift waarmee uit gegevens van een steekproef een schatting van een parameter wordt berekend, noemt men een schattingsvoorschrift of kortweg een schatter. Nu kan men vaak uit een populatie op veel manieren een representatieve steekproef trekken. Men zal dan ook, bij het gebruik van steeds dezelfde schatter, bij elke steekproef een andere schatting van de parameter kunnen vinden. Het is te hopen dat deze verschillende schattingen niet teveel uiteenlopen. Een maat voor de variatie in de schattingen is de standaardafwijking van alle mogelijke schattingen. Deze standaardafwijking heet de standaardfout van de gebruikte schatter. Bij elke schatting die wordt gerapporteerd, behoort de standaardfout vermeld te worden. Het behoeft geen betoog dat een standaardfout niet zonder meer beschikbaar is; immers, om hem te berekenen zou men moeten beschikken over alle mogelijke steekproeven. Veel standaardfouten worden dan ook geschat met behulp van hulpmiddelen uit de mathematische statistiek en de kansrekening. De statistiek leert dat veel standaardfouten omgekeerd evenredig zijn met de wortel van het aantal personen in de steekproef. Om een standaardfout te halveren, moet men dan ook in het algemeen een vier keer zo grote steekproef trekken.

2.6.3 Aselecte steekproeven

De eenvoudigste steekproef is de aselecte steekproef. Zo'n steekproef ter grootte n bestaat uit n personen uit de bemonsterde populatie. Men kan op veel manieren zo'n steekproef samenstellen; dat wil zeggen dat men allerlei n-tallen uit de populatie kan kiezen. Als elk van die n-tallen dezelfde kans heeft om getrokken te worden, spreekt men van het trekken van een aselecte steekproef ter grootte n. Aan de hand van statistische en economische criteria kan men de vereiste omvang van de steekproef bepalen. Zulke criteria zijn bijvoorbeeld: de kans op onjuiste uitspraken en de kosten van het vergaren van responsen. De aselecte steekproef is om veel redenen aantrekkelijk. Zo is de kans groot dat de steekproef een goede representatie biedt van de populatie. Als, bijvoorbeeld, een populatie voor de helft uit vrouwen bestaat, dan is de kans erg klein om bij aselect getrokken steekproeven een steekproef te verkrijgen met louter vrouwen er in. Van belang is dat het bepalen van schatters en standaardfouten bij aselecte steekproeven doorgaans redelijk eenvoudig is.

Aan de aselecte steekproef kleven echter wel enige bezwaren. Het voornaamste bezwaar is dat er geen rekening wordt gehouden met heterogeniteit in de populatie. De populatie bestaat dikwijls uit deelgroepen personen die onderling meer op elkaar lijken dan personen uit verschillende deelgroepen. Aan het verschijnsel van homogeniteit van deelgroepen wordt aandacht geschonken in paragraaf 2.6.6. Als er sprake is van homogene deelgroepen, kan men gebruik maken van een gestratificeerde steekproef.

2.6.4 Gestratificeerde steekproeven

Men maakt gebruik van gestratificeerde steekproeven als men onderkent dat de populatie bestaat uit deelgroepen die in veel opzichten van elkaar verschillen. Vaak wil men, naast uitspraken over de gehele populatie, uitspraken doen over deze deelgroepen. Die deelgroepen, strata genoemd, kunnen zoveel verschillen dat men elk stratum op een aparte manier moet benaderen. Zo maakt men bij bevolkingsonderzoeken vaak onderscheid tussen de strata urbaan of stedelijk enerzijds en ruraal of landelijk anderzijds. Niet alleen leven personen in beide strata op verschillende wijze, ook brengt elk stratum zijn eigen wijze van onderzoeken met zich mee. Te denken valt aan de verschillen in afstand en reistijd tussen twee personen in de stad en die tussen twee personen op het land. De aselecte steekproeftrekking beschouwt personen als de eenheden waarvan men een steekproef trekt. De gestratificeerde steekproef- trekking bestaat uit het trekken van een steekproef uit elk der strata.

Dikwijls is het om administratieve en logistieke redenen niet mogelijk steekproeven van personen te trekken. Zo komt het vaak voor dat men wel beschikt over een lijst met adressen van gemeenschappen maar niet over adressen van personen. Bij gemeenschappen kan men denken aan huishoudens en scholen. In zo'n geval trekt men een aselecte steekproef van gemeenschappen en onderzoekt dan alle in een gemeenschap aangetroffen personen, of trekt weer een steekproef van personen uit elke gemeenschap. In het laatste geval spreekt men van getrapte steekproeftrekking.

2.6.5 Getrapte steekproeven

Als men een bevolkingsonderzoek wil doen in een omvangrijke regio, verdeelt men vaak de regio in deelgebieden en trekt dan een steekproef van deelgebieden. De deelgebieden vormen nu de eenheden van de steekproef. Deelgebieden worden doorgaans 'clusters' genoemd. Alle personen uit een deelgebied of cluster worden onderzocht, of een steekproef van personen. De onderzoekers kunnen een deelgebied in een keer bezoeken, wat reistijd en kosten bespaart. Ook kan men denken aan leerlingen die gegroepeerd zijn in klassen en klassen die weer gegroepeerd zijn in scholen. Leerlingen uit dezelfde klas lijken in veel opzichten op elkaar omdat ze in dezelfde omstandigheden verkeren. Als men de reacties van een leerling op een instrument kent, kan men vaak al een redelijk goede voorspelling maken van de reacties van de klasgenoten. Men zou dan ook kunnen volstaan met het trekken van een steekproef uit elke klas. Om logistieke redenen is dat vaak niet mogelijk. Een school stelt bijvoorbeeld een lesuur en een gehele klas ter beschikking; dan is het niet praktisch om een steekproef van leerlingen uit de klas te trekken. Zonder hogere kosten kan men alle leerlingen uit de klas in het onder- zoek betrekken.

Diverse vormen van steekproeftrekken kunnen desgewenst gecombineerd worden. Zo kan men in elk stratum van een gestratificeerde steekproef een getrapte steekproef trekken.

2.6.6 Intraklassecorrelatie

De onderlinge gelijkenis van personen uit hetzelfde cluster van een getrapte steekproef, ook wel homogeniteit van het cluster genoemd, kan men uitdrukken in een bepaalde maat die de intraklassecorrelatiecoëfficiënt wordt genoemd. In deze paragraaf spreken we over de getrapte steekproef. De intraklassecorrelatiecoëfficiënt is gedefinieerd als

de proportie van de variantie van een variabele in een populatie die is toe te schrijven aan het effect van de clusters. Aan deze definitie ligt een uit de variantie-analyse bekende decompositie van scores ten grondslag. Elke score wordt geschreven als de som van een algemeen gemiddelde, een clustereffect, en een residu.

Het is van groot belang, te weten hoe groot de intraklassecorrelatiecoëfficiënt in een steekproef is. Natuurlijk zal deze grootheid veelal geschat moeten worden; vaak kan men er voor teruggrijpen op eerder onderzoek. Het voert te ver, in dit hoofdstuk in te gaan op het schatten van de intraklassecorrelatiecoëfficiënt. Wel willen we de lezer een indruk geven van de invloed die deze coëfficiënt heeft op het vaststellen van de omvang van de te trekken steekproef. We veronderstellen daartoe dat we het gemiddelde van een kenmerk in een populatie willen schatten met een bepaalde nauwkeurigheid. Een relatieve maat voor de nauwkeurigheid van een schatter is de precisie. De precisie van een schatter is de verhouding van de standaardfout van de schatter en de standaardafwijking van de variabele in de populatie. Zonder de waarden van de standaardfout en de standaardafwijking te kennen, kan men bijvoorbeeld toch als eis formuleren dat de standaardfout ten hoogste een tiende is van de standaardafwijking van de variabele. De precisie wordt aangeduid met het symbool π; de intraklassecorrelatie met het symbool ρ. Merk op dat een kleine respectievelijk grote waarde van π overeenkomt met een grote respectievelijk kleine precisie. Een eenvoudig voorbeeld moge het begrip precisie verduidelijken. Veronderstel dat men het gemiddelde van een variabele wil schatten met een precisie van 0.10. De standaardafwijking van de variabele is niet bekend. Het is bekend dat de standaardfout van een geschat gemiddelde gelijk is aan de standaardafwijking van de variabele gedeeld door de wortel uit het aantal personen in de steekproef. De standaardfout duiden we aan met het symbool SE. Omdat we gesteld hebben dat π gelijk is aan 0.10, kunnen we schrijven: $SE/\sigma = 0.10$. Hieruit volgt dat $SE = 0.10\sigma$. Omdat in het onderhavige geval geldt dat $SE = \sigma/\sqrt{n}$, krijgen we de vergelijking $\sigma/\sqrt{n} = 0.10\sigma$. Als we deze vergelijking oplossen, vinden we dat de steekproef moet bestaan uit $n = 100$ personen om het gemiddelde te schatten met de gewenste precisie.

Als nu in een getrapte steekproef elk der clusters bestaat uit m personen en elk getrokken cluster in zijn geheel wordt beschouwd, dan kan men afleiden dat men c clusters in de steek- proef moet hebben waarbij c gelijk is aan: $\pi^{-2}m^{-1}\{1+(m-1)\rho\}$. De afleiding van dit resultaat is te vinden in Cochran (1977). De formule geldt alleen als de populatie heel erg groot is; wij geven haar alleen voor illustratieve doeleinden. Als de intraklassecorrelatie gelijk is aan 1, blijkt c gelijk te zijn aan π^{-2}. Het doet er niet meer toe hoe groot een cluster is: als men er een waarneming uit heeft gedaan, heeft men ze immers allemaal. Als echter de intraklassecorrelatie gelijk is aan 0, blijkt c

gelijk te zijn aan $\pi^{-2}m^{-1}$. In dat geval is het aantal te trekken clusters omgekeerd evenredig met de omvang van elk der clusters.

In de praktijk neemt men vaak intraklassecorrelaties waar tussen 0.05 en 0.20. Bij wijze van voorbeeld is in tabel 2.1 voor verschillende combinaties van clustergrootte, precisie en intraklassecorrelatie aangegeven hoeveel clusters men in de steekproef moet hebben om een gemiddelde te schatten met de gegeven precisie.

Tabel 2.1

Aantal te trekken clusters bij gegeven precisie, intraklassecorrelatie en clustergrootte

	π					
	0.05		0.075		0.10	
ρ	$m=4$	$m=20$	$m=4$	$m=20$	$m=4$	$m=20$
0	100	20	45	9	25	5
0.05	115	39	52	18	29	10
0.10	130	58	58	26	33	15
0.15	145	77	65	35	37	20
0.20	160	96	72	43	40	24
0.25	175	115	78	52	44	29

Uit de tabel blijkt dat het aantal te trekken clusters toeneemt als de intraklassecorrelatie toeneemt. Dat komt doordat een relatief grote intraklassecorrelatie betekent dat elke persoon in een cluster relatief weinig nieuwe informatie aandraagt: als men er een heeft geobserveerd, kan men al vrij goed voorspellen wat andere observaties uit dezelfde cluster zullen opleveren. Ook blijkt uit de tabel dat het aantal te trekken clusters toeneemt als π afneemt en dus de precisie toeneemt. Dat komt overeen met de eerder genoemde eigenschap van een standaard- fout, kleiner te worden als het aantal observaties groter wordt. Tenslotte blijkt dat men, bij dezelfde intraklassecorrelatie en precisie, minder clusters nodig heeft naarmate de clusters groter zijn. Dit effect neemt af naarmate de intraklassecorrelatie toeneemt, om de eerder al genoemde reden van verlies aan informatieve waarde van elke waarneming.

2.7 Proefopzetten

Zoals gezegd, is het vaak niet mogelijk een persoon alle stimuli voor te leggen waar men belang in stelt. Ook hier leggen tijd en geld hun beperkingen op. Men moet dan

procedures bedenken waarmee men zo goed mogelijk de informatie inwint die men wil hebben. Zulke procedures worden toewijzingsprocedures of proefopzetten genoemd. We beperken ons hier tot enige algemene beschouwingen. Veronderstel dat, bijvoorbeeld vanwege een beperkt budget of vanwege de beperkte tijd waarin men over een persoon kan beschikken, het totale aantal te verzamelen responsen vastligt. De vraag rijst dan op welke wijze men de aantallen personen en stimuli in het uit te voeren onderzoek moet kiezen. Als de stimuli op de een of andere wijze op elkaar lijken, waardoor men uit responsen op de ene stimulus een redelijk goede voorspelling kan maken van responsen op de andere stimulus, heeft het niet veel zin alle stimuli aan personen voor te leggen. Men beperkt dan het aantal aan te bieden stimuli, en trekt een grotere steekproef van personen.

Omdat het meestal niet mogelijk is alle personen alle stimuli aan te bieden, rijst de vraag hoe men de stimuli over de personen moet verdelen. Doorgaans verdeelt men de te onder- zoeken stimuli in een aantal elkaar uitsluitende groepjes stimuli en de personen in elkaar uit- sluitende groepjes personen. Aan elk groepje personen wijst men een van de groepjes stimuli toe; men spreekt van multiple matrix sampling. Het verdient aanbeveling de verdeling van groepjes stimuli over groepjes personen evenwichtig te houden: alle stimuli en alle personen moeten ongeveer evenveel te doen hebben. Enerzijds voorkomt men hiermee dat sommige personen veel meer werk moeten verrichten dan andere; anderzijds bewerkstelligt men ermee dat grootheden die met statistische methoden worden geschat, niet erg uiteenlopen in de met schattingen nu eenmaal gepaard gaande standaardfouten. Daarom maakt men in de psycho- metrie veel gebruik van onvolledige proefopzetten. Dat zijn proefopzetten waarin stimuli zodanig aan personen worden aangeboden dat niet elke persoon alle stimuli voorgelegd krijgt.

Men kan vaak met vrucht gebruik maken van aanwezige kennis om stimuli toe te wijzen aan personen. Op theoretische gronden of op grond van eerder onderzoek stelt men vast dat de reacties van bepaalde personen op bepaalde stimuli op voorhand goed te voorspellen zijn. Het is dan zonde van de moeite en het geld zulke stimuli toch aan die personen aan te bieden. Zo kan men besluiten items die men op voorhand erg gemakkelijk acht, niet voor te leggen aan leerlingen die men op voorhand heel knap vindt: men durft de veronderstelling wel aan dat zulke leerlingen zulke items goed zullen beantwoorden.

Men kan vaststellen dat onvolledige proefopzetten eerder regel dan uitzondering zijn in psychometrisch onderzoek, op grond van de geschetste overwegingen en omdat in praktijk budgetten voor onderzoek beperkt zijn.

2.8 Stimuli

Stimuli kunnen vele vormen aannemen, van ongestructureerde vragenlijsten tot welomschreven opdrachten en toetsen die bestaan uit een aantal met elkaar samenhangende items. Welke soort stimuli men gebruikt, is natuurlijk afhankelijk van het soort probleem dat men bestudeert. Stimuli worden geacht operationalisaties te zijn van het te onderzoeken gedrag, ze moeten valide zijn. Zo ligt het voor de hand leerlingen optelopgaven voor te leggen indien men wil weten in hoeverre leerlingen getallen kunnen optellen.

In de praktijk is het operationaliseren van gedrag in stimuli geen eenvoudige zaak. In het onderwijs maakt men veel gebruik van items: vragen die door leerlingen beantwoord moeten worden. Maar ook komt het voor dat door personen vertoonde gedragingen door een of meer beoordelaars of keurmeesters worden beoordeeld. Voorbeelden daarvan zijn het kunstrijden op de schaats, het Eurovisie Songfestival en de verkiezing van Miss World. De beoordelaars beschikken over een beoordelingsschema of beoordelingsmodel; voor Miss World bevat dit model een lijst met ideale maten. In het beoordelingsmodel staat vermeld welke interpretatie aan een waarneming moet worden gegeven.

Omdat het construeren van goede stimuli erg moeilijk is, zal men doorgaans niet met een enkele stimulus volstaan als operationalisatie van het te onderzoeken gedrag. Er is dus reden genoeg om meer stimuli aan te bieden; door vaker stimuli van hetzelfde soort aan te bieden, voert men als het ware een meting herhaaldelijk uit. Men verhoogt op deze manier de betrouwbaarheid van de meting. Daarbij veronderstelt men dat niet de reactie op elke stimulus van belang is maar dat het waargenomen responspatroon betekenis heeft. De veel gehoorde uitroep "Deze vraag meet toch geen intelligentie!" snijdt dan ook geen hout; slechts de combinatie van antwoorden heeft betekenis. Die betekenis ontleent een responspatroon aan een meetmodel.

2.9 Meetmodellen

Door gebruik te maken van een meetmodel kan men een responspatroon betekenis geven, dat wil zeggen interpreteren. Een voorbeeld van een meetmodel is de Guttmanschaal (Guttman, 1950). Dit model veronderstelt dat het mogelijk is items te ordenen naar moeilijkheidsgraad
en personen naar vaardigheidsniveau. Ook veronderstelt het model dat de moeilijkheidsgraden en de vaardigheidsniveaus op dezelfde schaal zijn uitgedrukt; personen en

items liggen op dezelfde schaal. Daarmee is ook een relatie gegeven tussen elk der personen en elk der items. Personen die op de schaal rechts van het item liggen, zullen het item juist beantwoorden; de andere personen geven een fout antwoord. Als juiste antwoorden worden gecodeerd met een 1 en foute antwoorden met een 0, en men de items rangschikt van gemakkelijk naar moeilijk en de personen van dom naar knap, zal men het volgende kunnen vaststellen. Aangezien elke persoon het juiste antwoord geeft op de items die links van hem liggen en het foute antwoord op de items die rechts van hem liggen, kunnen er alleen maar de volgende antwoordpatronen voorkomen: allemaal enen, allemaal nullen, of een aantal enen die gevolgd worden door een aantal nullen. Natuurlijk weet men niet of er aan de veronderstellingen van het meetmodel is voldaan. Het meetmodel krijgt zin doordat men van de andere kant begint. Men probeert, als men de antwoorden van personen op items heeft geregistreerd, de items en de personen zo te rangschikken dat de resulterende antwoordpatronen de door het meetmodel vereiste structuur hebben. Als dat lukt, heeft men een verklaring van het vertoonde gedrag gevonden. Die verklaring is gegeven in de veronderstellingen van het meetmodel. In dit voorbeeld van een meetmodel laten we een aantal belangrijke kwesties onbesproken. Zo zal men in de praktijk altijd antwoordpatronen vinden die niet de door het model vereiste samenstelling hebben. Men kan dan het model voor onhoudbaar verklaren. Maar ook kan men het meetmodel omwerken tot een probabilistisch of kansmodel: men eist dan alleen maar dat de kans op van het model afwijkende antwoordpatronen een zekere waarde niet overschrijdt. Zulke probabilistische meetmodellen komen in dit boek uitgebreid aan de orde.

Een verzameling stimuli, te zamen met een door een meetmodel verschaft interpretatie- kader, noemt men een meetinstrument. Een vragenlijst die naar een aantal socio-economische eigenschappen van personen vraagt, behoeft geen meetinstrument te zijn. Men kan de groep personen naar een aantal concrete zaken classificeren en daarmee volstaan. Zo'n inventarisatie kan een praktisch nut dienen maar levert zonder een model geen kennis en inzicht op.

Bij een meetinstrument is er doorgaans sprake van een niet direct waar te nemen eigenschap maar van een latente variabele: de moeilijkheidsgraad van een vraag of het vaardigheidsniveau van een persoon. Als iemand veel van de hem voorgelegde optelitems goed beantwoordt, concludeert men daaruit dat hij beschikt over een grote mate van optelvaardigheid. Het is van belang er op te wijzen dat een psychometrisch meetmodel niet noodzakelijkerwijze een psychologische theorie weergeeft. Zelfs als een Guttmanschaal blijkt te passen bij een tabel met antwoordpatronen, weet men nog niet waarom sommige items gemakkelijker zijn dan andere. De gevonden rangschikking van items en personen kan echter van groot nut zijn bij het formuleren van een theorie.

3

Klassieke testtheorie en generaliseerbaarheidstheorie

De klassieke testtheorie beschrijft het verschijnsel meetfout en procedures om de grootte van meetfouten te bepalen. Het uitgangspunt van de klassieke testtheorie is een meting x_{vt} die verkregen is door een meetinstrument t voor te leggen aan een persoon v. Zoals is uiteen- gezet in het vorige hoofdstuk, wordt een meting altijd gecodeerd als een getal. Zo'n gecodeerde meting noemt men een score. De klassieke testtheorie houdt zich niet bezig met de aard, het schaalniveau en de interpretatie van een score. Zij houdt zich met slechts een enkel probleem bezig, en wel met de meetfout waarmee een score x_{vt} behept is. De meetfout wordt geacht op te treden doordat men bij het meten niet alle factoren in de hand heeft die op een meting van invloed zijn. Zulke factoren verstoren de meetprocedure en zorgen er voor dat men niet de meting krijgt die men graag had willen hebben maar een daar enigszins van afwijkende score. Verstorende factoren kunnen zijn gelegen in de te meten persoon, in het meetinstrument, en in de meetsituatie. Een voorbeeld van de eerste soort is de bloeddruk: deze vertoont in de loop van de dag zulke grote fluctuaties dat een enkele meting eigenlijk onvoldoende is. Een voorbeeld van de tweede soort verstorende factoren is de thermometer. Dat instrument wisselt warmte uit met het te meten voorwerp, waardoor de thermometer niet de exacte temperatuur van het voorwerp aangeeft. Een voorbeeld van een verstoring in de meetsituatie is het eindexamen dat wordt afgenomen in een schoolgebouw waarnaast een heistelling palen de grond in boort.

De belangrijkste parameters uit de klassieke testtheorie zijn correlaties en standaard-afwijkingen. Het gebruik van dergelijke parameters brengt met zich mee dat alle uitspraken van de klassieke testtheorie over personen en over meetinstrumenten gerelateerd zijn aan een bepaalde populatie. Zo kan men eigenschappen van een meetinstrument die bepaald zijn in een populatie, niet zonder meer voor geldend houden in een andere populatie. Voor een aantal meetproblemen schiet de klassieke testtheorie dan ook tekort. De wens, te kunnen beschikken over parameters van

personen en meetinstrumenten die niet aan een populatie gebonden zijn, heeft geleid tot de itemresponstheorie. Deze theorie wordt behandeld in hoofdstuk 4.

De klassieke testtheorie wordt eerst, in de paragrafen 3.1 tot en met 3.6, in abstracte termen beschreven. In de paragrafen 3.7 tot en met 3.10 worden diverse grootheden concreet geïllustreerd aan de hand van een voorbeeld. Daarbij worden ook grootheden behandeld die optreden bij het construeren van toetsen. De toets uit het voorbeeld is klein gehouden om het de lezer mogelijk te maken het rekenwerk te volgen. Een uitbreiding van de klassieke testtheorie, de generaliseerbaarheidstheorie, wordt in de paragrafen 3.11 tot en met 3.14 besproken.

3.1 Ware score

De waargenomen score is door de verstorende factoren niet altijd de meting die we zouden willen hebben. De klassieke testtheorie veronderstelt nu dat het effect van de verstorende factoren beschouwd kan worden als een aselecte trekking uit een kansverdeling. In feite is dit de enige veronderstelling die de klassieke testtheorie kent. De afleiding die nu volgt is gebaseerd op Novick (1966). Uit de zojuist genoemde veronderstelling kan men de gehele klassieke testtheorie opbouwen. Als de bij de meting x_{vt} optredende meetfout wordt aangeduid met ε_{vt}, veronderstelt de klassieke testtheorie dat deze meetfout een realisatie is van een toevalsvariabele E_{vt}. Deze toevalsvariabele draagt twee subscripten om aan te geven dat zij varieert binnen de combinatie van de vaste persoon v en het vaste meetinstrument t. Beschouw nu de voor de meetfout gecorrigeerde meting $\tau_{vt} = x_{vt} - \varepsilon_{vt}$. Men kan dan ook schrijven: $x_{vt} = \tau_{vt} + \varepsilon_{vt}$. Deze uitdrukking schrijft de score x_{vt} als een ontbinding, een decompositie, in twee termen. De eerste term, τ_{vt}, zou men kunnen opvatten als de meting die men had willen verkrijgen. Maar de gegeven ontbinding is niet uniek. Men kan namelijk bij de term τ_{vt} een willekeurige constante c optellen en deze constante van de term ε_{vt} aftrekken zonder dat het resultaat verandert: $x_{vt} = \tau_{vt} + \varepsilon_{vt} = (\tau_{vt} + c) + (\varepsilon_{vt} - c)$. In feite is dit een geval van een vergelijking met twee onbekenden. Om met de gegeven decompositie uit de voeten te kunnen, moet men normeren. Daaronder verstaat men het kiezen en vastleggen van een waarde voor de constante c. In de klassieke testtheorie heeft men voor de volgende normering gekozen. Aangezien E_{vt} een toevalsvariabele is met realisaties ε_{vt}, en τ_{vt} een vaste waarde heeft, is x_{vt} een realisatie van een toevalsvariabele X_{vt}. Voor de constante c is in de klassieke testtheorie de verwachte waarde van de toevalsvariabele E_{vt} gekozen: $c = \mathscr{E}(E_{vt})$. De verwachte waarde van een toevalsvariabele kan men in dit boek opvatten als het

gemiddelde van een hele grote steekproef van trekkingen uit de verdeling van die variabele. De verwachte waarde van een constante is gelijk aan die constante. Met de gekozen normering kan men nu de toevalsvariabele X_{vt} schrijven als: $X_{vt} = \{\tau_{vt} + \mathscr{E}(E_{vt})\} + \{E_{vt} - \mathscr{E}(E_{vt})\}$. Daaruit volgt onmiddellijk dat $\mathscr{E}(X_{vt}) = \tau_{vt} + \mathscr{E}(E_{vt})$. Ook deze decompositie moet genormeerd worden. In de klassieke testtheorie stelt men daartoe $\mathscr{E}(E_{vt})$ gelijk aan 0. Het resultaat is de volgende belangrijke uitdrukking:

$$\mathscr{E}(X_{vt}) = \tau_{vt}. \tag{3.1}$$

Het rechterlid van (3.1) heet in de klassieke testtheorie de ware score van persoon *v* op meet- instrument *t.* Men dient te beseffen dat de door (3.1) gedefinieerde ware score een wis- kundige constructie is en niet noodzakelijkerwijze gelijk is aan de score die verkregen zou zijn als er geen verstorende factoren aanwezig waren. Het kan bijvoorbeeld goed zijn dat de toevalsvariabele X_{vt} alleen maar gehele waarden kan aannemen; dat sluit echter niet uit dat de verwachte waarde van die variabele, de ware score, een gebroken getal is.

3.2 De centrale formule van de klassieke testtheorie

De ware score is, omdat hij is gedefinieerd als een verwachte waarde, een maat voor de centrale tendentie van de scores: hij geeft aan om welke waarde de verkregen metingen variëren. Het is van groot belang, te weten in welke mate de metingen rondom de ware score variëren. Bekende maten voor de variatie van een toevalsvariabele zijn de variantie en de standaardafwijking van die variabele. De variantie van een toevalsvariabele is gelijk aan de verwachte waarde van het kwadraat van het verschil tussen een score en de daarbij behorende ware score. Voor de toevalsvariabele X_{vt} schrijft men de variantie als volgt: $\sigma^2_{X_{vt}} = \mathscr{E}\{(X_{vt} - \tau_{vt})^2\}$. Omdat geldt dat $X_{vt} - \tau_{vt}$ gelijk is aan E_{vt} en omdat $\mathscr{E}(E_{vt})$ gelijk is aan 0, kan men de zojuist geschreven variantie ook schrijven als: $\sigma^2_{X_{vt}} = \mathscr{E}\{(E_{vt})^2\}$. De laatste uitdrukking kan men natuurlijk ook schrijven als: $\sigma^2_{E_{vt}}$.

Merk op dat de in deze paragraaf genoemde varianties alle betrekking hebben op de variatie van toevalsvariabelen die zijn gedefinieerd voor een vaste persoon *v* en een vast meetinstrument *t*. Om de varianties te kunnen schatten, zou men moeten beschikken over herhaalde metingen van *v* met *t*, verkregen onder identieke omstandigheden. Door de eerder genoemde verstorende factoren is het echter niet mogelijk, herhaalde metingen te verkrijgen onder identieke omstandigheden. In plaats

van herhaalde metingen te gebruiken, gaat de klassieke testtheorie er toe over meer personen tegelijk te beschouwen. Het is duidelijk dat nu kenmerken van een populatie $\mathcal{P}$ van personen een rol gaan spelen.

Beschouw een willekeurig uit de populatie $\mathcal{P}$ getrokken persoon. Om aan te geven dat de persoon willekeurig is getrokken, duiden we die persoon aan met een $\star$. Zodra we de persoon $\star$ hebben getrokken, geldt alles wat hierboven gezegd is. Men kan denken aan een tweestapsprocedure: eerst trekt men willekeurig een persoon $\star$ uit de populatie $\mathcal{P}$, en dan trekt men een meetfout $\varepsilon_{\star t}$ uit de verdeling van de toevalsvariabele $E_{\star t}$. Bij de persoon $\star$ behoort een ware score $\tau_{\star t}$. Men kan nu ook zeggen dat er drie nieuwe toevalsvariabelen zijn gemaakt: $T_{\star t}$, $E_{\star t}$ en $X_{\star t}$. De laatste twee variabelen variëren zowel over personen als binnen de aselect gekozen persoon; de eerste varieert alleen over personen. De betrekking tussen de drie toevalsvariabelen kan men schrijven als: $X_{\star t} = T_{\star t} + E_{\star t}$. Omdat we in het vervolg steeds een enkel meetinstrument en een enkele populatie beschouwen, laten we waar dat mogelijk is de subscripten weg. De laatst geschreven betrekking kan men dan schrijven als:

$$X = T + E \ . \tag{3.2}$$

Formule (3.2) is de centrale formule van de klassieke testtheorie. Men kan er, jammer genoeg, niet aan zien dat de toevalsvariabele T alleen over personen varieert maar niet binnen een persoon, en dat de toevalsvariabelen X en E zowel tussen de personen als binnen elke persoon variëren. In het bovenstaande is daarom uiteengezet hoe deze formule tot stand komt.

3.3 Betrouwbaarheid

Uit (3.2) kan men enige interessante betrekkingen afleiden. In de eerste plaats geldt dat de verwachte waarde van de toevalsvariabele E over de populatie $\mathcal{P}$ gelijk is aan 0: $\mathcal{E}_{\mathcal{P}}\mathcal{E}(E) = \mathcal{E}_{\mathcal{P}}(0) = 0$. Er zijn twee verwachtingen genomen: in de eerste plaats de verwachting over de meetfouten binnen een persoon, en in de tweede plaats de verwachting over personen van de verwachte meetfout. Dit komt overeen met het feit dat E zowel binnen een persoon als over personen varieert.

In de tweede plaats kan men afleiden dat de correlatie tussen de variabelen T en E gelijk is aan 0. Immers, voor elke persoon v in $\mathcal{P}$ geldt dat $\mathcal{E}(E_{vt}) = 0$. Dit geldt dan ook voor een willekeurig uit de populatie $\mathcal{P}$ getrokken persoon $\star$. A fortiori geldt dit voor elke persoon $\star$ uit $\mathcal{P}$ die een ware score gelijk aan $\tau_{\star t}$ heeft: $\mathcal{E}(E_{\star t}|\tau_{\star t}) = 0$. Dit geldt natuurlijk voor elke waarde van $\tau_{\star t}$. De uitdrukking $\mathcal{E}(E_{\star t}|\tau_{\star t})$ heet: de regressie

van E op T. Aangezien de regressie van E op T gelijk is aan 0, is ook de correlatie tussen E en T gelijk aan 0.

In de derde plaats kan men uit de decompositie van X die gegeven is in (3.2), de volgende decompositie afleiden van de variantie σ_X^2 van de variabele X:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2. \tag{3.3}$$

De drie varianties zijn de varianties van respectievelijk de waargenomen toetsscores, de ware toetsscores en de meetfouten. Men noemt de drie varianties doorgaans: geobserveerde variantie, ware variantie en foutenvariantie.

Een van de voornaamste grootheden in de klassieke testtheorie is de betrouwbaarheid. Deze grootheid, die wordt voorgesteld door het symbool ρ_{XT}^2, is als volgt gedefinieerd:

$$\rho_{XT}^2 = \sigma_T^2 / \sigma_X^2 = \sigma_T^2 / \{\sigma_T^2 + \sigma_E^2\}. \tag{3.4}$$

Zolang de geobserveerde variantie groter is dan 0, neemt de betrouwbaarheid waarden aan tussen 0 en 1. De betrouwbaarheid is gelijk aan 0 als er geen ware variantie is: men meet alleen maar meetfouten met het meetinstrument. De betrouwbaarheid is gelijk aan 1 als er geen sprake is van meetfouten: $\sigma_E^2 = 0$, wat overeenkomt met $\sigma_X^2 = \sigma_T^2$. Elke geobserveerde score van een persoon is dan gelijk aan de ware score van die persoon. In het uitzonderlijke geval dat σ_X^2 gelijk is aan 0, is de betrouwbaarheid niet gedefinieerd.

Waarom de betrouwbaarheid wordt aangeduid met het symbool ρ_{XT}^2, wordt duidelijk als men de correlatie beschouwt tussen de geobserveerde scores X en de ware scores T. De teller van deze correlatie is gelijk aan de covariantie tussen X en T:

$$Cov(X,T) = \mathcal{E}[\{X - \mathcal{E}(X)\} \times \{T - \mathcal{E}(T)\}] =$$

$$\mathcal{E}([\{T - \mathcal{E}(T)\} + \{E - \mathcal{E}(E)\}] \times \{T - \mathcal{E}(T)\}) =$$

$$\mathcal{E}\{T - \mathcal{E}(T)\}^2 + \mathcal{E}[\{T - \mathcal{E}(T)\} \times \{E - \mathcal{E}(E)\}] = \sigma_T^2 + Cov(T,E) =$$

$$\sigma_T^2 + \sigma_T \sigma_E \rho_{TE} = \sigma_T^2.$$

In deze afleiding is gebruik gemaakt van het eerder gegeven resultaat dat de correlatie tussen T en E, hier aangeduid met ρ_{TE}, gelijk is aan 0. De noemer van de correlatie X en T is gelijk aan $\sigma_X \sigma_T$. We zien dan dat de correlatie ρ_{XT} tussen de geobserveerde

scores X en de ware scores T gelijk is aan σ_T/σ_X; deze uitdrukking is gelijk aan de wortel uit de in (3.4) gegeven uitdrukking voor de betrouwbaarheid.

3.4 Standaardmeetfout

De wortel uit de foutenvariantie σ_E^2 heet de standaardmeetfout. Uit (3.4) kan men afleiden dat de standaardmeetfout σ_E kan worden bepaald uit de geobserveerde variantie en de betrouwbaarheid: $\sigma_E = \sigma_X(1-\rho_{XT}^2)^{½}$. De standaardmeetfout is uitgedrukt in de schaal- eenheid van het meetinstrument. Men kan twee standaardmeetfouten van verschillende meetinstrumenten dan ook niet zomaar met elkaar vergelijken. De betrouwbaarheid daaren- tegen is louter een getal; men kan de betrouwbaarheden van twee toetsen wel onderling vergelijken. De standaardmeetfout wordt voornamelijk gebruikt om uit een geobserveerde score een intervalschatting voor de ware score te bepalen.

Men heeft het wel als een bezwaar van de klassieke testtheorie gezien dat er een enkele standaardmeetfout is die wordt toegepast bij elke score x_{vt}. Het wordt onrealistisch geacht aan te nemen dat een toets op elk scoreniveau even nauwkeurig meet. Aan dit bezwaar wordt tegemoet gekomen in de itemresponstheorie die in hoofdstuk 4 wordt besproken. Ook binnen de klassieke testtheorie heeft men dit bezwaar erkend. Er zijn diverse procedures ontwikkeld om voor verschillende scoreniveaus een eigen standaardmeetfout te bepalen. Een overzicht van deze procedures vindt men bij Feldt, Steffen en Gupta (1985). Een van die procedures is ontwikkeld door Thorndike (1951).

De methode van Thorndike maakt gebruik van het begrip parallelle metingen. Dit begrip wordt besproken in paragraaf 3.6.1. Een paar eigenschappen van parallelle metingen worden hier gebruikt. Veronderstel dat het mogelijk is, het meetinstrument te verdelen in twee parallelle deeltoetsen. Voor zulke parallelle deeltoetsen, met scorevariabelen X_1 en X_2, geldt dat $\mathscr{E}(X_1) = \mathscr{E}(X_2)$ en $\sigma_{X_1}^2 = \sigma_{X_2}^2$. Bovendien geldt dat de bijbehorende meetfouten E_1 en E_2 onderling onafhankelijk, en dus ongecorreleerd zijn. De standaardafwijking van de verschilscore $X_1 - X_2$ kan men nu schrijven:

$$\sigma_{(X_1-X_2)} = \sigma_{(E_1-E_2)} = (\sigma_{E_1}^2 + \sigma_{E_2}^2)^{½} = \sigma_E. \qquad (3.5)$$

In deze afleiding is gebruik gemaakt van het feit dat de correlatie tussen de meetfouten E_1 en E_2 gelijk is aan 0, van het feit dat $\sigma_{E_1}^2 = \sigma_{E_2}^2$, en van het feit dat $\sigma_{E_1}^2 = ½\sigma_E^2$. Met (3.5) kan men de standaardmeetfout van een meetinstrument schatten. Thorndike

stelt voor, (3.5) toe te passen op deelgroepen van personen die dezelfde score hebben. Zulke groepen noemt men wel scoregroepen. Het is dan mogelijk, met behulp van (3.5) standaardmeetfouten te schatten in verschillende scoregroepen afzonderlijk. In de praktijk zal het vaak nodig zijn, scoregroepen samen te nemen om te komen tot groepen met een voldoende aantal waarnemingen voor het nauwkeurig schatten van de standaardmeetfout.

3.5 Schattingen van de ware score

Een voor de hand liggende schatter van de ware score τ is de waargenomen score x. De waargenomen score is een zuivere schatter van de ware score. Men noemt een schatter zuiver als zijn verwachte waarde gelijk is aan de te schatten parameter. De vraag rijst hoe precies de geobserveerde score als schatter van de ware score is. Onder de veronderstelling dat de meetfout binnen elke persoon een normale verdeling heeft met gemiddelde 0 en standaard- afwijking σ_E, bestaat er een intervalschatting van de ware score. Dit interval bestaat uit de getallen $\hat{\tau}$ waarvoor geldt dat de volgende nulhypothese bij een van te voren vastgesteld significantieniveau niet wordt verworpen:

$$H_0\text{: } x - z \times \sigma_E \leq \hat{\tau} \leq x + z \times \sigma_E \qquad (3.6)$$

waarin z de standaardnormale afwijking is die behoort bij het gekozen significantieniveau. Als dit bijvoorbeeld vastgesteld is op de waarde 0.05, is de waarde van z gelijk aan 1.96. Merk op dat (3.6) een schattingsvoorschrift is. Men kiest eerst de getallen z en $\hat{\tau}$, terwijl σ_E bekend is verondersteld. Dan neemt men de realisatie x_{vt} van de toevalsvariabele X waar, en vult de verkregen waarde in (3.6) in. Als de gegeven ongelijkheden worden geschonden, besluit men dat het van te voren gekozen getal $\hat{\tau}$ geen goede schatting is van de ware score. Alle getallen $\hat{\tau}$ waarvoor de ongelijkheden in (3.6) niet geschonden zijn, vormen gezamenlijk een intervalschatting voor de ware score die behoort bij de geobserveerde score x. In de praktijk berekent men natuurlijk, zodra de score x is geobserveerd, de intervalgrenzen $x \pm z \times \sigma_E$. Het zo verkregen interval heet in de statistiek een betrouwbaarheidsinterval voor de ware score; de naam heeft niets te maken met het begrip betrouwbaarheid uit de klassieke testtheorie.

Een tweede schatter voor de ware score is de zogenoemde Kelley-schatter (Kelley, 1947; Lord & Novick, 1968). Deze schatter levert een kleinere standaardfout op, maar daarvoor betaalt men wel een prijs. Men moet namelijk veronderstellen dat de regressie

van T op X lineair is. Men kan afleiden dat deze regressie de volgende gedaante heeft:

$$\mathcal{E}(T \mid X = x) = (\rho^2_{XT})\, x + (1 - \rho^2_{XT})\, \bar{x} \qquad (3.7)$$

waarin $\bar{x}$ de gemiddelde geobserveerde score is van de steekproef van personen uit de populatie $\mathcal{P}$ aan wie men de toets heeft afgenomen (zie voor de afleiding Lord en Novick, 1968, p. 65). Zoals Kelley (1947, p. 409) zegt: "This is an interesting equation in that it expresses the estimate of true ability as a weighted sum of two separate estimates - one based upon the individual's observed score, [x], and the other based upon the mean of the group to which he belongs, ... If the test is highly reliable, much weight is given to the test score and little to the group mean, and vice versa." De standaardfout van de Kelley-schatter is gelijk aan $\sigma_E\,(\rho^2_{XT})^{\frac{1}{2}}$, de spreiding van het verschil $T - \mathcal{E}(T \mid X = x)$. In de regressie-analyse noemt men deze spreiding wel de spreiding om de regressielijn. Als men de standaardfout van de Kelley-schatter substitueert voor σ_E in (3.6) verkrijgt men een andere intervalschatter voor de ware score. Deze schatter leidt tot kleinere intervallen dan de schatter uit (3.6) omdat de gebruikte standaardfout kleiner is dan de in (3.6) als standaardfout gebruikte standaard-meetfout.

In de praktijk zal men niet vaak schattingen van ware scores tegenkomen. De reden daarvan is, dat toetsscores doorgaans relatief worden geïnterpreteerd. Niet de waarde van de score zelf is van belang, maar zijn rangnummer in de verdeling van scores in de populatie $\mathcal{P}$. De beschreven schatters van de ware score leiden tot dezelfde rangorde van personen als de geobserveerde scores; daarom heeft men geen geschatte ware scores nodig. Anders wordt het als een score wordt gerelateerd aan een op voorhand gegeven criterium. Zo'n criterium is bijvoorbeeld een getal waarboven een score moet liggen om als voldoende aangemerkt te worden. Dan bestaat de mogelijkheid, door het gebruik van geschatte ware scores het aantal classificatiefouten te verminderen.

In veel boeken en artikelen over de klassieke testtheorie ziet men verwarring optreden tussen de begrippen standaardfout en standaardmeetfout. De standaardfout, die eigenlijk 'standaardfout van een schatting' (standard error of estimate) heet, is een maat voor de nauwkeurigheid van een schatter. Men kan de nauwkeurigheid van een schatter opvoeren door een grotere steekproef te trekken (hoofdstuk 2). De standaardmeetfout daarentegen is een kenmerk van een toets; het groter maken van een steekproef van aan de toets onderworpen personen heeft op de standaardmeetfout geen enkele invloed. Om de standaardmeetfout kleiner te maken moet men de betrouwbaarheid van de toets groter maken. Een van de middelen daartoe is, de toets met een aantal items te verlengen. Het verlengen van een toets wordt besproken in

paragraaf 3.6.2. De verwarring tussen de begrippen standaardfout en standaardmeetfout wordt wellicht verklaard door het feit dat de standaardmeetfout de rol speelt van standaardfout in (3.6).

3.6 Het schatten van de betrouwbaarheid en de standaardmeetfout

Er zijn diverse procedures ontwikkeld om de betrouwbaarheid en de standaardmeetfout van een toets te schatten. Men kan die grootheden immers niet precies bepalen omdat men in de praktijk alleen maar kan beschikken over een steekproef van personen uit de populatie $\mathcal{P}$. In de volgende paragrafen bespreken we methoden om de betrouwbaarheid en de standaardmeetfout te schatten uit parallelle metingen, uit twee afnames van de toets, uit toetsverlenging, en uit coëfficiënt alpha als een ondergrens van de betrouwbaarheid. In paragraaf 3.11 zullen we zien dat men ook de betrouwbaarheid kan schatten door middel van een variantie-analyse van itemscores.

3.6.1 Parallelle metingen

Een belangrijk begrip dat is toegevoegd aan de klassieke testtheorie is dat van de parallelle meting. Men beschikt niet alleen over de realisaties van de geobserveerde toetsscore X maar ook over die van een toetsscore X' die voldoet aan de volgende eigenschappen: $\mathcal{E}(X') = \mathcal{E}(X)$ en $\sigma^2_{X'} = \sigma^2_X$ in elke deelpopulatie van $\mathcal{P}$. Metingen die aan deze eigenschappen voldoen, noemt men parallelle metingen, of ook wel streng parallelle metingen. Beschouw nu de correlatie $\rho_{XX'}$ tussen parallelle metingen. De teller hiervan is gelijk aan:

$$\mathrm{Cov}(X, X') = \mathrm{Cov}(T + E, T + E') = \mathrm{Cov}(T, T) + \mathrm{Cov}(E, E') = \sigma^2_T + \mathrm{Cov}(E, E').$$

Nu wordt er verondersteld dat de bij beide metingen optredende meetfouten E en E' onder- ling onafhankelijk zijn; de meetfouten zijn niet gecorreleerd. Een correlatie ongelijk aan nul zou duiden op de aanwezigheid van een factor die beide metingen systematisch beïnvloedt. Bij parallelle metingen veronderstelt men dat zo'n factor er niet is. De meetfouten worden geacht experimenteel onafhankelijk te zijn. Experimentele onafhankelijkheid brengt met zich mee dat de meetouten niet gecorreleerd zijn. Er geldt dus: $\mathrm{Cov}(E, E') = 0$, en dus $\mathrm{Cov}(X, X') = \sigma^2_T$. De noemer van de correlatie tussen X en X' is gelijk aan: $\sigma_X \sigma_{X'} = \sigma_X \sigma_X = \sigma^2_X$. We zien hieruit dat de correlatie tussen parallelle metingen, $\rho_{XX'}$, gelijk is aan de betrouwbaarheid van

de meting X en ook aan die van de meting X'. Dit verklaart het gebruik van het symbool $\rho_{XX'}$ voor de betrouwbaarheid in veel boeken en artikelen over de klassieke testtheorie.

In de praktijk is het niet eenvoudig, parallelle metingen te construeren. Soms slaagt men er in metingen te maken die wel een paar, maar niet alle eigenschappen van parallelle metingen hebben. In tabel 3.1 zijn enige vormen van parallelliteit opgesomd, die afnemen in de strengheid van de eisen.

Tabel 3.1
Enige vormen van parallelliteit

Soort parallelliteit	Eigenschappen
Parallelliteit	$\mathscr{E}(X) = \mathscr{E}(X'),\ \sigma_X^2 = \sigma_{X'}^2$
Tau-equivalentie	$\mathscr{E}(X) = \mathscr{E}(X')$
Essentiële tau-equivalentie	$\mathscr{E}(X) = \mathscr{E}(X') + \kappa\ (\kappa \neq 0)$
Congenerieke parallelliteit	$T = \lambda T' + \kappa,\ (\lambda \neq 0)$

In deze tabel zijn κ en λ constanten die van de meetinstrumenten afhangen. De genoemde eigenschappen gelden in elke deelpopulatie van $\mathscr{P}$. Dat betekent onder meer dat voor elke persoon de ware scores op de parallelle toetsen aan elkaar gelijk zijn, en dus dat $\sigma^2(T) = \sigma^2(T')$. Uit tabel 3.1 ziet men dat men als eerste de veronderstelling laat vallen dat parallelle toetsen dezelfde geobserveerde variantie hebben en dus dezelfde foutenvariantie. Daarna verruimt men de relatie die tussen de ware scores van de beide toetsen bestaat: voor essentieel tau-equivalente metingen verschillen de ware scores een constante, terwijl voor congenerieke metingen de ware scores lineaire transformaties zijn van elkaar. Of aan de diverse vormen van parallelliteit is voldaan, kan men onderzoeken met methoden voor lineaire-structuurmodellen. Zulke methoden zijn beschreven in Bollen (1989).

In de praktijk zal men vaak moeite hebben, meetinstrumenten te maken die aan een van de genoemde definities van parallelliteit voldoen. Daarom heeft men, om de betrouwbaarheid en de standaardmeetfout van een meting X te schatten, methoden bedacht die geen gebruik maken van parallelle metingen. Een van die methoden bestaat eruit, de toets tweemaal af te nemen bij dezelfde personen. Andere methoden vereisen wel dat het mogelijk is het meetinstrument in stukken te verdelen. Bij toetsen die items bevatten, en ook als er diverse beoordelaars zijn, kan men spreken over onderdelen of deeltoetsen.

3.6.2 Test-hertestmethode

Als men niet kan beschikken over parallelvormen van een toets, kan men onder bepaalde omstandigheden dezelfde toets twee keer afnemen bij dezelfde personen. In feite beschouwt men de toets als parallel aan zichzelf. De procedure veronderstelt dat er geen leereffecten kunnen optreden tussen de twee toetsmomenten, en dat tussen die momenten in de populatie niet wezenlijk van karakter verandert. De betrouwbaarheid van de toets kan men dan eenvoudig schatten uit de correlatie tussen de twee verkregen toetsscores.

3.6.3 Toetsverlenging

Een van de methoden om de betrouwbaarheid te schatten, bestaat er uit het meetinstrument op de een of andere wijze in k parallelle delen te verdelen. Elk paar deeltoetsen heeft dezelfde correlatie ρ; deze correlatie is dan ook per definitie de betrouwbaarheid van elk der deeltoetsen. Deze betrouwbaarheid ρ wordt bekend verondersteld. In de praktijk kan dit het geval zijn als men een nieuwe toets wil samenstellen uit bestaande toetsen; een dergelijke samengestelde toets noemt men wel een verlengde toets. Als toetsscore op de verlengde toets kiest men de som van de scores op de deeltoetsen. Men kan dan het volgende afleiden. De geobserveerde variantie kan men als volgt schrijven:

$$\sigma_X^2 = \sigma^2\left(\sum_i^k X_i\right) = \sum_i^k \sigma_{X_i}^2 + \sum_{i\neq j} \mathrm{Cov}(X_i, X_j) = k\sigma_{X_i}^2 + \sum_{i\neq j} \sigma_{X_i}\sigma_{X_j}\rho =$$

$$= k\sigma_{X_i}^2 + k(k-1)\sigma_{X_i}^2\rho = k\sigma_{X_i}^2\left[1 + (k-1)\rho\right].$$

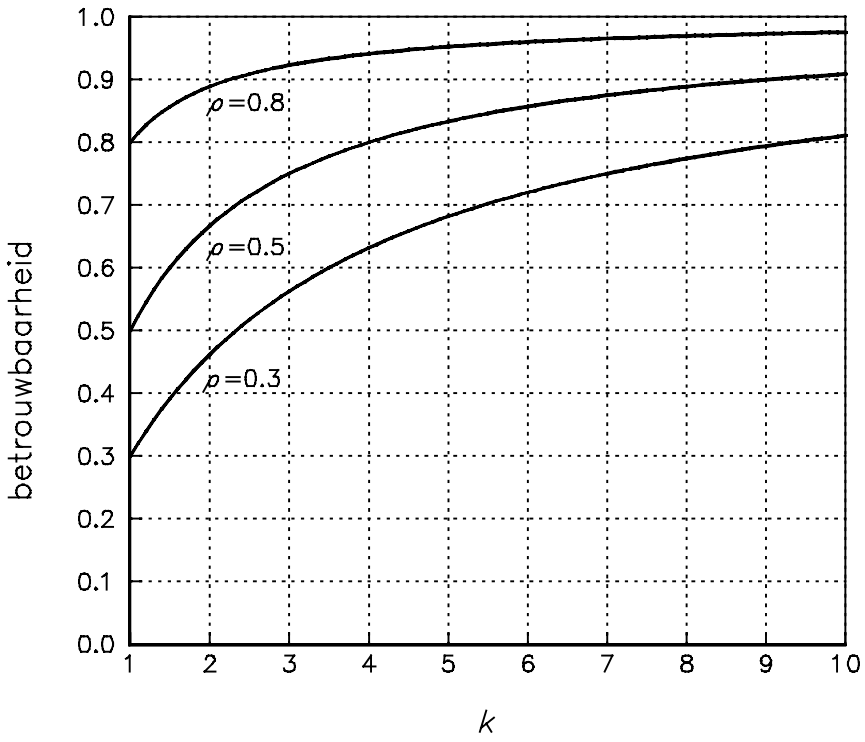
Evenzo kan men de ware variantie schrijven als:

$$\sigma_T^2 = \sigma^2\left(\sum_i^k T_i\right) = \sum_i^k \sigma_{T_i}^2 + \sum_{i\neq j} \mathrm{Cov}(T_i, T_j) = k\sigma_{T_i}^2 + k(k-1)\sigma_{T_i}^2 = k^2\sigma_{T_i}^2.$$

Als men deze twee uitdrukkingen substitueert in formule (3.4), verkrijgt men het volgende resultaat:

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} = \frac{k^2\sigma_{T_i}^2}{k\sigma_{X_i}^2\left[1 + (k-1)\rho\right]} = \frac{k\rho}{1 + (k-1)\rho}. \qquad (3.8)$$

Formule (3.8) is de Spearman-Brown-formule voor toetsverlenging (Brown, 1910; Spearman, 1910). Zij speelt een rol bij het samenstellen van toetsen uit gegeven deeltoetsen of items, vooral om te bepalen of men aan een toets in wording nog delen moet toevoegen om een bepaalde betrouwbaarheid te kunnen bewerkstelligen. In figuur 3.1 is voor een aantal waarden van ρ de betrouwbaarheid uitgezet tegen het aantal deeltoetsen k.



Figuur 3.1
Het verband tussen de lengte en de betrouwbaarheid van een toets

In de praktijk wordt de Spearman-Brown-formule voornamelijk gebruikt bij het construeren van toetsen. Een toets met k items blijkt een betrouwbaarheid ρ te hebben. Met behulp van de Spearman-Brown-formule kan men dan uitrekenen hoeveel maal men k items aan de toets moet toevoegen om een gewenste betrouwbaarheid $\rho' > \rho$ te bereiken.

3.6.4 Coëfficiënt alpha

De Spearman-Brown-formule veronderstelt dat men de betrouwbaarheid van de deeltoetsen kent. Aangezien dat in de praktijk dikwijls niet het geval is, kan men gebruik maken van de volgende ongelijkheid:

$$\rho_{XT}^2 \geq \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} \sigma_{X_i}^2}{\sigma_X^2}\right]. \qquad (3.9)$$

Het rechterlid van ongelijkheid (3.9) heet coëfficiënt alpha, of ook wel Cronbachs alpha (Cronbach, 1951). Merk op dat coëfficiënt alpha louter te schatten grootheden bevat. Met deze coëfficiënt is dus een ondergrens voor de betrouwbaarheid van een meetinstrument gegeven. De afleiding van coëfficiënt alpha bestaat uit een aantal stappen. In de eerste stap vormen we alle paren deeltoetsen, berekenen in elk paar de som van de ware varianties, en leiden voor de som van deze sommen een ongelijkheid af:

$$\sigma_{(T_i - T_j)}^2 = \sigma_{T_i}^2 + \sigma_{T_j}^2 - 2\,\mathrm{Cov}(T_i, T_j) \geq 0 \Rightarrow \sum_{i \neq j}\left[\sigma_{T_i}^2 + \sigma_{T_j}^2\right] \geq 2\sum_{i \neq j}\mathrm{Cov}(T_i, T_j).$$

De eerste ongelijkheid geldt omdat het linkerlid een variantie is, en dus nooit negatief kan zijn. In de tweede stap berekenen we opnieuw de som van sommen van ware varianties, maar nu met inbegrip van de oneigenlijke paren waarin elke deeltoets met zichzelf wordt gecombineerd. Voor de zo verkregen som leiden we weer een ongelijkheid af, waarbij de in de eerste stap afgeleide ongelijkheid wordt gebruikt:

$$\sum_i \sum_j \left[\sigma_{T_i}^2 + \sigma_{T_j}^2\right] = 2k\sum_i \sigma_{T_i}^2 = 2\sum_i \sigma_{T_i}^2 + \sum_{i \neq j}\left[\sigma_{T_i}^2 + \sigma_{T_j}^2\right] \geq$$

$$2\sum_i \sigma_{T_i}^2 + 2\sum_{i \neq j}\mathrm{Cov}(T_i, T_j) \Rightarrow (k-1)\sum_i \sigma_{T_i}^2 \geq \sum_{i \neq j}\mathrm{Cov}(T_i, T_j).$$

In de derde stap leiden we een eenvoudige ongelijkheid af voor de ware variantie:

$$\sigma_T^2 = \sigma^2\left(\sum_i T_i\right) = \sum_i \sigma_{T_i}^2 + \sum_{i \neq j}\mathrm{Cov}(T_i, T_j) \geq$$

$$\geq \frac{k}{k-1}\sum_{i \neq j}\mathrm{Cov}(T_i, T_j).$$

De som in het rechterlid van deze ongelijkheid kan als volgt worden herschreven:

$$\sum_{i \neq j}\mathrm{Cov}(T_i, T_j) = \sum_{i \neq j}\mathrm{Cov}(X_i, X_j) = \sigma_X^2 - \sum_i \sigma_{X_i}^2.$$

Als we alle ongelijkheden substitueren in formule (3.4), is het resultaat de volgende ongelijkheid:

$$\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_X} \geq \frac{k}{k-1}\left[1 - \frac{\sum_i \sigma^2_{X_i}}{\sigma^2_X}\right]. \qquad (3.10)$$

Als men coëfficiënt alpha beschouwt als een schatter van de betrouwbaarheid, kan men de standaardmeetfout schatten met: $\hat{\sigma}_E = \hat{\sigma}_X\sqrt{(1-\alpha)}$.

In het rechterlid van (3.10), dat gelijk is aan coëfficiënt alpha, ziet men de varianties optreden van de verschillende deeltoetsen. Er is niet verondersteld dat deze varianties aan elkaar gelijk zijn. In feite is het voldoende dat de deeltoetsen essentieel tau-equivalent zijn, als gedefinieerd in tabel 3.1.

Coëfficiënt alpha wordt wel een maat voor de interne consistentie van een toets genoemd. Men noemt een toets intern consistent als de items in de toets niet alle een correlatie van 0 met elkaar hebben. Men kan laten zien dat coëfficiënt alpha op de volgende manier kan worden geschreven:

$$\alpha = \frac{\bar{c}\langle X_i, X_j\rangle}{\sigma^2_{\bar{X}}}. \qquad (3.11)$$

In (3.11) is de teller, $\bar{c}\langle X_i, X_j\rangle$, gelijk aan het gemiddelde van de covarianties tussen alle paren itemscores: $\bar{c}(X_i, X_j) = [k(k-1)]^{-1}\sum_{i\neq j}\mathrm{Cov}(X_i, X_j)$. De noemer is gelijk aan de variantie van het gemiddelde van de itemscores: $\bar{X} = k^{-1}\sum_{i=1}^{k} X_i$. Als alle items onderling perfect correleren, zijn alle varianties van de itemscores aan elkaar gelijk, zijn de covarianties tussen de items gelijk aan deze varianties, en is de gemiddelde itemscore gelijk aan elk der itemscores. Uit (3.11) blijkt dat coëfficiënt alpha in dat geval gelijk is aan 1. Een enkele keer komt men in de literatuur de opvatting tegen dat een toets met een hoge interne consistentie, dus met een hoge waarde van coëfficiënt alpha, een enkele factor in de zin van de factoranalyse meet. Dat deze opvatting op een misverstand berust, is overtuigend aangetoond door Green en Lissitz (1977).

3.7 Toets- en itemanalyse

De toets- en itemanalyse is de praktische uitvoering van het schatten van de in de voorafgaande paragrafen beschreven grootheden. Aangezien in de praktijk toetsen

bestaan uit opgaven of items, worden ook kengetallen voor items berekend. Deze laatste grootheden spelen een belangrijke rol in het proces van toetsconstructie. Zij vormen niet alleen de bouwstenen van schattingsformules voor de betrouwbaarheid en de standaardmeetfout, maar zijn ook op zichzelf beschouwd van belang om eigenschappen van items te beschrijven. Doorgaans bepaalt men de kengetallen van items en toetsen in een proefafname: een concepttoets wordt aan een groep personen afgenomen, en op basis van de verkregen gegevens worden de grootheden van de items en de toets geschat. Zonodig worden er items herzien of wordt de samenstelling van de toets veranderd.

In deze paragraaf worden eerst de toets- en itemindices van een toets met meerkeuzevragen besproken. Daarna komen de indices van een toets met open vragen aan de orde voor zover deze niet besproken zijn bij de toets met meerkeuzevragen. In paragraaf 3.8 worden de betrouwbaarheid en de standaardmeetfout apart besproken. Omdat de toets- en itemindices veelal gebaseerd zijn op steekproeven, is paragraaf 3.9 gewijd aan standaardfouten van de geschatte toets- en itemindices. In paragraaf 3.10 tenslotte schenken we aandacht aan normen en richtlijnen voor diverse toets- en itemindices.

Aangezien er in een toets- en itemanalyse voortdurend sprake is van schattingen van grootheden op basis van de gegevens van een steekproef van personen, zal dikwijls de conventie worden gevolgd, de schatters aan te duiden met gewone letters. Zo zal een (schatter van de) variantie worden geschreven als s^2 en niet als $\hat{\sigma}^2$.

3.7.1 Toets- en itemindices bij toetsen met meerkeuzevragen

Toetsen met meerkeuzevragen bestaan uit vragen of items waarbij een persoon het goede antwoord moet kiezen uit verschillende alternatieven. We gaan er van uit dat elk goed beantwoord item 1 scorepunt oplevert en elk fout beantwoord item 0 scorepunten. De som van de itemscores vormt de toetsscore van een persoon. De toets- en itemindices worden besproken aan de hand van een toets die een tweekeuze-item en twee driekeuze-items bevat. De toets is door vier personen gemaakt. Dit is weliswaar geen realistische situatie maar het stelt de lezer in staat de indices na te rekenen. De itemantwoorden staan in tabel 3.2. In de kop van deze tabel zijn de goede antwoorden, samen wel de sleutel genoemd, vermeld. De itemantwoorden zijn met behulp van de sleutel omgezet in itemscores. Deze staan samen met de toetsscores in tabel 3.3.

Tabel 3.2

Antwoorden per persoon en per item (tussen haakjes de sleutel)

Tabel 3.3

Itemscores en toetsscores

persoon	item		
	1(B)	2(A)	3(C)
1	B	A	C
2	B	A	A
3	B	B	B
4	A	C	A

persoon	item			toetsscore
	1	2	3	
1	1	1	1	3
2	1	1	0	2
3	1	0	0	1
4	0	0	0	0
som	3	2	1	6

De resultaten van de toets- en itemanalyse van de gegevens uit tabel 3.3 staan in tabel 3.4. De indices uit deze tabel worden in de volgende deelparagraaf besproken.

Tabel 3.4

Resultaten toets- en itemanalyse van de toets met meerkeuzevragen

	p- en a-waarden			discriminatie-indices				r_{ir}- en r_{ar}-waarden		
item	A	B	C	s_i	r_{it}	r_{ir}	eff	A	B	C
1	0.25	0.75*		0.43	0.77	0.52	0.30	–0.52	0.52*	
2	0.50*	0.25	0.25	0.50	0.89	0.71	0.40	0.71*	0.00	–0.82
3	0.50	0.25	0.25*	0.43	0.77	0.52	0.30	–0.30	–0.17	0.52*

aantal personen : 4 gemiddelde p-waarde : 0.50
gemiddelde toetsscore : 1.50 betrouwbaarheid (KR-20) : 0.75
standaardafwijking : 1.12 standaardmeetfout : 0.56

3.7.2 Itemindices bij toetsen met meerkeuzevragen

In tabel 3.4 staan de waarden voor de moeilijkheid van een item en de aantrekkelijkheid van de afleiders onder de kop 'p- en a-waarden'. Bij elk alternatief is de fractie personen vermeld die het alternatief heeft gekozen. De fractie waarbij een ster (*) staat, hoort bij het goede antwoord en wordt de p-waarde van het item genoemd. De p-waarde wordt berekend door het aantal personen dat het item goed heeft, te delen door het aantal personen dat het item heeft gemaakt. De bij de afleiders of foute antwoorden vermelde fracties worden de a-waarden van het item genoemd en worden berekend door het aantal personen dat een afleider heeft gekozen te delen door het aantal personen dat het item heeft gemaakt. Bij item 2 in ons voorbeeld, een driekeuze-item, zien we bij de alternatieven A, B en C respectievelijk de waarden

0.50*, 0.25 en 0.25 staan. Dit betekent dat alternatief A het goede antwoord is met een p-waarde van 0.50. De a-waarden van de alternatieven B en C zijn beide gelijk aan 0.25.

Een p-waarde ligt per definitie tussen 0 en 1. Bij een p-waarde gelijk aan 0 hebben alle personen het item fout; bij een p-waarde gelijk aan 1 hebben alle personen het item goed. Het kan voorkomen dat een item een afleider heeft met een a-waarde die groter is dan de p-waarde. Dit kan er op wijzen dat een afleider niet fout is of dat het als goed bestempelde alternatief wellicht niet goed is. In het algemeen geeft een hoge a-waarde ons informatie over het item die in combinatie met andere informatie tot een definitief oordeel over de kwaliteit van het item moet leiden.

Onder het kopje 's_i' is de standaardafwijking van de items vermeld. De standaardafwijking van een item, s_i, wordt bij dichotome scores berekend als: $s_i = \sqrt{p\,q} = \sqrt{p(1-p)}$, waarin p de p-waarde van het item is en q gelijk is aan 1 - p. Wanneer alle personen een item goed dan wel fout hebben, is de standaardafwijking gelijk aan 0. De standaardafwijking is maximaal als p = 0.50, dus als de ene helft van de personen het item fout heeft en de andere helft het item goed. In dat geval is $s_i = \sqrt{0.5(1-0.5)} = 0.5$.

Omdat een item een onderdeel van een toets is, zijn er diverse indices ontwikkeld om de samenhang tussen een itemscore en de toetsscore weer te geven. Een index die veel gebruikt wordt is de r_{it}. De r_{it} is de produkt-moment-correlatie tussen de itemscore en de toetsscore. Deze correlatie wordt bij dichotoom gescoorde items wel puntbiseriële correlatie genoemd: het is de correlatie tussen een dichotome en een continu geachte variabele. Een produkt- moment-correlatie neemt waarden aan tussen +1 en -1. Een correlatie van +1 betekent dat er een perfect positief lineair verband bestaat tussen twee variabelen, in ons geval tussen de itemscore en de toetsscore. Dat de r_{it}-waarden in tabel 3.4 zo hoog zijn, heeft te maken met het feit dat de toets uit slechts drie items bestaat. Bij toetsen van veertig of meer items is een r_{it} van 0.50 al hoog (zie tabel 3.12).

De r_{it} wordt een discriminatie-index genoemd omdat zij aangeeft in hoeverre een item onderscheid maakt tussen personen met hoge toetsscores en personen met lage toetsscores. Een hoge r_{it} betekent dat veel personen met een hoge toetsscore het item goed hebben beantwoord en veel personen met een lage toetsscore het item fout hebben beantwoord. Later zullen we zien dat een hoge r_{it} ook betekent dat het item relatief veel bijdraagt aan de betrouwbaarheid van de toets (zie paragraaf 3.8.1).

Hiervoor zagen we dat de r_{it} een produkt-moment-correlatie is. Die kan met een van de algemene formules voor een correlatie berekend worden. Afgeleid kan worden dat voor dichotome scores de r_{it} van een item ook geschreven kan worden als:

$$r_{it} = \frac{\bar{x}_g - \bar{x}_f}{s_x} \sqrt{p(1-p)}\ , \tag{3.12}$$

waarin:
$\bar{x}_g$ = gemiddelde toetsscore van de personen die het item goed hebben,
$\bar{x}_f$ = gemiddelde toetsscore van de personen die het item fout hebben,
s_x = standaardafwijking van de toetsscores.

De teller in het deel voor het wortelteken in (3.12) maakt duidelijk waarom we de r_{it} een discriminatie-index noemen: hoe groter het verschil tussen $\bar{x}_g$ en $\bar{x}_f$, des te groter de r_{it}.

Naast de r_{it} is de r_{ir} een veel gebruikte discriminatie-index. De r_{ir} is een soortgelijke index als de r_{it}. Gaat het bij de r_{it} om de correlatie tussen itemscores en toetsscores, bij de r_{ir} gaat het om de correlatie tussen itemscores en restscores. De restscore van een persoon is gelijk aan zijn toetsscore minus de score op het desbetreffende item. Een persoon heeft dus evenzoveel restscores als er items zijn in de toets.

Zowel aan de r_{it} als aan de r_{ir} kleven bezwaren. De r_{it} geeft een geflatteerd beeld van de samenhang tussen de score op een item en de toetsscore, omdat de itemscore onderdeel is van de toetsscore. We correleren dus het item voor een deel met zichzelf. De r_{ir} ondervangt dit bezwaar, maar heeft als bezwaar dat de restscore waarmee een item gecorreleerd wordt, met het item varieert. De r_{ir}-waarden van eenzelfde toets zijn daardoor onderling niet te vergelijken. Als echter het aantal items in een toets veertig of meer is, zijn beide bezwaren van geen belang meer.

Nog een andere maat om het discriminerend vermogen van een item te karakteriseren is het effectieve gewicht dat te vinden is onder het kopje 'eff'. Onder het effectieve gewicht verstaan we de bijdrage van een item aan de spreiding van toetsscores. Hoe hoger het effectieve gewicht van een item is, des meer spreiding in de toetsscores toegeschreven kan worden aan het item. Het volgende kan worden afgeleid (Gulliksen, 1950; Ferguson & Takane, 1989):

$$\sum_{i=1}^{k} r_{it} s_i = s_x, \tag{3.13}$$

waarin k het aantal items is.

Het effectieve gewicht van item i is gedefinieerd als:

$$\frac{r_{it} \times s_i}{s_x}. \tag{3.14}$$

De teller in (3.14) wordt de itembetrouwbaarheidsindex genoemd en is een onderdeel van de formule om de betrouwbaarheid van de toets te schatten (zie paragraaf 3.8.1). Uit (3.14) volgt dat de som van de effectieve gewichten gelijk is aan 1. In ons voorbeeld van tabel 3.4 heeft item 2 een effectief gewicht van 0.40; dat betekent dat het item voor 40% bijdraagt aan de standaardafwijking van de toetsscores. Een andere interpretatie van het effectieve gewicht wordt gegeven door regressie-analyse. Als men de lineaire regressievergelijking van de itemscore op de toetsscore opstelt, blijkt de regressiecoëfficiënt gelijk te zijn aan het effectieve gewicht van het item.

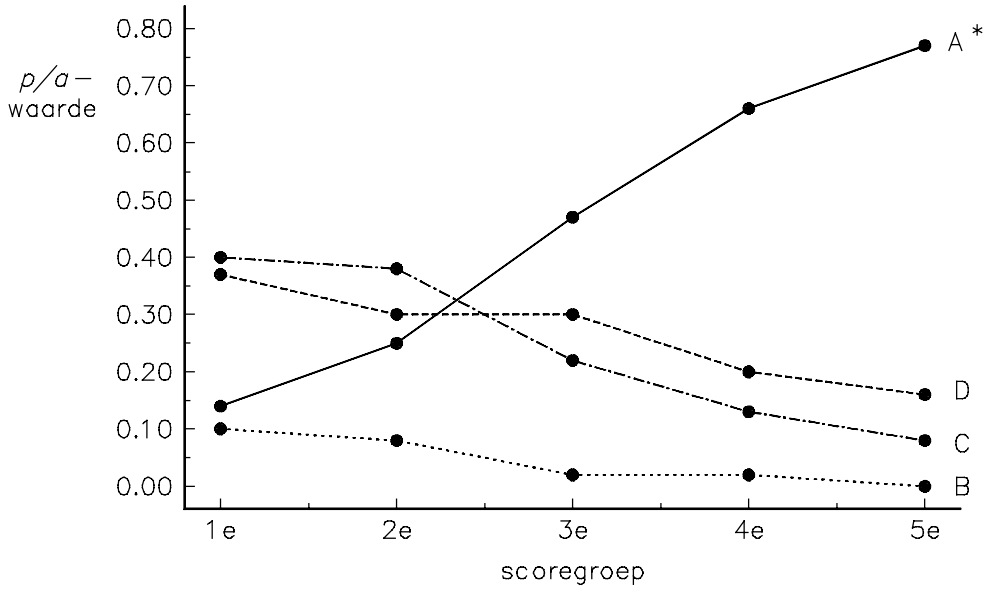
Bij een toets met meerkeuzevragen is het mogelijk, naast een discriminatie-index voor het goede antwoord discriminatie-indices voor de afleiders (foute antwoorden) te berekenen. In tabel 3.4 kunnen we zien dat er bij elk item r_{ar}-waarden zijn vermeld naast de r_{ir}-waarde. Per item zijn er uiteraard evenveel r_{ar}-waarden als er afleiders zijn. De r_{ar} wordt berekend door personen die het desbetreffende foute antwoord hebben gekozen een itemscore 1 en de anderen een itemscore 0 te geven. Vervolgens wordt de correlatie tussen het foute antwoord en de restscore berekend, waarbij de restscore per definitie dezelfde waarde heeft als bij de berekening van de r_{ir}. Omdat we toetsen met een hoge betrouwbaarheid nastreven, zijn items met positieve r_{ir}- en negatieve r_{ar}-waarden gewenst. Zulke waarden impliceren dat relatief veel personen met een hoge toetsscore het item goed hebben beantwoord en relatief veel personen met een lage toetsscore het item fout hebben beantwoord. Een positieve r_{ar} geeft aan dat relatief veel goede personen de desbetreffende afleider als het goede antwoord hebben aangemerkt. Soms kan dit een sleutelfout zijn: de verkeerde sleutel is per ongeluk opgegeven of bij nader inzien blijkt dat de afleider met de positieve r_{ar} het goede antwoord is.

Tabel 3.5

Per scoregroep de p- en a-waarden van een item

score	n	A*	B	C	D
0 - 18	123	0.14	0.10	0.40	0.37
19 - 22	124	0.25	0.08	0.38	0.30
23 - 29	124	0.47	0.02	0.22	0.30
30 - 35	124	0.66	0.02	0.13	0.20
36 - 47	124	0.77	0.00	0.08	0.16
0 - 47	619	0.46*	0.04	0.24	0.26
gem. score	26.0	30.8	18.8	21.0	23.5

Het discriminerend vermogen van een item kunnen we ook weergeven door de personen in een aantal scoregroepen op te delen en vervolgens per scoregroep de *p*- en *a*-waarden te berekenen. Als voorbeeld presenteren we in tabel 3.5 van een item de *p*- en *a*-waarden per scoregroep. In die tabel lezen we dat alternatief A het goede antwoord is met een *p*-waarde van 0.46. Van de afleiders is D het meest aantrekkelijk met een *a*-waarde van 0.26. Verder zien we dat de totale groep van 619 personen is opgesplitst in vijf bijna even grote scoregroepen. Bekijken we nu van het item de *p*-waarde per scoregroep, dan heeft het item in de minst vaardige groep, met scores tussen 0 en 18, een *p*-waarde van 0.14. De *p*-waarde van het item wordt groter met het vaardiger worden van de groep, en in de meest vaardige groep heeft het item een *p*-waarde van 0.77. Bij de afleiders is de tendens andersom; hoe vaardiger de groep, des te lager de *a*-waarde. Het item is dus een voorbeeld van een goed discriminerend item: de *p*-waarde van het item is in de groep van de beste personen veel hoger dan in de groep van de slechtste personen, en de *a*-waarden van het item zijn voor de slechtste personen hoger dan de *a*-waarden voor de beste personen. De *p*- en *a*-waarden uit tabel 3.5 zijn grafisch weergegeven in figuur 3.2. De keuze van het aantal scoregroepen is arbitrair. Om er echter voor te zorgen dat de standaardfout van een fractie niet te groot wordt, moet het aantal personen per scoregroep niet te klein zijn (zie tabel 3.8).



Figuur 3.2
Per scoregroep *p*- en *a*-waarden van het item uit tabel 3.5

3.7.3 Toetsindices bij toetsen met meerkeuzevragen

Behalve informatie over de drie afzonderlijke items uit de toets, bevat tabel 3.4 ook informatie die betrekking heeft op de toets als geheel. We kunnen in de tabel lezen dat vier personen, n = 4, de toets gemaakt hebben. Een maat voor de moeilijkheidsgraad van een toets is de gemiddelde toetsscore $\bar{x}$, die bij deze toets gelijk is aan 6/4=1.50. De standaardafwijking van de toetsscores, s_x, is een maat voor de spreiding van de toetsscores en kan als volgt berekend worden:

$$s_x = \left(\frac{\sum_{v=1}^{n} (x_v - \bar{x})^2}{n} \right)^{1/2}, \qquad (3.15)$$

waarin x_v de toetsscore is van persoon v.

De standaardafwijking kan volgens (3.13) ook verkregen worden door de itembetrouwbaar-
heidsindices te sommeren. Wanneer de standaardafwijking gelijk is aan 0, hebben alle personen dezelfde toetsscore. De standaardafwijking is maximaal wanneer de ene helft van de personen alle items goed heeft en de andere helft alle items fout.

De gemiddelde p-waarde, $\bar{p}$, is het gemiddelde van de p-waarden van de afzonderlijke items. Bij toetsen met meerkeuzevragen kan de gemiddelde p-waarde berekend worden hetzij door alle p-waarden op te tellen en de som te delen door het aantal items k, hetzij door de gemiddelde toetsscore te delen door het aantal items in de toets. In formulevorm:

$$\bar{p} = \frac{\sum_{i=1}^{k} p_i}{k} \text{ of } \bar{p} = \frac{\bar{x}}{k}. \qquad (3.16)$$

De toetsindices betrouwbaarheid en standaardmeetfout worden in paragraaf 3.8 besproken.

3.7.4 Toets- en itemindices bij toetsen met open vragen

Bij toetsen met open vragen moeten personen zelf het antwoord formuleren op de vragen die voorgelegd worden. Het is gebruikelijk dat er per vraag meer dan een

scorepunt behaald kan worden en dat de antwoorden door beoordelaars met behulp van een correctievoorschrift gescoord worden. In deze paragraaf gaan we er van uit dat beoordelaars geen factor zijn die de meetprocedure verstoren. In dat geval is er ook geen wezenlijk verschil tussen de analyse van een toets met open vragen en de analyse van een toets met meerkeuzevragen. Het enige verschil is dat er bij open vragen andere itemscores dan alleen maar 0 en 1 mogelijk zijn. Indien beoordelaars wel een storende factor zijn, dient er een analyse als beschreven in paragraaf 3.13 plaats te vinden.

In het voorbeeld in tabel 3.6 gaan we uit van vier open vragen die door zes personen beantwoord zijn. Op elke vraag kunnen maximaal twintig punten behaald worden.

Tabel 3.6

Itemscores en toetsscores

persoon	item				toetsscore
	1	2	3	4	
1	17	8	14	3	42
2	16	10	13	5	44
3	18	15	14	18	65
4	16	14	14	8	52
5	14	7	7	4	32
6	17	15	17	16	65
som	98	69	79	54	300

De resultaten van de toets- en itemanalyse staan in tabel 3.7. Aangezien de toets- en itemanalyse van open vragen voor een deel dezelfde indices bevat als de toets- en itemanalyse van meerkeuzevragen, komen hierna niet meer alle toets- en itemindices aan de orde. Alleen de voor open vragen specifieke indices worden besproken.

Tabel 3.7

Resultaten van de toets- en itemanalyse van de toets met open vragen

item	max. score	gem. score	p'	s_i	r_{it}	r_{ir}	eff
1	20.00	16.33	0.82	1.25	0.81	0.77	0.08
2	20.00	11.50	0.58	3.30	0.95	0.91	0.26
3	20.00	13.17	0.66	3.02	0.81	0.69	0.20
4	20.00	9.00	0.45	5.89	0.94	0.79	0.46

aantal personen : 6 gemiddelde p'-waarde : 0.63

gemiddelde toetsscore	: 50.00	betrouwbaarheid (alpha)	: 0.82
standaardafwijking	: 12.10	standaardmeetfout	: 5.12

3.7.5 Itemindices bij toetsen met open vragen

Bij een toets met open vragen kan het aantal te behalen scorepunten van vraag tot vraag variëren. Daarom is in tabel 3.7 een kolom met het opschrift 'max. score' opgenomen. In deze kolom staat het aantal punten dat op een item behaald kan worden. In het voorbeeld zijn bij alle items de maxima gelijk.

Een andere voor open vragen specifieke index staat in de kolom met opschrift 'gem. score'. In deze kolom staat de gemiddelde score die op elk van de items behaald is. Bij ongelijke maximale scores zijn de gemiddelde itemscores niet vergelijkbaar. Daarom wordt de p'-waarde berekend; deze staat in de kolom met het opschrift 'p''. De p'-waarde duidt de moeilijkheidsgraad van een item aan, en wordt berekend door de gemiddelde itemscore te delen door de maximale itemscore. Merk op dat we bij open vragen over de p'-waarde spreken en bij meerkeuzevragen over de p-waarde. De definitie van de twee grootheden is gelijk; het verschil in notatie heeft geen andere functie dan aan te geven om welke soort vraag het gaat.

3.7.6 Toetsindices bij toetsen met open vragen

Bij toetsen met open vragen worden dezelfde toetsindices berekend als bij toetsen met meerkeuzevragen. Om misverstanden te voorkomen, verdient de berekening van de gemiddelde p'-waarde enige toelichting. De gemiddelde p'-waarde wordt berekend door de gemiddelde toetsscore te delen door de maximaal te behalen toetsscore. In tegenstelling tot bij een toets met meerkeuzevragen mag de gemiddelde p'-waarde bij een toets met open vragen alleen maar op deze manier berekend worden en niet via de p'-waarden van de individuele vragen. Als men dat wel zou doen, zou men verschillen in maximaal te behalen itemscores veronachtzamen.

3.8 Betrouwbaarheid en standaardmeetfout

Bij de toets- en itemanalyse van de meerkeuzevragen is de KR-20 als betrouwbaarheidsmaat berekend en bij de toets- en itemanalyse van de open vragen coëfficiënt alpha. Hierna laten we zien dat de KR-20 een speciaal geval is van coëfficiënt alpha. In paragraaf 3.5 zijn twee manieren besproken om met behulp van de standaardmeetfout een intervalschatting voor de ware score te bepalen. Deze twee manieren worden in paragraaf 3.8.3 gebruikt om intervalschattingen te verkrijgen voor ware verschilscores.

3.8.1 Coëfficiënt alpha en de KR-20

Het is gebruikelijk, de betrouwbaarheid van een toets met coëfficiënt alpha te schatten. De formule voor coëfficiënt alpha is gegeven in het rechterlid van (3.9). Omdat bij dichotoom gescoorde vragen geldt dat $s_i^2 = p_i q_i$, kan coëfficiënt alpha voor dichotoom gescoorde items geschreven worden als:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} p_i q_i}{s_x^2}\right]. \qquad (3.17)$$

Formule (3.17) staat bekend als de KR-20 en is onafhankelijk van Cronbachs coëfficiënt alpha door Kuder en Richardson (1937) ontwikkeld. Vanwege (3.12) kan coëfficiënt alpha ook geformuleerd worden als:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} s_i^2}{\left(\sum_{i=1}^{k} r_{it} s_i\right)^2}\right]. \qquad (3.18)$$

Uit (3.18) laat zich het verband tussen de r_{it} en de betrouwbaarheid nog niet eenvoudig aflezen. Bij dichotoom gescoorde items liggen de itemvarianties in de praktijk tussen 0.21 en 0.25 ($0.3 < p < 0.7$). Indien we de itemvarianties nu als constant beschouwen voor alle items, kunnen we afleiden (Thorndike, 1982):

$$\alpha \approx \frac{k}{k-1}\left(1 - \frac{1}{k(\bar{r}_{it})^2}\right), \qquad (3.19)$$

waarin $\bar{r}_{it}$ het gemiddelde van de r_{it}-waarden is.

3.8.2 Verschilscores

In paragraaf 3.5 zijn schattingen van de ware score aan de orde geweest. Er is op gewezen dat het schatten van ware scores niet altijd nodig is. In de praktijk zou men willen weten of een toetsscore van 30 voor Kay en een toetsscore van 33 voor Wilko betekent dat de laatstgenoemde meer weet dan Kay. Daar kan men niet achter komen, omdat men de ware scores van Kay en Wilko niet kent. Wel kan men iets zeggen over het volgende probleem. Als men aselect twee personen uit de populatie trekt waarvan de waargenomen scores drie punten verschillen, kan men dan zeggen of dit verschil substantieel is? Statistisch gezien betekent dit dat we de nulhypothese willen toetsen dat de ware toetsscores van de twee aselect getrokken personen gelijk zijn. Noem deze ware scores τ_1 en τ_2, en de geobserveerde scores x_1 en x_2. Veronderstel dat de geobserveerde scores x_1 en x_2 normaal verdeeld zijn met verwachte waarden τ_1 respectievelijk τ_2, en beide met standaardafwijking σ_E. Dan is de verschilscore $x_1 - x_2$ normaal verdeeld met gemiddelde $\tau_1 - \tau_2$ en standaardafwijking $\sigma_E\sqrt{2}$. Naar analogie van (3.6) kunnen we een intervalschatting maken van het verschil $\delta = \tau_1 - \tau_2$. Dit interval bestaat uit alle waarden $\hat{\delta}$ waarvoor de volgende nulhypothese niet wordt verworpen:

$$H_0\text{: } (x_1 - x_2) - z \times \sigma_E\sqrt{2} \leq \hat{\delta} \leq (x_1 - x_2) + z \times \sigma_E\sqrt{2}\,.$$

Veronderstel dat de toets een standaardmeetfout σ_E heeft van 1, dan vindt men, bij een verschil van drie punten in geobserveerde scores, het 95%-betrouwbaarheidsinterval: $0.23 \leq \tau_1 - \tau_2 \leq 5.77$. Aangezien dit interval niet de waarde 0 bevat, zal men bij een waargenomen verschil van drie punten, de hypothese verwerpen dat de bijbehorende ware scores aan elkaar gelijk zijn.

Men kan ook een intervalschatting voor verschilscores bepalen op basis van de in paragraaf 3.5 genoemde Kelley-schatter. Men kan afleiden dat de verschilscore $\delta = \tau_1 - \tau_2$ een verwachte waarde heeft gelijk aan $\rho^2_{XT}(x_1 - x_2)$ en een standaardafwijking gelijk aan $(2\rho^2_{XT}\sigma^2_E)^{½}$. Voor een toets met een betrouwbaarheid van 0.80 en een standaardmeetfout van 1 is, bij een verschil in waargenomen scores van 3 punten, het 95%-betrouwbaarheidsinterval gelijk aan: $-0.08 \leq \tau_1 - \tau_2 \leq 4.88$. Nu zal men de nulhypothese van gelijke ware scores niet verwerpen. Merk op dat het laatst

gegeven betrouwbaarheidsinterval iets kleiner is dan het eerst gegeven interval: 4.96 tegenover 5.54.

3.9 Nauwkeurigheid van toets- en itemindices

Bij het berekenen van toets- en itemindices is het buitengewoon belangrijk dat men er zich rekenschap van geeft hoe nauwkeurig die indices geschat zijn. De statistiek geeft ons op deze vraag een antwoord omdat het mogelijk is betrouwbaarheidsintervallen te construeren. Zoals reeds eerder is aangegeven, is een betrouwbaarheidsinterval een stochastisch interval om een steekproefwaarde dat met een gegeven kans de te schatten populatiewaarde bevat. De p-waarde, de gemiddelde score, de r_{it}-waarde, de KR-20 en coëfficiënt alpha zijn allemaal voorbeelden van grootheden die gebaseerd zijn op steekproeven en daardoor behept met steekproeffouten. In de volgende paragrafen zullen we op deze steekproeffouten en op de constructie van betrouwbaarheidsintervallen ingaan.

3.9.1 Standaardfout van een p-waarde

De standaardfout s_p van een p-waarde wordt met de volgende formule berekend:

$$s_p = \left(\frac{p(1-p)}{n}\right)^{½}. \qquad (3.20)$$

In (3.20) staat n voor het aantal personen in de aselect getrokken steekproef. Nu zegt een vuistregel in de statistiek dat, indien $n > \{9 \times (1-p)/p\}$ bij $p \leq 0.50$ en $n > \{9 \times p/(1-p)\}$ bij $p \geq 0.50$, een p-waarde bij benadering normaal verdeeld is. Hiervan uitgaande, kunnen we een betrouwbaarheidsinterval construeren voor de werkelijke p-waarde. Veronderstel dat de geschatte p-waarde van een item 0.20 is en dat het item door 100 personen is gemaakt, dan is de bijbehorende standaardfout $\sqrt{0.2 \times 0.8/100} = 0.04$. We kunnen dan bijvoorbeeld de grenzen van het 95%-betrouwbaarheidsinterval berekenen. Uit de berekening volgt dat in 95% van de gevallen bij items met een geschatte p-waarde van 0.20 de werkelijke p-waarde tussen 0.12 en 0.28 zal liggen (0.12 = 0.20 - 1.96 × 0.04 en 0.28 = 0.20 +1.96 × 0.04). In tabel 3.8, die gebaseerd is op exacte berekeningen (De Jonge, 1963), kan men bij $p = 0.20$ en $n = 100$ aflezen dat de grenzen 0.13 en 0.29 zijn. De afwijkingen zijn minimaal.

Tabel 3.8

95%-betrouwbaarheidsintervallen voor fracties

steekproef -fractie p	aantal personen in de steekproef (n)									
	50		100		200		500		1000	
0.00	0.00	0.07	0.00	0.04	0.00	0.02	0.00	0.01	0.00	0.00
0.10	0.03	0.22	0.05	0.18	0.06	0.15	0.08	0.13	0.08	0.12
0.20	0.10	0.34	0.13	0.29	0.15	0.26	0.17	0.24	0.18	0.23
0.30	0.18	0.45	0.21	0.40	0.24	0.37	0.26	0.34	0.27	0.33
0.40	0.26	0.55	0.30	0.50	0.33	0.47	0.36	0.45	0.37	0.43
0.50	0.35	0.65	0.40	0.60	0.43	0.57	0.46	0.55	0.47	0.53
0.60	0.45	0.74	0.50	0.70	0.53	0.67	0.55	0.64	0.57	0.63
0.70	0.55	0.82	0.60	0.79	0.63	0.76	0.66	0.74	0.67	0.73
0.80	0.66	0.90	0.71	0.87	0.74	0.85	0.76	0.83	0.77	0.82
0.90	0.78	0.97	0.82	0.95	0.85	0.94	0.87	0.92	0.88	0.92
1.00	0.93	1.00	0.96	1.00	0.98	1.00	0.99	1.00	1.00	1.00

3.9.2 Standaardfout van een gemiddelde toetsscore en van een p'-waarde

De standaardfout $s_{\bar{x}}$ van de gemiddelde toetsscore $\bar{x}$ is gelijk aan:

$$s_{\bar{x}} = \frac{s_x}{\sqrt{n}}. \tag{3.21}$$

Neem als voorbeeld een toets die door 429 personen gemaakt is, en waarvan de gemiddelde toetsscore gelijk is aan 32.24 en de standaardafwijking van de toetsscores 6.29 is. De standaardfout bedraagt dan 0.30 en het 95%-betrouwbaarheidsinterval heeft de grenzen 31.64 en 32.84.

De standaardfout s_p van een p'-waarde is gelijk aan:

$$s_{p'} = \frac{s_i}{m\sqrt{n}}. \tag{3.22}$$

In (3.22) staat m voor de maximaal te behalen score op de vraag. Bij de toets met open vragen in tabel 3.7 heeft item 4 een p'-waarde van 0.45. We kunnen daarvan de standaardfout berekenen; deze bedraagt 0.12. Het 95%-betrouwbaarheidsinterval voor de werkelijke p'-waarde heeft de grenzen 0.14 en 0.76. Dit interval is groot omdat zo weinig personen het item gemaakt hebben.

3.9.3 Standaardfout van een r_{it}-waarde

De berekening van de standaardfout van een r_{it}-waarde is nogal gecompliceerd. In Iker en Perry (1960) staan benaderingsformules en tabellen voor de standaardfout.

Tabel 3.9

95%-betrouwbaarheidsintervallen voor r_{it}-waarden

r_{it}-waarde (steekproef)	aantal personen in de steekproef (n)							
	100		200		500		1000	
0.00	-0.20	0.20	-0.14	0.14	-0.08	0.08	-0.06	0.06
0.10	-0.10	0.30	-0.04	0.24	0.02	0.18	0.04	0.16
0.20	0.00	0.40	0.06	0.34	0.12	0.28	0.14	0.26
0.30	0.12	0.48	0.18	0.42	0.22	0.38	0.24	0.36
0.40	0.24	0.56	0.28	0.52	0.32	0.48	0.34	0.46
0.50	0.36	0.64	0.40	0.60	0.44	0.56	0.46	0.54
0.60	0.48	0.72	0.51	0.69	0.54	0.66	0.56	0.64

Tabel 3.9 is gebaseerd op Iker en Perry, en is van toepassing op p-waarden die tussen 0.20 en 0.80 liggen. In tabel 3.9 staan voor diverse waarden van de r_{it} en n de 95%-betrouwbaarheids- intervallen voor de werkelijke waarden van de r_{it} vermeld. Indien bijvoorbeeld bij een toets- en itemanalyse die gebaseerd is op 1000 personen, de r_{it}-waarde van een item 0.20 is, dan zijn de 95%-betrouwbaarheidsgrenzen van de werkelijke r_{it}-waarde 0.14 en 0.26.

3.9.4 Standaardfout van coëfficiënt alpha

Voor coëfficiënt alpha heeft Feldt (1965) de steekproefverdeling afgeleid waarop tabel 3.10 gebaseerd is. In deze tabel zijn bij diverse steekproefwaarden van coëfficiënt alpha de onder- en bovengrenzen vermeld van het 95%-betrouwbaarheidsinterval voor de werkelijke waarde van coëfficiënt alpha. De tabel mag alleen gebruikt worden indien een toets tien of meer vragen bevat. Als bijvoor- beeld de betrouwbaarheid van een toets die is afgenomen bij 500 personen gelijk is aan 0.70, dan loopt het 95%-betrouwbaarheidsinterval van 0.66 tot 0.74.

Tabel 3.10

95%-betrouwbaarheidsintervallen voor coëfficiënt alpha

α (steekproef)	aantal personen in de steekproef (n)							
	100		200		500		1000	
0.10	-0.17	0.33	-0.09	0.27	-0.02	0.21	0.02	0.18
0.20	-0.04	0.41	0.03	0.35	0.10	0.30	0.13	0.27
0.30	0.09	0.48	0.25	0.43	0.21	0.38	0.24	0.30
0.40	0.22	0.55	0.27	0.51	0.32	0.47	0.35	0.45
0.50	0.35	0.63	0.40	0.59	0.44	0.56	0.45	0.54
0.60	0.48	0.70	0.52	0.67	0.55	0.65	0.56	0.63
0.70	0.61	0.78	0.64	0.76	0.66	0.74	0.67	0.73
0.80	0.74	0.85	0.76	0.84	0.77	0.82	0.78	0.82
0.90	0.87	0.93	0.88	0.92	0.89	0.91	0.89	0.91

3.10 Normen voor toets- en itemindices

In de volgende paragrafen worden normen en richtlijnen voor toets- en itemindices geformuleerd. We moeten bedenken dat deze normen en richtlijnen opgesteld zijn met de gedachte dat we er naar moeten streven een toets met een zo hoog mogelijke betrouwbaarheid te construeren. Nogmaals dient er op gewezen te worden dat de indices bij kleine aantallen personen een relatief kleine precisie hebben, zodat voorzichtigheid geboden is bij de interpretatie van zulke indices.

3.10.1 Normen voor p- en p'-waarden

In de literatuur vinden we verschillende opvattingen over de optimale p-waarde van een item. Crocker en Algina (1986) stellen dat de optimale p-waarde halverwege de raadkans en 1.0 moet liggen. De veronderstelling hierbij is dat er geraden wordt als men niet weet wat het goede antwoord op een meerkeuze-item is. In formulevorm uitgedrukt: $p = 0.5 + 0.5/m$, waarin m het aantal alternatieven is en p de gewenste p-waarde. Naar aanleiding van een simulatie-onderzoek komt Lord (1952) tot een andere conclusie. De aanbevelingen van voornoemde auteurs over de optimale p-waarde van items met verschillende aantallen alternatieven staan in tabel 3.11.

De conclusie van een onderzoek van Feldt (1993) is, dat de optimale p-waarde tussen 0.57 en 0.67 moet liggen wanneer er geraden kan worden. Indien er geen reden is om aan te

Tabel 3.11

Optimale p-waarde bij items met 2-5 alternatieven

aantal alternatieven	optimale p-waarde (p=0.5+0.5/m)	optimale p-waarde (Lord)
2	0.75	0.85
3	0.67	0.77
4	0.63	0.74
5	0.60	0.70

nemen dat er geraden wordt, of als er niet geraden kan worden zoals bij open vragen, is de
optimale p-waarde gelijk aan 0.50. Het effect van de moeilijkheid van een item op de betrouwbaarheid blijkt echter verbazingwekkend klein te zijn, zelfs als de p-waarden variëren van 0.27 tot 0.79.

3.10.2 Normen voor r_{it}-waarden

Ook voor r_{it}-waarden vindt men in de literatuur geen absolute normen. Zoals bekend kan een produkt-moment-correlatie, dus ook een r_{it}-waarde, variëren tussen -1 en +1. Een r_{it}-waarde van 0.50 en hoger is echter in de praktijk bij toetsen met meer dan veertig items al erg hoog. Ebel en Frisbie (1986) komen tot de in tabel 3.12 vermelde normen voor de r_{it}-waarden.

Tabel 3.12

Normen voor r_{it}-waarden

r_{it}-waarde	itembeoordeling
0.40 en hoger	zeer goed
0.30 - 0.39	goed
0.20 - 0.29	twijfelachtig
0.19 en lager	slecht

Omdat de grootte van de r_{it} onder andere afhankelijk is van het aantal items in een toets, moet men strikt genomen bovenstaande normen alleen hanteren bij r_{it}-waarden die gecorrigeerd zijn voor toetslengte. De correctie kan uitgevoerd worden met een correctie-formule van Henrysson (1963). Vanwege het geringe effect kan de correctie achterwege blijven indien de items afkomstig zijn uit toetsen met veertig of meer items.

3.10.3 Normen voor de betrouwbaarheid

In de literatuur wordt 0.85 als vereiste ondergrens voor de betrouwbaarheid van een toets genoemd wanneer de vaardigheid van een groep personen op basis van slechts een enkele toets wordt bepaald. Wanneer de vaardigheid met meer toetsen of op verschillende momenten wordt getoetst zijn lagere ondergrenzen acceptabel, waarbij in de literatuur 0.65 wel als gewenste ondergrens wordt genoemd (Frisbie, 1988).

Een mogelijke norm voor de betrouwbaarheid zouden we kunnen ontlenen aan het percentage ten onrechte gezakte en ten onrechte geslaagde personen, ofwel het percentage niet-consistente beslissingen, bij een selectietoets (Dousma & Horsten, 1989). Met de ten onrechte gezakte en de ten onrechte geslaagde personen bedoelen we de personen waarvoor, indien ze een parallelle toets hadden afgelegd, de beslissing anders geweest had kunnen zijn. Het percentage niet-consistente beslissingen neemt toe als de betrouwbaarheid lager wordt en ook als het percentage gezakten stijgt, waarbij het percentage gezakten afhangt van de cesuur of grensscore. Tabel 3.13 laat de percentages niet-consistente beslissingen zien als functie van het percentage gezakten en van de betrouwbaarheid. Daarbij moet opgemerkt worden dat het gebruik van de tabel alleen zinvol is wanneer de toetsscores ongeveer normaal verdeeld zijn.

Tabel 3.13

Percentages niet-consistente beslissingen als functie van het percentage gezakten en de betrouwbaarheid

percentage gezakten	betrouwbaarheid						
	0.0	0.50	0.60	0.70	0.80	0.90	1.00
5	10	8	7	6	5	4	0
10	18	14	12	11	9	6	0
15	26	18	17	14	12	8	0
20	32	23	20	17	14	10	0
25	38	26	23	20	16	11	0
30	42	29	25	22	18	12	0
35	46	31	27	23	19	13	0
40	48	32	29	24	20	14	0
45	50	33	29	25	20	14	0
50	50	33	30	25	20	14	0

In tabel 3.13 kunnen we zien dat bij een toets met een betrouwbaarheid van 0.80 en met een percentage gezakten van 30, het percentage niet-consistente beslissingen gelijk aan 18 is. Dat wil dan zeggen dat 9% van de gezakten tot de geslaagden zou kunnen hebben behoord en 9% van de geslaagden tot de gezakten. Dus voor 18% van alle personen had de beslissing anders kunnen zijn.

3.11 Generaliseerbaarheidstheorie

De bespreking van de generaliseerbaarheidstheorie, (Cronbach, Gleser, Nanda & Rajaratnam, 1972), in dit hoofdstuk bestaat uit vier paragrafen. Het begrippenkader dat in de generaliseerbaarheidstheorie gehanteerd wordt en dat in belangrijke mate ontleend is aan de variantie-analytische literatuur, wordt in deze paragraaf besproken. In paragraaf 3.12 wordt de generaliseerbaarheidstheorie behandeld aan de hand van de analyse van de toets met meerkeuzevragen die in paragraaf 3.7 met de klassieke testtheorie geanalyseerd is. In paragraaf 3.13 wordt de generaliseerbaarheidstheorie verder toegelicht aan de hand van een analyse van een toets waarbij beoordelaars de antwoorden van personen op vragen beoordelen. In beide paragrafen wordt aandacht besteed aan verschillen tussen de klassieke testtheorie en generaliseerbaarheidstheorie. In paragraaf 3.14 komen kort een aantal andere aspecten van de generaliseerbaarheidstheorie aan de orde. Merk op dat de notatie die in de paragrafen 3.11 tot en met 3.14 gehanteerd wordt afwijkt van die uit voorgaande paragrafen. De reden hiervoor, is de notatie aan te laten sluiten bij de in de literatuur gebruikelijke notatie.

In de generaliseerbaarheidstheorie worden observaties of metingen beschreven in termen van de condities waaronder zij geobserveerd worden. Condities van een bepaalde soort worden aangeduid als 'facet'. De dertig meerkeuzevragen van een toets zijn volgens deze terminologie de dertig condities van het facet 'vragen'. En bij een toets bestaande uit tien open vragen waarbij de antwoorden door twee beoordelaars beoordeeld worden, spreken we over de tien condities van het facet 'vragen' en de twee condities van het facet 'beoordelaars'. Het door personen laten beantwoorden van vragen, kunnen we opvatten als een gestandaardiseerd experiment (Meerling, 1981). Een proefopzet waarin responsen of antwoorden van personen op (condities van het facet) vragen worden geobserveerd, wordt een een-facet-design genoemd. Een proefopzet waarin de observaties beoordelingen zijn van responsen van personen op (condities van het facet) vragen die beoordeeld worden door (condities van het facet) beoordelaars, wordt een twee-facet-design genoemd. Het aantal observaties dat per

persoon verkregen wordt, is afhankelijk van het design dat gebruikt wordt. Wanneer we aan tien personen een toets van dertig vragen voorleggen, een zogenaamd gekruist een-facet-design (personen × vragen), hebben we per persoon dertig observaties. Zouden we echter aan elke persoon drie andere vragen voorleggen, dan hebben we per persoon slechts drie observaties. Wanneer we aan tien personen een toets van tien vragen voorleggen en de responsen op de tien vragen laten beoordelen door twee beoordelaars, een zogenaamd gekruist twee-facet-design (personen × vragen × beoordelaars), krijgen we twintig observaties per persoon. Zouden we echter vijf vragen door de eerste beoordelaar en vijf andere vragen door de tweede beoordelaar laten beoordelen, dan krijgen we tien observaties per persoon.

Voor het bepalen van de rekenvaardigheid van personen, kunnen we antwoorden van personen op meerkeuzevragen observeren. De verzameling van alle denkbare observaties die naar onze mening acceptabel of geschikt zijn voor het geven van een oordeel over personen, wordt in de generaliseerbaarheidstheorie het universum genoemd. Uiteraard zouden we het bepalen van de rekenvaardigheid van personen willen baseren op de observaties of scores verkregen op alle vragen uit het universum, de universumscores. Om praktische redenen kunnen we de personen echter niet meer dan een steekproef van bijvoorbeeld dertig vragen uit het universum voorleggen. Het bepalen van de rekenvaardigheid baseren we op de scores die op de dertig vragen behaald worden, de geobserveerde scores. De nauwkeurigheid waarmee we menen te kunnen generaliseren van geobserveerde scores naar universumscores, dat wil zeggen de geobserveerde scores kunnen opvatten als universumscores, wordt 'generaliseerbaarheid' genoemd. Als maat voor de generaliseerbaarheid wordt de generaliseerbaarheidscoëfficiënt gebruikt. Deze coëfficiënt heeft een benedengrens van 0 en een bovengrens van 1.

In het geval van de meerkeuzevragen bestaat het universum alleen uit het facet vragen. Bestaat het universum niet uit meerkeuzevragen maar uit open vragen waarvan de antwoorden door beoordelaars beoordeeld moeten worden, dan kunnen we de beoordeling door alle in aanmerking komende beoordelaars laten verrichten. In dit geval bestaat het universum uit twee facetten: het facet 'open vragen' en het facet 'beoordelaars.' De universumscores zijn gelijk aan de scores die verkregen zouden zijn na het beoordelen van alle antwoorden op alle open vragen door alle beoordelaars. Aangezien we in de praktijk de beoordeling zullen moeten beperken tot een klein aantal beoordelaars, zijn de geobserveerde scores van de personen de scores verkregen na het beoordelen van de open vragen door dit kleine aantal beoordelaars.

De voorbeelden laten zien dat voor het generaliseren naar een universum een duidelijke beschrijving van het universum een voorwaarde is. Deze beschrijving bevat

in de eerste plaats de facetten waaruit het universum bestaat. In het eerste voorbeeld bestaat het universum alleen uit het facet 'vragen'. In het tweede voorbeeld bestaat het universum uit de facetten 'vragen' en 'beoordelaars'. In de tweede plaats moet een beschrijving van het universum uitsluitsel geven over de condities die binnen het universum vallen. Dit heeft te maken met het belangrijke onderscheid dat in de variantie-analyse aangeduid wordt met de termen 'random' en 'fixed'. In het eerste voorbeeld zijn de vragen uit de toets opgevat als een aselecte of random steekproef uit een zeer grote verzameling of 'oneindig universum' van vragen. In het tweede voorbeeld zijn vragen en beoordelaars opgevat als een random steekproef uit een oneindig universum van vragen en beoordelaars. In het voorbeeld van de meerkeuzevragen impliceert een random facet dat we vinden dat ook dertig andere vragen in aanmerking hadden kunnen komen om de rekenvaardigheid van personen te bepalen. Deze twee (of meer) toetsen van dertig vragen worden in de generaliseerbaarheidstheorie random parallelle toetsen genoemd. Voor het voorbeeld van de open vragen betekent een random facet 'open vragen' en een random facet 'beoordelaars' dat we vinden dat ook tien andere open vragen en twee andere beoordelaars in aanmerking hadden kunnen komen om de vaardigheid te bepalen. Zouden we in het tweede voorbeeld vinden dat slechts twee bepaalde beoordelaars in aanmerking komen, dan spreken we van een fixed facet 'beoordelaars'. Bij een fixed facet hebben we alle condities van een facet in ons design opgenomen en hoeven dan ook niet te generaliseren naar het universum. Later zullen we zien dat het onderscheid tussen random en fixed facetten consequenties voor de generaliseerbaarheid heeft.

3.12 Design met een facet

In een gekruist een-facet-design wordt de geobserveerde score van een persoon op een item, X_{pv}, uitgedrukt als een decompositie in vier componenten:

$$\begin{aligned} X_{pv} &= \mu && = \text{algemeen gemiddelde} \\ &+ \mu_p - \mu && = \text{persoonseffect} \\ &+ \mu_v - \mu && = \text{itemeffect} \\ &+ X_{pv} - \mu_p - \mu_v + \mu && = \text{residu} \end{aligned} \tag{3.23}$$

In (3.23) is de eerste component, het algemene gemiddelde, gedefinieerd als $\mu \equiv \mathcal{E}_p \mathcal{E}_v X_{pv}$, de gemiddelde score (= verwachting over personen en items) verkregen na het beantwoorden van alle items uit het universum door alle personen uit de

populatie. Het algemene gemiddelde geeft dezelfde constante bijdrage aan de geobserveerde score van alle personen.

De universumscore van een persoon is hier gedefinieerd als $\mu_p \equiv \mathscr{E}_v X_{pv}$, de gemiddelde score (= verwachting over items) van een persoon verkregen na het beantwoorden van alle items uit het universum van items. De tweede component, het persoonseffect $\mu_p - \mu$, is gelijk aan het verschil tussen de universumscore van een persoon en het algemene gemiddelde. Personen met een positief persoonseffect hebben een score die hoger is dan het algemene gemiddelde terwijl personen met een negatief persoonseffect een score hebben die lager is dan het algemene gemiddelde. Verschillen in vaardigheid tussen personen kunnen we weergeven als verschillen tussen hun persoonseffecten.

De moeilijkheidsgraad van een item is gedefinieerd als $\mu_v \equiv \mathscr{E}_p X_{pv}$, de gemiddelde score (= verwachting over personen) van een item na het beantwoorden van het item door alle personen uit de populatie. De derde component, het itemeffect $\mu_v - \mu$, is gelijk aan het verschil tussen de moeilijkheidsgraad van een item en het algemene gemiddelde. Een item met een positief itemeffect is gemakkelijker dan een item met een negatief itemeffect. Verschillen in moeilijkheidsgraad tussen items kunnen we weergeven als verschillen tussen hun itemeffecten.

De vierde component, de foutencomponent of het residu, is het verschil tussen X_{pv} en de eerste drie componenten. Zoals we in het voorbeeld van tabel 3.15 zullen zien, beschikken we bij het gekruiste een-facet-design maar over een enkele observatie voor elke combinatie van persoon en vraag. Dit betekent dat we het persoons- × itemeffect niet kunnen onderscheiden van andere foutenbronnen. Behalve het persoons- × itemeffect bevat het residu alle foutencomponenten die de geobserveerde score doen afwijken van de som van de eerste drie componenten.

Met uitzondering van het algemene gemiddelde, hebben de componenten in (3.23) een verdeling. Uit de wijze waarop de effecten in (3.23) gedefinieerd zijn, volgt dat hun gemiddelden gelijk zijn aan nul. De definitie van het gemiddelde van het persoonseffect bijvoorbeeld luidt $\mathscr{E}_p(\mu_p - \mu) = \mathscr{E}_p(\mu_p) - \mathscr{E}_p(\mu) = \mu - \mu = 0$. De drie componenten hebben ook elk een eigen variantie die we aanduiden met variantiecomponent. De variantiecomponenten voor respectievelijk personen, items en het residu zijn gedefinieerd als:

$$\sigma_p^2 = \mathscr{E}_p(\mu_p - \mu)^2, \tag{3.24}$$

$$\sigma_v^2 = \mathscr{E}_v(\mu_v - \mu)^2, \text{ en} \tag{3.25}$$

$$\sigma^2_{pv,e} = \mathcal{E}_p \mathcal{E}_v (X_{pv} - \mu_p - \mu_v + \mu)^2. \quad (3.26)$$

De notatie van de variantiecomponent voor het residu laat zien dat de component uit een variantiecomponent personen × vragen en een variantiecomponent voor de fouten (error) bestaat.

De variantie van de geobserveerde scores is gedefinieerd als

$$\sigma^2_X = \sigma^2_{(X_{pv})} = \mathcal{E}_p \mathcal{E}_v (X_{pv} - \mu)^2,$$

en deze totale variantie is gelijk aan de som van de drie variantiecomponenten, ofwel

$$\sigma^2_X = \sigma^2_p + \sigma^2_v + \sigma^2_{pv,e}. \quad (3.27)$$

3.12.1 Generaliseerbaarheidsstudie

Om schattingen van de variantiecomponenten van effecten te verkrijgen, dienen we een onderzoek, of wat wel genoemd wordt een generaliseerbaarheidsstudie of G-studie, uit te voeren. Het schatten gebeurt met behulp van procedures uit de variantie-analyse. We bespreken hieronder een gekruist design waarbij n_p personen en n_v items of vragen aselecte steekproeven zijn uit respectievelijk een populatie van personen en een universum van items. Tabel 3.14 bevat de variantie-analysetabel van dit gekruist random-effecten-design.

Tabel 3.14
Variantie-analysetabel van een gekruist design met twee random effecten

Effecten	Kwadraten-sommen	Vrijheids-graden	Gemiddelde kwadratensommen	Verwachte gemiddelde kwadratensommen
Personen *(p)*	SS_p	$df_p = n_p - 1$	$MS_p = SS_p/df_p$	$\mathcal{E}(MS_p) = \sigma^2_{pv,e} + n_v\sigma^2_p$
Items *(v)*	SS_v	$df_v = n_v - 1$	$MS_v = SS_v/df_v$	$\mathcal{E}(MS_v) = \sigma^2_{pv,e} + n_p\sigma^2_v$
Residu *(pv,e)*	$SS_{pv,e}$	$df_{pv,e} = (n_p - 1) \times (n_v - 1)$	$MS_{pv,e} = SS_{pv,e}/df_{pv,e}$	$\mathcal{E}(MS_{pv,e}) = \sigma^2_{pv,e}$

Schattingen van de variantiecomponenten krijgen we door het oplossen van vergelijkingen voor de verwachte gemiddelde kwadratensommen (expected mean squares). Daartoe worden de verwachte gemiddelde kwadratensommen gelijkgesteld aan de geobserveerde gemiddelde kwadratensommen (mean squares) en de exacte waarden van de variantiecomponenten vervangen door de geschatte waarden. Dit resulteert in de volgende vergelijkingen:

$$MS_{pv,e} = \hat{\sigma}^2_{pv,e},$$

$$MS_v = \hat{\sigma}^2_{pv,e} + n_p\hat{\sigma}^2_v, \text{ ofwel } \hat{\sigma}^2_v = (MS_v - MS_{pv,e})/n_p,$$

$$MS_p = \hat{\sigma}^2_{pv,e} + n_v\hat{\sigma}^2_p, \text{ ofwel } \hat{\sigma}^2_p = (MS_p - MS_{pv,e})/n_v.$$

Omdat de gemiddelde kwadratensom voor het residu gelijk is aan de schatting van de variantiecomponent voor het residu, $\hat{\sigma}^2_{pv,e} = MS_{pv,e}$, kunnen we de vergelijking voor de gemiddelde kwadratensom voor de items schrijven als $\hat{\sigma}^2_v = (MS_v - \hat{\sigma}^2_{pv,e})/n_p$. Door in deze vergelijking de gemiddelde kwadratensom van de items, berekend door het uitvoeren van een variantie-analyse, en de geschatte waarde voor de variantiecomponent van het residu in te vullen, verkrijgen we een schatting van de variantiecomponent voor items. Door herschrijven van de vergelijking voor de gemiddelde kwadratensom van de personen als $\hat{\sigma}^2_p = (MS_p - \hat{\sigma}^2_{pv,e})/n_v$, verkrijgen we op analoge wijze een schatting van de variantiecomponent voor personen.

In tabel 3.14 zien we, dat we om de drie variantiecomponenten te kunnen schatten, over de kwadratensommen (sums of squares) dienen te beschikken. Daartoe vervangen we de drie parameters μ, μ_p en μ_v in (3.14) door hun geobserveerde equivalenten, wat resulteert in de volgende decompositie:

$$X_{pv} = \bar{X} + (\bar{X}_p - \bar{X}) + (\bar{X}_v - \bar{X}) + (X_{pv} - \bar{X}_p - \bar{X}_v + \bar{X}). \qquad (3.28)$$

We illustreren de berekening van de kwadratensommen aan de hand van het voorbeeld in tabel 3.15. Deze tabel bevat de itemscores die vier personen op drie items behaald hebben. Daarnaast bevat de tabel de volgende statistische grootheden: de toetsgemiddelden, $\bar{X}_p$, van de vier personen, de itemgemiddelden, $\bar{X}_v$, van de drie items en het algemene gemiddelde, $\bar{X}$. Merk op dat het voorbeeld gelijk aan is aan het voorbeeld dat in paragraaf 3.7 bij de behandeling van de klassieke testtheorie besproken is. Voor de observaties en grootheden in deze tabel hebben we vergelijking (3.24) uitgeschreven in tabel 3.16.

De kwadratensom voor personen berekenen we door de getallen uit de kolom $(\bar{X}_p - \bar{X})$ van tabel 3.16 te kwadrateren en dan te sommeren.

Tabel 3.15

De itemscores van vier personen op drie items, de gemiddelde score per persoon en per item en het algemene gemiddelde

	Item			
Persoon	1	2	3	$\bar{X}_p$
1	1	1	1	1.00
2	1	1	0	.67
3	1	0	0	.33
4	0	0	0	.00
$\bar{X}_v$	.75	.50	.25	0.50 = $\bar{X}$

Op analoge wijze verkrijgen we de kwadratensom voor de items uit de kolom $(\bar{X}_v - \bar{X})$, en die voor het residu uit de kolom $(X_{pv} - \bar{X}_p - \bar{X}_v + \bar{X})$.

Tabel 3.16

Vergelijking (3.28) uitgeschreven voor de observaties en grootheden uit tabel 3.15

$X_{pv} =$	$\bar{X}$	$+(\bar{X}_p - \bar{X})$	$+ (\bar{X}_v - \bar{X})$	$+ (X_{pv} - \bar{X}_p - \bar{X}_v + \bar{X})$
$X_{11} = 1 =$	.500	+ .500	+ .250	— .250
$X_{12} = 1 =$	.500	+ .500	+ .000	+ .000
$X_{13} = 1 =$	.500	+ .500	— .250	+ .250
$X_{21} = 1 =$	.500	+ .167	+ .250	+ .083
$X_{22} = 1 =$	.500	+ .167	+ .000	+ .333
$X_{23} = 0 =$	.500	+ .167	— .250	— .417
$X_{31} = 1 =$	.500	— .167	+ .250	+ .417
$X_{32} = 0 =$	.500	— .167	+ .000	— .333
$X_{33} = 0 =$	.500	— .167	— .250	— .083
$X_{41} = 0 =$	.500	— .500	+ .250	— .250
$X_{42} = 0 =$	.500	— .500	+ .000	+ .000
$X_{43} = 0 =$	.500	— .500	— .250	+ .250

Voor de berekening van de totale kwadratensom brengen we in vergelijking (3.28) het algemene gemiddelde naar het linkerlid waardoor we in tabel 3.16 een nieuwe kolom, $(X_{pv} - \bar{X})$, krijgen. De getallen in deze kolom worden gekwadrateerd en daarna gesommeerd. De totale kwadratensom, SS_{tot}, is gelijk aan de som van de drie andere kwadratensommen en wordt geschreven als:

$$\Sigma_p \Sigma_v (X_{pv} - \bar{X})^2 = n_v \Sigma_p (\bar{X}_p - \bar{X})^2 + n_p \Sigma_v (\bar{X}_v - \bar{X})^2 + \Sigma_p \Sigma_v (X_{pv} - \bar{X}_p - \bar{X}_v + \bar{X})^2,$$

of:

$$\Sigma_p \Sigma_v (X_{pv} - \bar{X})^2 = SS_p + SS_v + SS_{pv,e}.$$

Tabel 3.17 bevat de resultaten van de generaliseerbaarheidsstudie voor de data uit tabel 3.15.
We laten het aan de lezer over de resultaten in tabel 3.17 na te rekenen. In de laatste kolom van de tabel staan de schattingen van de variantiecomponenten voor de drie effecten. Aangezien de grootte van de componenten afhangt van de scoreschaal die gebruikt wordt, geeft de absolute grootte van de variantiecomponenten ons geen bruikbare informatie.

Tabel 3.17
Resultaten generaliseerbaarheidsstudie voor data uit tabel 3.15

Effecten	Kwadraten-sommen	Vrijheids-graden	Gemiddelde kwadratensommen	Schattingen van variantiecomponenten
Personen *(p)*	1.667	3	0.555	$\hat{\sigma}^2_p$ = 0.139 (45.5%)
Items *(v)*	0.500	2	0.250	$\hat{\sigma}^2_v$ = 0.028 (9%)
Residu *(pv,e)*	0.833	6	0.139	$\hat{\sigma}^2_{pv,e}$ = 0.139 (45.5%)

Vandaar dat we voor elke component de procentuele bijdrage aan de totale variantie vermelden. In verband met de interpretatie van de variantiecomponenten willen we er met verwijzing naar de definities (3.24)-(3.27) nog eens benadrukken dat de variantie-componenten het resultaat zijn van de decompositie van de geschatte totale variantie van scores van afzonderlijke personen op afzonderlijke items. Dit betekent dus dat $\hat{\sigma}^2_v$ en $\hat{\sigma}^2_{pv,e}$ geen variantiecomponenten van gemiddelde of totaalscores zijn. Merk op dat we de items dichotoom gescoord hebben, zodat de variantiecomponenten in de tabel nooit groter kunnen zijn dan 0.25. De variantiecomponent voor de personen, de

geschatte universumscore-variantie, bedraagt bijna de helft van de totale variantie. De geschatte variantiecomponent voor de items is relatief klein. De geschatte variantiecomponent voor het residu is ook relatief groot. Deze variantiecomponent bestaat uit de interactiecomponent personen × vragen en andere foutenvariantie. Wanneer het residu louter uit de interactiecomponent zou bestaan, zou dit betekenen dat de rangorde van de personen niet voor alle items gelijk is. Dit zou in het voorbeeld het geval geweest zijn wanneer de eerste persoon het derde item fout en de vierde persoon het derde item goed beantwoord zou hebben.

3.12.2 Decisiestudie

Tot nu toe had de bespreking uitsluitend betrekking op de decompositie van een score van een persoon op een item uit het universum van items. Een persoon krijgt echter altijd een toets voorgelegd die uit een aantal items bestaat. Decisies of beslissingen over een persoon zijn dan ook altijd gebaseerd op de gemiddelde score of de totaalscore die behaald is op dat aantal items. In ons voorbeeld bestaat de toets uit drie random getrokken rekenitems uit het universum van rekenitems. Een andere toets met ook drie random getrokken items uit hetzelfde universum zouden we ook geschikt gevonden hebben voor het meten van de rekenvaardigheid. Dit betekent dat het universum waar in dit geval naar gegeneraliseerd wordt, het universum van random parallelle toetsen met drie items is.

Het lineaire model voor de decompositie van de gemiddelde score van een persoon op een toets met n_v items, aangeduid met X_{pV}, luidt:

$$X_{pV} = \mu + (\mu_p - \mu) + (\mu_V - \mu) + (X_{pV} - \mu_p - \mu_V + \mu). \qquad (3.29)$$

Vergelijking (3.29) is gelijk aan vergelijking (3.23) met dit verschil dat we in (3.29) de score, behaald op een enkel item, vervangen hebben door de gemiddelde score behaald op n_v items. In de notatie van (3.29) wordt een hoofdletter *V* gebruikt om aan te geven dat het de gemiddelde score van n_v items betreft. In (3.29) wordt de universumscore gedefinieerd als $\mu_p = \mathcal{E}_V X_{pV}$, de verwachte waarde van X_{pV} over random parallelle toetsen. De definities van de variantiecomponenten zijn gelijk aan die van (3.24), (3.25) en (3.26) met dien verstande dat *v* vervangen is door *V*. Het spreekt vanzelf dat door bij (3.24) de verwachting over *V* te nemen, de universumscorevariantie σ_p^2 niet verandert. De twee andere variantiecomponenten zijn: $\sigma_V^2 = \sigma_v^2 / n_v$ en $\sigma_{pV,e}^2 = \sigma_{pV,e}^2 / n_v$. Deze twee variantiecomponenten hebben betrekking op de populatie van personen en

het universum van random parallelle toetsen. De variantiecomponent $\sigma^2_V = \sigma^2_v / n_v$ moet geïnterpreteerd worden als de variantie van de verdeling van gemiddelde scores van random parallelle toetsen. De totale variantie, $\sigma^2_X = \sigma^2_{(XpV)}$ is gelijk aan $\sigma^2_X = \sigma^2_p + \sigma^2_V + \sigma^2_{pV,e}$. Wat het voorgaande betekent voor ons voorbeeld, hebben we samengevat in tabel 3.18.

In tabel 3.18 zien we hoe groot de variantiecomponenten die we in de generaliseerbaarheids-studie (G-studie) geschat hebben, in een zogenaamde decisiestudie (D-studie) worden wanneer de toets uit n_v items bestaat. Voor een gekruist een-facet-random-effect design zijn twee decisies of beslissingen van belang: de beslissing of we de toets voor het nemen van relatieve of absolute beslissingen zullen gebruiken en de beslissing uit hoeveel items we onze toets moeten laten bestaan.

Tabel 3.18

Resultaten decisiestudie voor data uit tabel 3.15

Effecten	Variantiecomponenten G-studie	Variantiecomponenten D-studie
Personen *(p)*	$\hat{\sigma}^2_p = 0.139$	$\hat{\sigma}^2_p = 0.139$
Items *(v)*	$\hat{\sigma}^2_v = 0.028$	$\hat{\sigma}^2_V = 0.028/3 = .009$
Residu *(pv,e)*	$\hat{\sigma}^2_{pv,e} = 0.139$	$\hat{\sigma}^2_{pV,e} = 0.139/3 = .046$

Het doel van een toets kan zijn, vast te stellen hoe de prestatie van een persoon zich verhoudt tot de prestaties van andere personen. Wanneer beslissingen over personen gebaseerd zijn op wat personen presteren in relatie tot andere personen, spreken we van relatieve beslissingen. De mate waarin we er met de toets in slagen personen van elkaar te onderscheiden, drukken we uit in een generaliseerbaarheidscoëfficiënt voor relatieve beslissingen. Voor het gekruiste één-facet-random-effect-design is de schatting van deze generaliseerbaarheidscoëfficiënt, een ratio van variantiecomponenten, gedefinieerd als:

$$\hat{\rho}^2 = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + \frac{\hat{\sigma}^2_{pv,e}}{n_v}}. \qquad (3.30)$$

De noemer van (3.30) bevat de universumscorevariantie $\hat{\sigma}^2_p$ en de foutenvariantie $\hat{\sigma}^2_{pv,e}/n_v$. Merk op dat de variantiecomponent $\hat{\sigma}^2_v/n_v$ niet als foutenvariantie in de noemer van (3.30) voorkomt. De reden hiervoor is dat verschillen in gemiddelde scores van random parallelle toetsen geen rol spelen wanneer we personen met elkaar willen

vergelijken. Wanneer we willen beslissen of Jan beter kan rekenen dan Piet, dan maakt het niet uit of we ze een toets met makkelijke of een toets met moeilijke items voorleggen. Brennan (1992, p. 16) laat formeel zien dat verschillen tussen scores van personen de voor beiden gelijke itemcomponent doet wegvallen.

We kunnen aan (3.30) zien dat we de coëfficiënt kunnen verhogen door de toets uit meer items laten bestaan waardoor de foutenvariantie kleiner zal worden. Omdat (3.30) een schatting van de generaliseerbaarheidscoëfficiënt na toetsverlenging geeft, wordt de formule ook wel de 'stepped-up generalizability coëfficiënt' genoemd. In hoofdstuk 11 laten we zien hoe (3.30) herschreven en gebruikt kan worden als de Spearman-Brown-formule voor toetsverlenging uit de klassieke testtheorie.

In tabel 3.18 zien we dat voor de toets met drie items de universumscorevariantie gelijk is aan .139, en de foutenvariantie aan .139/3 = .046. De generaliseerbaarheidscoëfficiënt is gelijk aan .139/{.139 + .046} = 0.75. De generaliseerbaarheidscoëfficiënt kan op twee manieren geïnterpreteerd worden. De eerste interpretatie is dat de coëfficiënt bij benadering gelijk is aan de verwachte waarde van de gekwadrateerde correlatie tussen geobserveerde en universumscores. Daarnaast kan de coëfficiënt geïnterpreteerd worden als de correlatie tussen de scores van twee random parallelle toetsen, elk bestaande uit n_v items.

Met behulp van de gemiddelde kwadratensommen kunnen we (3.30) ook uitdrukken als:

$$\hat{\rho}^2 = \frac{MS_p - MS_{pv,e}}{MS_p}. \qquad (3.31)$$

Bewezen kan worden dat in het geval van dichotome scores (3.31) gelijk is aan de KR-20 en in het geval van polytome scores aan Cronbachs coëfficiënt alpha (Sirotnik, 1970).

Het doel van de toets kan ook zijn, vast te stellen of personen in staat zijn een bepaalde prestatie te leveren, bijvoorbeeld tachtig procent van de items uit het universum goed te beantwoorden. In deze situatie zijn we niet geïnteresseerd in wat een persoon presteert in vergelijking met andere personen, maar in het absolute prestatieniveau van de persoon. Beslissingen die gebaseerd zijn op het absolute prestatieniveau van een persoon worden absolute beslissingen genoemd. In dit geval spelen verschillen in toetsen wel degelijk een rol bij de beslissing of personen aan het gewenste prestatieniveau voldoen. Wanneer een toets namelijk uit makkelijke items bestaat, kan eerder aan het prestatieniveau voldaan worden dan wanneer de toets uit moeilijke items bestaat. Dit betekent dat wanneer met een toets absolute beslissingen over personen genomen worden, $\hat{\sigma}_v^2/n_v$ bijdraagt aan de foutenvariantie.

De schatting van de generaliseerbaarheidscoëfficiënt voor absolute beslissingen is gedefinieerd als:

$$\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_v^2}{n_v} + \frac{\hat{\sigma}_{pv,e}^2}{n_v}}. \qquad (3.32)$$

Door de variantiecomponenten uit tabel 3.18 in (3.32) in te vullen, verkrijgen we de generaliseerbaarheidscoëfficiënt voor de toets uit ons voorbeeld. De coëfficiënt is gelijk aan .139/{.139 + .028/3 + .139/3} = 0.72. Merk op dat de coëfficiënten voor relatieve en absolute beslissingen slechts weinig verschillen. Dit verschil wordt uiteraard nog kleiner als we de toets verlengen.

Het onderscheid tussen relatieve en absolute beslissingen wijst op een belangrijk verschil tussen de generaliseerbaarheidstheorie en de klassieke testtheorie. De assumptie van parallelle toetsen in de klassieke testtheorie impliceert namelijk dat de gemiddelde toetsscores gelijk zijn wat betekent dat $\hat{\sigma}_v^2/n_v$ per definitie gelijk is aan nul. Dit sluit aan op de praktijk dat met de klassieke testtheorie doorgaans alleen relatieve beslissingen maar geen absolute beslissingen over personen genomen worden.

3.13 Design met twee facetten

Hiervoor hebben we de verschillende fasen van de analyse van een-facet-design besproken. Aangezien de analyse van een twee-facet-design op vergelijkbare wijze verloopt, kan de bespreking van de diverse fasen relatief kort zijn. Een voorbeeld van een gekruist twee-facet- design is een design waarbij de antwoorden op vragen van personen beoordeeld worden door beoordelaars. In een gekruist twee-facet-design wordt de geobserveerde score van een persoon *p* op een item *v*, toegekend door een beoordelaar *b*, X_{pvb}, uitgedrukt als een decompositie van de score in zeven componenten:

$$\begin{aligned}
X_{pvb} &= \mu && \text{(algemene gemiddelde)}\\
&+ \mu_p - \mu && \text{(persoonseffect)}\\
&+ \mu_v - \mu && \text{(itemeffect)}\\
&+ \mu_b - \mu && \text{(beoordelaarseffect)}\\
&+ \mu_{pv} - \mu_p - \mu_v + \mu && \text{(persoons- × itemeffect)}\\
&+ \mu_{pb} - \mu_p - \mu_b + \mu && \text{(persoons- × beoordelaarseffect)}\\
&+ \mu_{vb} - \mu_v - \mu_b + \mu && \text{(item- × beoordelaarseffect)}\\
&+ X_{pvb} - \mu_{pv} - \mu_{pb} - \mu_{vb} + \mu_p + \mu_v + \mu_b - \mu. && \text{(residu)}
\end{aligned} \qquad (3.33)$$

In (3.33) is het algemene gemiddelde gedefinieerd als $\mu = \mathcal{E}_p \mathcal{E}_v \mathcal{E}_b X_{pvb}$, de gemiddelde score (= verwachting over personen, vragen en beoordelaars) na beoordeling van alle antwoorden van alle personen uit de populatie op alle vragen uit het universum door alle beoordelaars uit het universum van beoordelaars. De universumscore van een persoon is gedefinieerd als $\mu_p = \mathcal{E}_v \mathcal{E}_b X_{pvb}$, de gemiddelde score (= verwachting over items en beoordelaars) van een persoon na beoordeling van de antwoorden op alle vragen uit het universum door alle beoordelaars uit het universum. De strengheid van een beoordelaar is gedefinieerd als $\mu_b = \mathcal{E}_p \mathcal{E}_v X_{pvb}$, de gemiddelde score (= verwachting over personen en items) van een beoordelaar na beoordeling van de antwoorden op alle vragen uit het universum door alle personen uit de populatie. De parameter μ_{pv} is gedefinieerd als $\mu_{pv} = \mathcal{E}_b X_{pvb}$, de gemiddelde score (= verwachting over beoordelaars) van een persoon op een vraag na beoordeling van het antwoord door alle beoordelaars uit het universum. De definities van de parameters μ_v, μ_{pb} en μ_{vb} zijn respectievelijk $\mu_v = \mathcal{E}_p \mathcal{E}_b X_{pvb}$, $\mu_{pb} = \mathcal{E}_v X_{pvb}$ en $\mu_{vb} = \mathcal{E}_p X_{pvb}$. De definities van de variantiecomponenten voor personen, vragen en beoordelaars zijn respectievelijk $\sigma_p^2 = \mathcal{E}_p(\mu_p - \mu)^2$, $\sigma_b^2 = \mathcal{E}_b(\mu_b - \mu)^2$ en $\sigma_v^2 = \mathcal{E}_v(\mu_v - \mu)^2$. Voor wat betreft de overige variantiecomponenten volstaan we met het geven van de definitie voor het persoons-× itemeffect: $\sigma_{pv}^2 = \mathcal{E}_p \mathcal{E}_v(\mu_{pv} - \mu_p - \mu_v + \mu)^2$.

De totale variantie is gelijk aan:

$$\sigma_X^2 = \sigma_p^2 + \sigma_v^2 + \sigma_b^2 + \sigma_{pv}^2 + \sigma_{pb}^2 + \sigma_{vb}^2 + \sigma_{pvb,e}^2. \qquad (3.34)$$

In het twee-facet-design met slechts een observatie voor elke combinatie van persoon, vraag en beoordelaar, bestaat de variantiecomponent voor het residu, $\sigma_{pvb,e}^2$, uit de niet te scheiden variantiecomponenten voor de interactie personen × vragen × beoordelaars en voor de fouten. Daarnaast worden er in (3.34) nog vijf andere variantiecomponenten voor mogelijke foutenbronnen onderscheiden: de twee variantiecomponenten voor de twee hoofdeffecten en de drie variantiecomponenten voor de drie eerste-orde-interactie-effecten.

De mogelijkheid om door toepassing van designs met meer facetten verschillende foutenbronnen te onderscheiden, is het belangrijkste verschil tussen de generaliseerbaarheids-theorie en de klassieke testtheorie. In voorgaande paragrafen zagen we dat in de klassieke testtheorie geen onderscheid gemaakt wordt tussen de verschillende storende factoren die de toetsscore van een persoon beïnvloeden en dat alle foutenbronnen door een enkele variantie-component gerepresenteerd worden.

3.13.1 Generaliseerbaarheidsstudie

De tabellen 3.19 en 3.20 bevatten alle informatie die nodig is om een generaliseerbaarheidsstudie uit te voeren. Tabel 3.19 geeft de variantie-analysetabel van een gekruist twee-facet-design met drie random effecten. In tabel 3.20 staat hoe men de kwadratensommen kan berekenen en hoe de zeven variantiecomponenten geschat kunnen worden.

Aan de hand van het voorbeeld, ontleend aan Thorndike (1982, p. 161), in tabel 3.21 laten we zien hoe de berekening van de kwadratensommen verloopt. Daartoe dienen we de zeven parameters in (3.33) te vervangen door hun geobserveerde equivalenten. Dit resulteert in de volgende decompositie:

$$X_{pvb} = \bar{X} + (\bar{X}_p - \bar{X}) + (\bar{X}_v - \bar{X}) + (\bar{X}_b - \bar{X}) + \bar{X}_{pv\sim} + \bar{X}_{pb\sim} + \bar{X}_{vb\sim} + X_{pvb\sim} \qquad (3.35)$$

In (3.35) staat $\bar{X}_{pv\sim}$ als afkorting voor $\bar{X}_{pv} - \bar{X}_p - \bar{X}_v + X$. De betekenis van afkortingen voor de andere interactietermen staat in tabel 3.20.

Tabel 3.19

Variantie-analysetabel van een gekruist design met drie random effecten en schattingen van variantiecomponenten

Effecten	Kwadraten-sommen	Vrijheidsgraden	Gemiddelde kwadratensommen	
Personen (*p*)	SS_p	$df_p = n_p - 1$	$MS_p = SS_p/df_p$	$\hat{\sigma}^2_p = (MS_p$
Items (*v*)	SS_v	$df_v = n_v - 1$	$MS_v = SS_v/df_v$	$\hat{\sigma}^2_v = (MS_v$
Beoordelaars (*b*)	SS_b	$df_b = n_b - 1$	$MS_b = SS_b/df_b$	$\hat{\sigma}^2_b = (MS_b$
Personen x items (*pv*)	SS_{pv}	$df_{pv} = (n_p - 1)(n_v - 1)$	$MS_{pv} = SS_{pv}/df_{pv}$	$\hat{\sigma}^2_{pv} = (MS_{pv}$
Personen x beoordelaars (*pb*)	SS_{pb}	$df_{pb} = (n_p - 1)(n_b - 1)$	$MS_{pb} = SS_{pb}/df_{pb}$	$\hat{\sigma}^2_{pb} = (MS_{pb}$
Items x beoordelaars (*vb*)	SS_{vb}	$df_{vb} = (n_v - 1)(n_b - 1)$	$MS_{vb} = SS_{vb}/df_{vb}$	$\hat{\sigma}^2_{vb} = (MS_{vb}$
Residu (*pvb,e*)	$SS_{pvb,e}$	$df_{pvb,e} = (n_p - 1)(n_v - 1)(n_b - 1)$	$MS_{pvb,e} = SS_{pvb,e}/df_{pvb,e}$	$\hat{\sigma}^2_{pvb,e} = MS_{pvb,e}$

Tabel 3.20

Definities van kwadratensommen en schattingen van variantiecomponenten

SS_p	$= n_v n_b \sum_p (\bar{X}_p - \bar{X})^2$	$= MS_{pvb,e}$
SS_v	$= n_p n_b \sum_v (\bar{X}_v - \bar{X})^2$	$(MS_{vb} - MS_{pvb,e})/n_p$
SS_b	$= n_p n_v \sum_b (\bar{X}_b - \bar{X})^2$	$(MS_{pb} - MS_{pvb,e})/n_v$

$$SS_{pv} = n_b \sum_p \sum_v (\bar{X}_{pv\sim})^2 = n_b \sum_p \sum_v (\bar{X}_{pv} - \bar{X}_p - \bar{X}_v + \bar{X})^2 \qquad \hat{\sigma}^2_{pv} = (MS_{pv} - MS_{pvb,e})/n_b$$

$$SS_{pb} = n_v \sum_p \sum_b (\bar{X}_{pb\sim})^2 = n_v \sum_p \sum_b (\bar{X}_{pb} - \bar{X}_p - \bar{X}_b + \bar{X})^2 \qquad \hat{\sigma}^2_b = (MS_b - MS_{vb} - MS_{pb} + MS_{pvb,e})/(n_p n_v)$$

$$SS_{vb} = n_p \sum_v \sum_b (\bar{X}_{vb\sim})^2 = n_p \sum_v \sum_b (\bar{X}_{vb} - \bar{X}_v - \bar{X}_b + \bar{X})^2 \qquad \hat{\sigma}^2_v = (MS_v - MS_{vb} - MS_{pv} + MS_{pvb,e})/(n_p n_b)$$

$$SS_{pvb,e} = \sum_p \sum_v \sum_b (X_{pvb\sim})^2 = \sum_p \sum_v \sum_b (X_{pvb} - \bar{X}_{pv} - \bar{X}_{pb} - \bar{X}_{vb} + \bar{X}_p + \bar{X}_v + \bar{X}_b - \bar{X})^2 \qquad \hat{\sigma}^2_p = (MS_p - MS_{pb} - MS_{pv} + MS_{pvb,e})/(n_v n_b)$$

$$SS_{tot} = \sum_p \sum_v \sum_b (X_{pvb} - \bar{X})^2$$

Tabel 3.21

De itemscores van zes personen op vier items en twee beoordelaars, per beoordelaar de gemiddelde score per item en per persoon, de gemiddelde score per beoordelaar, de gemiddelde score van elke persoon en het algemene gemiddelde

	Beoordelaar 1					Beoordelaar 2					
Pers.	Item: 1	2	3	4	Gem.	Item: 1	2	3	4	Gem.	$\bar{X}_p$
1	9	6	6	2	5.75	8	2	8	1	4.75	5.25
2	9	5	4	0	4.50	7	5	9	5	6.50	5.50
3	8	9	5	8	7.50	10	6	9	10	8.75	8.13
4	7	6	5	4	5.40	9	8	9	4	7.70	6.50
5	7	3	2	3	3.75	7	4	5	1	4.25	4.00
6	10	8	7	7	8.00	7	7	10	9	8.25	8.13
Gem.	8.33	6.17	4.83	4.00	5.83	8.00	5.33	8.33	5.00	6.67	$\bar{X}$ = 6.25

Tabel 3.21 bevat de itemscores die twee beoordelaars aan de antwoorden op vier items aan zes personen toegekend hebben. Voor persoon 1 uit deze tabel hebben we (3.35) uitgeschreven in tabel 3.22.

Tabel 3.22

Vergelijking (3.35) uitgeschreven voor persoon 1 uit tabel 3.21

X_{pvb} =	$\bar{X}$	+ $(\bar{X}_p - \bar{X})$	+ $(\bar{X}_v - \bar{X})$	+ $(\bar{X}_b - \bar{X})$	+ $\bar{X}_{pv}$~	+ $\bar{X}_{pb}$~	+ $\bar{X}_{vb}$~	+ X_{pvb}~
X_{111} = 9 =	6.25	— 1.00	+ 1.92	— 0.42	+ 1.33	+ 0.92	+ 0.58	— 0.58
X_{112} = 8 =	6.25	— 1.00	+ 1.92	+ 0.42	+ 1.33	— 0.92	— 0.58	+ 0.58
X_{121} = 6 =	6.25	— 1.00	— 0.50	— 0.42	— 0.75	+ 0.92	+ 0.83	+ 0.67
X_{122} = 2 =	6.25	— 1.00	— 0.50	+ 0.42	— 0.75	— 0.92	— 0.83	— 0.67
X_{131} = 6 =	6.25	— 1.00	+ 0.33	— 0.42	+ 1.42	+ 0.92	— 1.33	— 0.17
X_{132} = 8 =	6.25	— 1.00	+ 0.33	+ 0.42	+ 1.42	— 0.92	+ 1.33	+ 0.17
X_{141} = 2 =	6.25	— 1.00	— 1.75	— 0.42	— 2.00	+ 0.92	— 0.08	+ 0.08
X_{142} = 1 =	6.25	— 1.00	— 1.75	+ 0.42	— 2.00	— 0.92	+ 0.08	— 0.08

Voor het berekenen van de kwadratensommen moeten we vergelijking (3.35) ook nog uitschrijven voor de vijf andere personen, wat een uitbreiding betekent van tabel 3.22 met de decomposities van veertig itemscores. De zeven kwadratensommen worden verkregen door de getallen in de desbetreffende kolommen van tabel 3.22 te kwadrateren en te sommeren. Beschikken we over de kwadratensommen, dan kunnen we schattingen van de variantie-componenten eenvoudig berekenen met behulp van tabel 3.20. Wellicht ten overvloede merken we op dat de standaardfouten van variantiecomponenten bij kleine aantallen personen en condities zeer groot zijn (Brennan, 1992, p. 104). De steekproef uit de populatie moet uit minstens honderd personen bestaan teneinde acceptabele standaardfouten te verkrijgen (Smith, 1978). De resultaten van de generaliseerbaarheidsstudie voor het voorbeeld staan vermeld in tabel 3.23.

Tabel 3.23

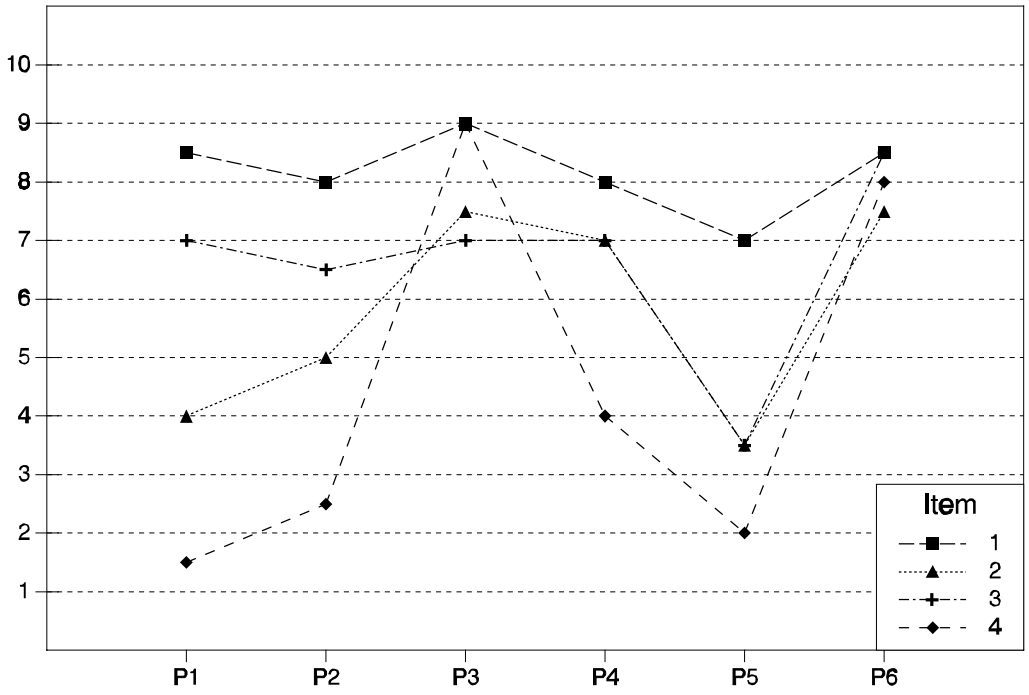
Resultaten generaliseerbaarheidsstudie voor data uit tabel 3.21

Effecten	Kwadraten-sommen	Vrijheids-graden	Gemiddelde kwadraten-sommen	Schattingen van variantie-componenten	
Personen *(p)*	109.75	5	21.95	$\hat{\sigma}^2_p$	= 2.16 (28%)
Items *(v)*	85.17	3	28.39	$\hat{\sigma}^2_v$	= 1.26 (15%)
Beoordelaars *(b)*	8.33	1	8.33	$\hat{\sigma}^2_b$	= –0.15 (0%)

Personen × items *(pv)*	59.08	15	3.94	$\hat{\sigma}^2_{pv}$ = 0.98 (12%)
Personen × beoordelaars *(pb)*	13.42	5	2.68	$\hat{\sigma}^2_{pb}$ = 0.18 (2%)
Items × beoordelaars *(vb)*	33.83	3	11.28	$\hat{\sigma}^2_{vb}$ = 1.55 (19%)
Residu *(pvb,e)*	29.42	15	1.96	$\hat{\sigma}^2_{pvb,e}$ = 1.96 (24%)

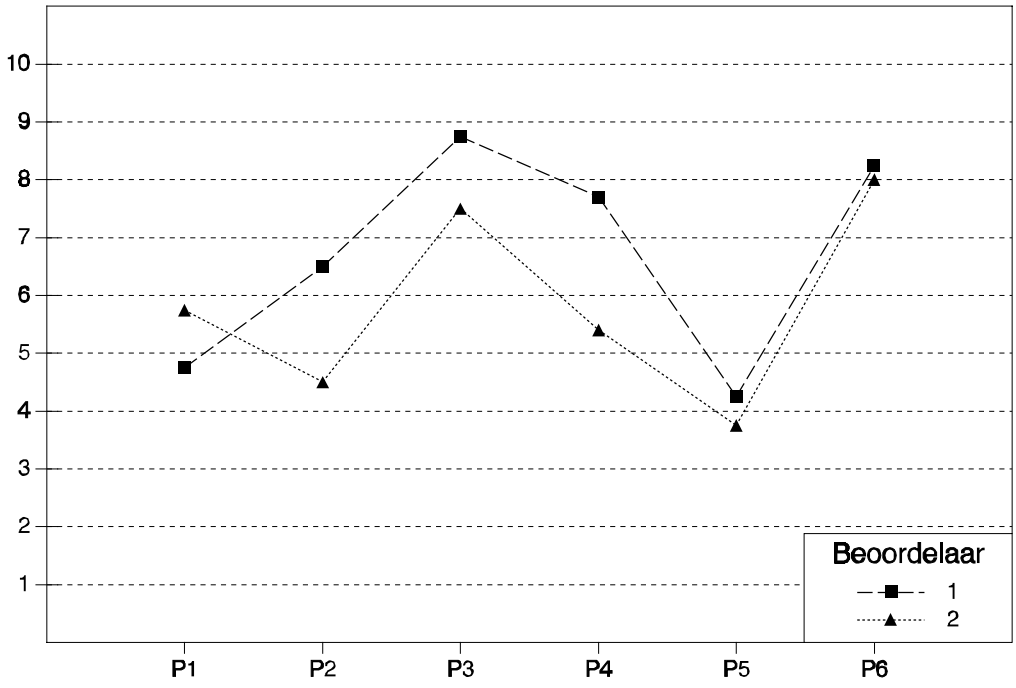
De laatste kolom van tabel 3.23 bevat de schattingen van de variantiecomponenten en hun procentuele bijdrage aan de totale variantie. We zien dat de variantiecomponent van de beoordelaars negatief is. Hoewel in theorie variantiecomponenten niet negatief kunnen zijn, kunnen schattingen van variantiecomponenten wel negatief zijn. Negatieve schattingen hebben veelal twee mogelijke oorzaken. Relatief grote negatieve componenten zijn meestal het gevolg van het gebruik van het verkeerde model. Een relatief grote negatieve component van beoordelaars had er in ons voorbeeld op kunnen wijzen dat het lineaire model in (3.33) niet het juiste model was om de data te analyseren. Relatief kleine negatieve componenten zijn meestal het gevolg van het gebruik van een te kleine steekproef. Dit laatste is waarschijnlijk de oorzaak van de negatieve component in ons voorbeeld. Aangezien negatieve componenten niet mogelijk zijn, worden negatieve schattingen vervangen door nul. Merk op dat er andere schattingsmethoden voor variantiecomponenten zijn die niet leiden tot negatieve schattingen. Een daarvan is de restrictieve grootste-aannemelijkheidschattingsmethode. De relatief grote bijdrage van de variantiecomponent voor de items is met name het gevolg van het grote verschil in moeilijkheidsgraad tussen item 1 en item 4. De gemiddelde itemscore van item 1 is 8.17, terwijl die van item 4 gelijk is aan 4.50.

De bijdrage van de interactiecomponent personen × items is veel groter dan die van de interactiecomponent personen × beoordelaars. Interactie tussen personen en items betekent dat personen niet consistent antwoorden op de verschillende items. Interactie tussen personen en beoordelaars houdt in dat personen niet consistent beoordeeld worden door verschillende beoordelaars. In figuur 3.3. hebben we de interactie personen × items grafisch gepresenteerd.



Figuur 3.3 Interactie personen × items

In figuur 3.3 is voor elk item een lijn getrokken die de gemiddelde itemscores, $\bar{X}_{pv}$, van personen, P1-P6, met elkaar verbindt. We zien dat de vier lijnen elkaar bij verschillende personen kruisen, wat betekent dat het niet dezelfde persoon is die de hoogste of laagste score op elk item behaalt. Lijnen die elkaar kruisen wijzen er op dat er sprake is van interactie. Merk op dat in tabel 3.22 de berekening van de variantiecomponent voor de interactie tussen personen en items gebaseerd is op $\bar{X}_{pv}\sim = \bar{X}_{pv} - \bar{X}_{p} - \bar{X}_{v} + \bar{X}$. We hadden de interactie tussen personen en items ook met behulp van $X_{pv}\sim$ in plaats van $\bar{X}_{pv}$ kunnen afbeelden. Wanneer de vier lijnen parallel lopen is, de kwadratensom personen × items, en dus ook de variantiecomponent, gelijk aan nul.



Figuur 3.4 Interactie personen × beoordelaars

Om mogelijke interactie tussen personen en beoordelaars te onderzoeken, is in figuur 3.4 voor elk item een lijn getrokken die de gemiddelde beoordelaarsscores, $\bar{X}_{pb}$, van personen met elkaar verbindt. We zien dat de twee lijnen elkaar bij de eerste persoon kruisen maar bij de andere vijf personen nagenoeg parallel lopen. Dit betekent dat de twee beoordelaars de eerste persoon niet, maar de vijf andere personen wel op dezelfde wijze onderscheiden. De variantiecomponent voor de interactie tussen personen en beoordelaars blijkt dan ook gering te zijn.

De interactie items × beoordelaars is de grootste eerste-orde-interactie, met name veroorzaakt door de derde vraag. Die vraag heeft van de eerste beoordelaar een lage beoordeling, gemiddelde score 4.83, en van de tweede beoordelaar een hoge beoordeling, gemiddelde score 8.33, ontvangen.

3.13.2 Decisiestudie

In ons voorbeeld bestaat de toets uit vier random getrokken items uit het universum van items en twee random getrokken beoordelaars uit het universum van beoordelaars die de antwoorden op de items beoordelen. Een andere toets met vier random getrokken items en twee random getrokken beoordelaars zou ook acceptabel geweest

zijn. Het universum waar in dit geval naar gegeneraliseerd wordt, is het universum van random parallelle toetsen met vier items en twee beoordelaars.

De schatting van de generaliseerbaarheidscoëfficiënt voor relatieve beslissingen is voor het gekruiste twee-facet-random-effect-design gedefinieerd als:

$$\hat{\rho}^2 = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + \frac{\hat{\sigma}^2_{pv}}{n_v} + \frac{\hat{\sigma}^2_{pb}}{n_b} + \frac{\hat{\sigma}^2_{pvb,e}}{n_v n_b}}. \qquad (3.36)$$

Naast de universumscorevariantie, bevat de noemer van (3.36) drie variantiecomponenten die interacties met personen betreffen. Hiervoor zagen we dat een relatief grote variantie- component voor de interactie tussen personen en items inhoudt dat bijvoorbeeld Jan niet op ieder item meer presteert dan Piet. Het maakt voor het nemen van relatieve beslissingen dan ook wel degelijk uit welke items aan welke personen voorgelegd worden. Een bepaald item wordt namelijk door Jan als gemakkelijk en door Piet als moeilijk opgevat, terwijl bij een ander item het omgekeerde het geval is. De variantiecomponent voor de interactie tussen personen en items dient dan ook beschouwd te worden als foutenvariantie. Ook de variantiecomponent voor de interactie tussen personen en beoordelaars, dat wil zeggen dat het van de beoordelaar afhangt of Jan beter is dan Piet, dient als foutenvariantie beschouwd te worden. De variantiecomponent voor het residu is per definitie foutenvariantie. Voor de toets uit ons voorbeeld is de generaliseerbaarheidscoëfficiënt gelijk aan: 2.16/{2.16 + 0.99/4 + 0.18/2 + 1.96/8} = .79.

De schatting van de generaliseerbaarheidscoëfficiënt voor absolute beslissingen is voor het gekruiste twee-facet-random-effect design gedefinieerd als:

$$\hat{\rho}^2 = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + \frac{\hat{\sigma}^2_v}{n_v} + \frac{\hat{\sigma}^2_b}{n_b} + \frac{\hat{\sigma}^2_{pv}}{n_v} + \frac{\hat{\sigma}^2_{pb}}{n_b} + \frac{\hat{\sigma}^2_{pvb,e}}{n_v n_b}}. \qquad (3.37)$$

Bij het nemen van absolute beslissingen maakt het niet alleen uit of er makkelijke of moeilijke vragen aan de personen voorgelegd worden, maar ook of die vragen door milde of strenge beoordelaars beoordeeld worden. Vandaar dat in (3.37) naast de variantiecomponenten voor de drie interacties ook de variantiecomponenten voor de items en voor de beoordelaars beschouwd worden als foutenvariantie. De generaliseerbaarheidscoëfficiënt voor absolute beslissingen is gelijk aan 2.16/{2.16 + 1.26/4 + 0.0/2 + 0.99/4 + 0.18/2 + 1.96/8} = .71 voor de toets uit ons voorbeeld.

3.14 Andere aspecten van de generaliseerbaarheidstheorie

Formule (3.36) laat zien dat we de generaliseerbaarheidscoëfficiënt kunnen verhogen door de toets te verlengen, wat neerkomt op het vergroten van het aantal items of het aantal beoordelaars. Voor het realiseren van dezelfde generaliseerbaarheidscoëfficiënt hebben we meer condities nodig van een facet met een relatief grote variantiecomponent die bijdraagt aan de foutenvariantie, dan condities van een facet met een relatief kleine variantiecomponent. We verwijzen naar hoofdstuk 11 voor een bespreking van toetsverlenging bij designs met meer facetten.

De generaliseerbaarheidscoëfficiënt kan ook verhoogd worden door een random facet op te vatten als een fixed facet. Dat een facet fixed is, wil zeggen dat een toets alle condities van een facet bevat. Beschouwen we in ons voorbeeld de items als fixed facet, dan generaliseren we niet meer naar het universum van random parallelle toetsen met vier items en twee beoordelaars, maar naar het universum van random parallelle toetsen met twee beoordelaars. Het spreekt vanzelf dat door het beperken van het universum waar naar gegeneraliseerd wordt, de beslissingen over personen nauwkeuriger kunnen zijn. Voor een bespreking van designs met fixed facets verwijzen we naar Shavelson en Webb (1991, pp. 65-82).

De bespreking in voorgaande paragrafen heeft zich beperkt tot gekruiste designs met een enkel facet en met twee facetten. Binnen de generaliseerbaarheidstheorie kunnen echter ook designs met meer dan twee facetten geanalyseerd worden. Daarnaast kunnen ook zogenaamde genestelde designs geanalyseerd worden. Ons voorbeeld met twee facetten zou een genesteld design zijn wanneer de eerste en de tweede vraag door de eerste beoordelaar beoordeeld worden en de derde en vierde vraag door de tweede beoordelaar. In dat geval zeggen we dat de vragen genesteld zijn binnen de beoordelaars. Genestelde designs komen vooral voor bij niet-experimenteel onderzoek (Feldt & Brennan, 1989). In het algemeen heeft het gebruik van gekruiste designs de voorkeur, omdat het met de resultaten van de generaliseerbaarheidsstudie van gekruiste designs mogelijk is na te gaan hoe de resultaten voor een genesteld design geweest zouden zijn. Het omgekeerde is niet het geval.

In de voorbeelden die tot nu toe besproken zijn, hadden de beslissingen steeds betrekking op personen. In veel onderzoek, met name onderzoek op het gebied van het onderwijs, zijn we echter niet of niet uitsluitend geïnteresseerd in (verschillen tussen) personen maar ook in klassen, leerdoelen of andere meetobjecten. Om aan te geven dat elk facet uit een design het meetobject kan zijn, introduceerden Cardinet, Tourneur en Allal (1981) het zogenaamde symmetrieprincipe. Uitgaande van dat principe laten zij

zien hoe binnen het kader van de generaliseerbaarheidstheorie een grote verscheidenheid aan onderzoeksvragen beantwoord kan worden.

De meest gebruikte schatting van de universumscore van een persoon is de geobserveerde gemiddelde score van een persoon. In Cronbach e.a. (1972) worden echter ook varianten van Kelley's formule (zie paragraaf 3.5) voor schattingen van universumscores besproken. Hoe schattingen van universumscores verkregen kunnen worden met behulp van lineaire predictiefuncties wordt beschreven door Jarjoura (1983).

Tenslotte dient opgemerkt te worden dat met de generaliseerbaarheidstheorie niet alleen univariate maar ook multivariate modellen, dat wil zeggen modellen waarbij de personen een aantal universumscores hebben, geanalyseerd kunnen worden. Voor een bespreking van modellen uit de multivariate generaliseerbaarheidstheorie verwijzen we naar Cronbach e.a. (1972), Shavelson en Webb (1981) en Brennan (1992).

3

Klassieke testtheorie en generaliseerbaarheidstheorie

De klassieke testtheorie beschrijft het verschijnsel meetfout en procedures om de grootte van meetfouten te bepalen. Het uitgangspunt van de klassieke testtheorie is een meting x_{vt} die verkregen is door een meetinstrument t voor te leggen aan een persoon v. Zoals is uiteen- gezet in het vorige hoofdstuk, wordt een meting altijd gecodeerd als een getal. Zo'n gecodeerde meting noemt men een score. De klassieke testtheorie houdt zich niet bezig met de aard, het schaalniveau en de interpretatie van een score. Zij houdt zich met slechts een enkel probleem bezig, en wel met de meetfout waarmee een score x_{vt} behept is. De meetfout wordt geacht op te treden doordat men bij het meten niet alle factoren in de hand heeft die op een meting van invloed zijn. Zulke factoren verstoren de meetprocedure en zorgen er voor dat men niet de meting krijgt die men graag had willen hebben maar een daar enigszins van afwijkende score. Verstorende factoren kunnen zijn gelegen in de te meten persoon, in het meetinstrument, en in de meetsituatie. Een voorbeeld van de eerste soort is de bloeddruk: deze vertoont in de loop van de dag zulke grote fluctuaties dat een enkele meting eigenlijk onvoldoende is. Een voorbeeld van de tweede soort verstorende factoren is de thermometer. Dat instrument wisselt warmte uit met het te meten voorwerp, waardoor de thermometer niet de exacte temperatuur van het voorwerp aangeeft. Een voorbeeld van een verstoring in de meetsituatie is het eindexamen dat wordt afgenomen in een schoolgebouw waarnaast een heistelling palen de grond in boort.

De belangrijkste parameters uit de klassieke testtheorie zijn correlaties en standaard-afwijkingen. Het gebruik van dergelijke parameters brengt met zich mee dat alle uitspraken van de klassieke testtheorie over personen en over meetinstrumenten gerelateerd zijn aan een bepaalde populatie. Zo kan men eigenschappen van een meetinstrument die bepaald zijn in een populatie, niet zonder meer voor geldend houden in een andere populatie. Voor een aantal meetproblemen schiet de klassieke testtheorie dan ook tekort. De wens, te kunnen beschikken over parameters van

personen en meetinstrumenten die niet aan een populatie gebonden zijn, heeft geleid tot de itemresponstheorie. Deze theorie wordt behandeld in hoofdstuk 4.

De klassieke testtheorie wordt eerst, in de paragrafen 3.1 tot en met 3.6, in abstracte termen beschreven. In de paragrafen 3.7 tot en met 3.10 worden diverse grootheden concreet geïllustreerd aan de hand van een voorbeeld. Daarbij worden ook grootheden behandeld die optreden bij het construeren van toetsen. De toets uit het voorbeeld is klein gehouden om het de lezer mogelijk te maken het rekenwerk te volgen. Een uitbreiding van de klassieke testtheorie, de generaliseerbaarheidstheorie, wordt in de paragrafen 3.11 tot en met 3.14 besproken.

3.1 Ware score

De waargenomen score is door de verstorende factoren niet altijd de meting die we zouden willen hebben. De klassieke testtheorie veronderstelt nu dat het effect van de verstorende factoren beschouwd kan worden als een aselecte trekking uit een kansverdeling. In feite is dit de enige veronderstelling die de klassieke testtheorie kent. De afleiding die nu volgt is gebaseerd op Novick (1966). Uit de zojuist genoemde veronderstelling kan men de gehele klassieke testtheorie opbouwen. Als de bij de meting x_{vt} optredende meetfout wordt aangeduid met ε_{vt}, veronderstelt de klassieke testtheorie dat deze meetfout een realisatie is van een toevalsvariabele E_{vt}. Deze toevalsvariabele draagt twee subscripten om aan te geven dat zij varieert binnen de combinatie van de vaste persoon v en het vaste meetinstrument t. Beschouw nu de voor de meetfout gecorrigeerde meting $\tau_{vt} = x_{vt} - \varepsilon_{vt}$. Men kan dan ook schrijven: $x_{vt} = \tau_{vt} + \varepsilon_{vt}$. Deze uitdrukking schrijft de score x_{vt} als een ontbinding, een decompositie, in twee termen. De eerste term, τ_{vt}, zou men kunnen opvatten als de meting die men had willen verkrijgen. Maar de gegeven ontbinding is niet uniek. Men kan namelijk bij de term τ_{vt} een willekeurige constante c optellen en deze constante van de term ε_{vt} aftrekken zonder dat het resultaat verandert: $x_{vt} = \tau_{vt} + \varepsilon_{vt} = (\tau_{vt} + c) + (\varepsilon_{vt} - c)$. In feite is dit een geval van een vergelijking met twee onbekenden. Om met de gegeven decompositie uit de voeten te kunnen, moet men normeren. Daaronder verstaat men het kiezen en vastleggen van een waarde voor de constante c. In de klassieke testtheorie heeft men voor de volgende normering gekozen. Aangezien E_{vt} een toevalsvariabele is met realisaties ε_{vt}, en τ_{vt} een vaste waarde heeft, is x_{vt} een realisatie van een toevalsvariabele X_{vt}. Voor de constante c is in de klassieke testtheorie de verwachte waarde van de toevalsvariabele E_{vt} gekozen: $c = \mathscr{E}(E_{vt})$. De verwachte waarde van een toevalsvariabele kan men in dit boek opvatten als het

gemiddelde van een hele grote steekproef van trekkingen uit de verdeling van die variabele. De verwachte waarde van een constante is gelijk aan die constante. Met de gekozen normering kan men nu de toevalsvariabele X_{vt} schrijven als: $X_{vt} = \{\tau_{vt} + \mathscr{E}(E_{vt})\} + \{E_{vt} - \mathscr{E}(E_{vt})\}$. Daaruit volgt onmiddellijk dat $\mathscr{E}(X_{vt}) = \tau_{vt} + \mathscr{E}(E_{vt})$. Ook deze decompositie moet genormeerd worden. In de klassieke testtheorie stelt men daartoe $\mathscr{E}(E_{vt})$ gelijk aan 0. Het resultaat is de volgende belangrijke uitdrukking:

$$\mathscr{E}(X_{vt}) = \tau_{vt}. \tag{3.1}$$

Het rechterlid van (3.1) heet in de klassieke testtheorie de ware score van persoon *v* op meet- instrument *t.* Men dient te beseffen dat de door (3.1) gedefinieerde ware score een wis- kundige constructie is en niet noodzakelijkerwijze gelijk is aan de score die verkregen zou zijn als er geen verstorende factoren aanwezig waren. Het kan bijvoorbeeld goed zijn dat de toevalsvariabele X_{vt} alleen maar gehele waarden kan aannemen; dat sluit echter niet uit dat de verwachte waarde van die variabele, de ware score, een gebroken getal is.

3.2 De centrale formule van de klassieke testtheorie

De ware score is, omdat hij is gedefinieerd als een verwachte waarde, een maat voor de centrale tendentie van de scores: hij geeft aan om welke waarde de verkregen metingen variëren. Het is van groot belang, te weten in welke mate de metingen rondom de ware score variëren. Bekende maten voor de variatie van een toevalsvariabele zijn de variantie en de standaardafwijking van die variabele. De variantie van een toevalsvariabele is gelijk aan de verwachte waarde van het kwadraat van het verschil tussen een score en de daarbij behorende ware score. Voor de toevalsvariabele X_{vt} schrijft men de variantie als volgt: $\sigma^2_{X_{vt}} = \mathscr{E}\{(X_{vt} - \tau_{vt})^2\}$. Omdat geldt dat $X_{vt} - \tau_{vt}$ gelijk is aan E_{vt} en omdat $\mathscr{E}(E_{vt})$ gelijk is aan 0, kan men de zojuist geschreven variantie ook schrijven als: $\sigma^2_{X_{vt}} = \mathscr{E}\{(E_{vt})^2\}$. De laatste uitdrukking kan men natuurlijk ook schrijven als: $\sigma^2_{E_{vt}}$.

Merk op dat de in deze paragraaf genoemde varianties alle betrekking hebben op de variatie van toevalsvariabelen die zijn gedefinieerd voor een vaste persoon *v* en een vast meetinstrument *t*. Om de varianties te kunnen schatten, zou men moeten beschikken over herhaalde metingen van *v* met *t*, verkregen onder identieke omstandigheden. Door de eerder genoemde verstorende factoren is het echter niet mogelijk, herhaalde metingen te verkrijgen onder identieke omstandigheden. In plaats

van herhaalde metingen te gebruiken, gaat de klassieke testtheorie er toe over meer personen tegelijk te beschouwen. Het is duidelijk dat nu kenmerken van een populatie $\mathcal{P}$ van personen een rol gaan spelen.

Beschouw een willekeurig uit de populatie $\mathcal{P}$ getrokken persoon. Om aan te geven dat de persoon willekeurig is getrokken, duiden we die persoon aan met een $\star$. Zodra we de persoon $\star$ hebben getrokken, geldt alles wat hierboven gezegd is. Men kan denken aan een tweestapsprocedure: eerst trekt men willekeurig een persoon $\star$ uit de populatie $\mathcal{P}$, en dan trekt men een meetfout $\varepsilon_{\star t}$ uit de verdeling van de toevalsvariabele $E_{\star t}$. Bij de persoon $\star$ behoort een ware score $\tau_{\star t}$. Men kan nu ook zeggen dat er drie nieuwe toevalsvariabelen zijn gemaakt: $T_{\star t}$, $E_{\star t}$ en $X_{\star t}$. De laatste twee variabelen variëren zowel over personen als binnen de aselect gekozen persoon; de eerste varieert alleen over personen. De betrekking tussen de drie toevalsvariabelen kan men schrijven als: $X_{\star t} = T_{\star t} + E_{\star t}$. Omdat we in het vervolg steeds een enkel meetinstrument en een enkele populatie beschouwen, laten we waar dat mogelijk is de subscripten weg. De laatst geschreven betrekking kan men dan schrijven als:

$$X = T + E \ . \tag{3.2}$$

Formule (3.2) is de centrale formule van de klassieke testtheorie. Men kan er, jammer genoeg, niet aan zien dat de toevalsvariabele T alleen over personen varieert maar niet binnen een persoon, en dat de toevalsvariabelen X en E zowel tussen de personen als binnen elke persoon variëren. In het bovenstaande is daarom uiteengezet hoe deze formule tot stand komt.

3.3 Betrouwbaarheid

Uit (3.2) kan men enige interessante betrekkingen afleiden. In de eerste plaats geldt dat de verwachte waarde van de toevalsvariabele E over de populatie $\mathcal{P}$ gelijk is aan 0: $\mathcal{E}_{\mathcal{P}}\mathcal{E}(E) = \mathcal{E}_{\mathcal{P}}(0) = 0$. Er zijn twee verwachtingen genomen: in de eerste plaats de verwachting over de meetfouten binnen een persoon, en in de tweede plaats de verwachting over personen van de verwachte meetfout. Dit komt overeen met het feit dat E zowel binnen een persoon als over personen varieert.

In de tweede plaats kan men afleiden dat de correlatie tussen de variabelen T en E gelijk is aan 0. Immers, voor elke persoon v in $\mathcal{P}$ geldt dat $\mathcal{E}(E_{vt}) = 0$. Dit geldt dan ook voor een willekeurig uit de populatie $\mathcal{P}$ getrokken persoon $\star$. A fortiori geldt dit voor elke persoon $\star$ uit $\mathcal{P}$ die een ware score gelijk aan $\tau_{\star t}$ heeft: $\mathcal{E}(E_{\star t}|\tau_{\star t}) = 0$. Dit geldt natuurlijk voor elke waarde van $\tau_{\star t}$. De uitdrukking $\mathcal{E}(E_{\star t}|\tau_{\star t})$ heet: de regressie

van E op T. Aangezien de regressie van E op T gelijk is aan 0, is ook de correlatie tussen E en T gelijk aan 0.

In de derde plaats kan men uit de decompositie van X die gegeven is in (3.2), de volgende decompositie afleiden van de variantie σ_X^2 van de variabele X:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2. \tag{3.3}$$

De drie varianties zijn de varianties van respectievelijk de waargenomen toetsscores, de ware toetsscores en de meetfouten. Men noemt de drie varianties doorgaans: geobserveerde variantie, ware variantie en foutenvariantie.

Een van de voornaamste grootheden in de klassieke testtheorie is de betrouwbaarheid. Deze grootheid, die wordt voorgesteld door het symbool ρ_{XT}^2, is als volgt gedefinieerd:

$$\rho_{XT}^2 = \sigma_T^2 / \sigma_X^2 = \sigma_T^2 / \{\sigma_T^2 + \sigma_E^2\}. \tag{3.4}$$

Zolang de geobserveerde variantie groter is dan 0, neemt de betrouwbaarheid waarden aan tussen 0 en 1. De betrouwbaarheid is gelijk aan 0 als er geen ware variantie is: men meet alleen maar meetfouten met het meetinstrument. De betrouwbaarheid is gelijk aan 1 als er geen sprake is van meetfouten: $\sigma_E^2 = 0$, wat overeenkomt met $\sigma_X^2 = \sigma_T^2$. Elke geobserveerde score van een persoon is dan gelijk aan de ware score van die persoon. In het uitzonderlijke geval dat σ_X^2 gelijk is aan 0, is de betrouwbaarheid niet gedefinieerd.

Waarom de betrouwbaarheid wordt aangeduid met het symbool ρ_{XT}^2, wordt duidelijk als men de correlatie beschouwt tussen de geobserveerde scores X en de ware scores T. De teller van deze correlatie is gelijk aan de covariantie tussen X en T:

$$Cov(X,T) = \mathcal{E}[\{X - \mathcal{E}(X)\} \times \{T - \mathcal{E}(T)\}] =$$

$$\mathcal{E}([\{T - \mathcal{E}(T)\} + \{E - \mathcal{E}(E)\}] \times \{T - \mathcal{E}(T)\}) =$$

$$\mathcal{E}\{T - \mathcal{E}(T)\}^2 + \mathcal{E}[\{T - \mathcal{E}(T)\} \times \{E - \mathcal{E}(E)\}] = \sigma_T^2 + Cov(T,E) =$$

$$\sigma_T^2 + \sigma_T \sigma_E \rho_{TE} = \sigma_T^2.$$

In deze afleiding is gebruik gemaakt van het eerder gegeven resultaat dat de correlatie tussen T en E, hier aangeduid met ρ_{TE}, gelijk is aan 0. De noemer van de correlatie X en T is gelijk aan $\sigma_X \sigma_T$. We zien dan dat de correlatie ρ_{XT} tussen de geobserveerde

scores X en de ware scores T gelijk is aan σ_T/σ_X; deze uitdrukking is gelijk aan de wortel uit de in (3.4) gegeven uitdrukking voor de betrouwbaarheid.

3.4 Standaardmeetfout

De wortel uit de foutenvariantie σ_E^2 heet de standaardmeetfout. Uit (3.4) kan men afleiden dat de standaardmeetfout σ_E kan worden bepaald uit de geobserveerde variantie en de betrouwbaarheid: $\sigma_E = \sigma_X(1-\rho_{XT}^2)^{½}$. De standaardmeetfout is uitgedrukt in de schaal- eenheid van het meetinstrument. Men kan twee standaardmeetfouten van verschillende meetinstrumenten dan ook niet zomaar met elkaar vergelijken. De betrouwbaarheid daaren- tegen is louter een getal; men kan de betrouwbaarheden van twee toetsen wel onderling vergelijken. De standaardmeetfout wordt voornamelijk gebruikt om uit een geobserveerde score een intervalschatting voor de ware score te bepalen.

Men heeft het wel als een bezwaar van de klassieke testtheorie gezien dat er een enkele standaardmeetfout is die wordt toegepast bij elke score x_{vt}. Het wordt onrealistisch geacht aan te nemen dat een toets op elk scoreniveau even nauwkeurig meet. Aan dit bezwaar wordt tegemoet gekomen in de itemresponstheorie die in hoofdstuk 4 wordt besproken. Ook binnen de klassieke testtheorie heeft men dit bezwaar erkend. Er zijn diverse procedures ontwikkeld om voor verschillende scoreniveaus een eigen standaardmeetfout te bepalen. Een overzicht van deze procedures vindt men bij Feldt, Steffen en Gupta (1985). Een van die procedures is ontwikkeld door Thorndike (1951).

De methode van Thorndike maakt gebruik van het begrip parallelle metingen. Dit begrip wordt besproken in paragraaf 3.6.1. Een paar eigenschappen van parallelle metingen worden hier gebruikt. Veronderstel dat het mogelijk is, het meetinstrument te verdelen in twee parallelle deeltoetsen. Voor zulke parallelle deeltoetsen, met scorevariabelen X_1 en X_2, geldt dat $\mathscr{E}(X_1) = \mathscr{E}(X_2)$ en $\sigma_{X_1}^2 = \sigma_{X_2}^2$. Bovendien geldt dat de bijbehorende meetfouten E_1 en E_2 onderling onafhankelijk, en dus ongecorreleerd zijn. De standaardafwijking van de verschilscore $X_1 - X_2$ kan men nu schrijven:

$$\sigma_{(X_1-X_2)} = \sigma_{(E_1-E_2)} = (\sigma_{E_1}^2 + \sigma_{E_2}^2)^{½} = \sigma_E. \qquad (3.5)$$

In deze afleiding is gebruik gemaakt van het feit dat de correlatie tussen de meetfouten E_1 en E_2 gelijk is aan 0, van het feit dat $\sigma_{E_1}^2 = \sigma_{E_2}^2$, en van het feit dat $\sigma_{E_1}^2 = ½\sigma_E^2$. Met (3.5) kan men de standaardmeetfout van een meetinstrument schatten. Thorndike

stelt voor, (3.5) toe te passen op deelgroepen van personen die dezelfde score hebben. Zulke groepen noemt men wel scoregroepen. Het is dan mogelijk, met behulp van (3.5) standaardmeetfouten te schatten in verschillende scoregroepen afzonderlijk. In de praktijk zal het vaak nodig zijn, scoregroepen samen te nemen om te komen tot groepen met een voldoende aantal waarnemingen voor het nauwkeurig schatten van de standaardmeetfout.

3.5 Schattingen van de ware score

Een voor de hand liggende schatter van de ware score τ is de waargenomen score x. De waargenomen score is een zuivere schatter van de ware score. Men noemt een schatter zuiver als zijn verwachte waarde gelijk is aan de te schatten parameter. De vraag rijst hoe precies de geobserveerde score als schatter van de ware score is. Onder de veronderstelling dat de meetfout binnen elke persoon een normale verdeling heeft met gemiddelde 0 en standaard- afwijking σ_E, bestaat er een intervalschatting van de ware score. Dit interval bestaat uit de getallen $\hat{\tau}$ waarvoor geldt dat de volgende nulhypothese bij een van te voren vastgesteld significantieniveau niet wordt verworpen:

$$H_0\text{: } x - z \times \sigma_E \leq \hat{\tau} \leq x + z \times \sigma_E \qquad (3.6)$$

waarin z de standaardnormale afwijking is die behoort bij het gekozen significantieniveau. Als dit bijvoorbeeld vastgesteld is op de waarde 0.05, is de waarde van z gelijk aan 1.96. Merk op dat (3.6) een schattingsvoorschrift is. Men kiest eerst de getallen z en $\hat{\tau}$, terwijl σ_E bekend is verondersteld. Dan neemt men de realisatie x_{vt} van de toevalsvariabele X waar, en vult de verkregen waarde in (3.6) in. Als de gegeven ongelijkheden worden geschonden, besluit men dat het van te voren gekozen getal $\hat{\tau}$ geen goede schatting is van de ware score. Alle getallen $\hat{\tau}$ waarvoor de ongelijkheden in (3.6) niet geschonden zijn, vormen gezamenlijk een intervalschatting voor de ware score die behoort bij de geobserveerde score x. In de praktijk berekent men natuurlijk, zodra de score x is geobserveerd, de intervalgrenzen $x \pm z \times \sigma_E$. Het zo verkregen interval heet in de statistiek een betrouwbaarheidsinterval voor de ware score; de naam heeft niets te maken met het begrip betrouwbaarheid uit de klassieke testtheorie.

Een tweede schatter voor de ware score is de zogenoemde Kelley-schatter (Kelley, 1947; Lord & Novick, 1968). Deze schatter levert een kleinere standaardfout op, maar daarvoor betaalt men wel een prijs. Men moet namelijk veronderstellen dat de regressie

van T op X lineair is. Men kan afleiden dat deze regressie de volgende gedaante heeft:

$$\mathcal{E}(T \mid X = x) = (\rho^2_{XT})\, x + (1 - \rho^2_{XT})\, \bar{x} \qquad (3.7)$$

waarin $\bar{x}$ de gemiddelde geobserveerde score is van de steekproef van personen uit de populatie $\mathcal{P}$ aan wie men de toets heeft afgenomen (zie voor de afleiding Lord en Novick, 1968, p. 65). Zoals Kelley (1947, p. 409) zegt: "This is an interesting equation in that it expresses the estimate of true ability as a weighted sum of two separate estimates - one based upon the individual's observed score, [x], and the other based upon the mean of the group to which he belongs, ... If the test is highly reliable, much weight is given to the test score and little to the group mean, and vice versa." De standaardfout van de Kelley-schatter is gelijk aan $\sigma_E\,(\rho^2_{XT})^{\frac{1}{2}}$, de spreiding van het verschil $T - \mathcal{E}(T \mid X = x)$. In de regressie-analyse noemt men deze spreiding wel de spreiding om de regressielijn. Als men de standaardfout van de Kelley-schatter substitueert voor σ_E in (3.6) verkrijgt men een andere intervalschatter voor de ware score. Deze schatter leidt tot kleinere intervallen dan de schatter uit (3.6) omdat de gebruikte standaardfout kleiner is dan de in (3.6) als standaardfout gebruikte standaard-meetfout.

In de praktijk zal men niet vaak schattingen van ware scores tegenkomen. De reden daarvan is, dat toetsscores doorgaans relatief worden geïnterpreteerd. Niet de waarde van de score zelf is van belang, maar zijn rangnummer in de verdeling van scores in de populatie $\mathcal{P}$. De beschreven schatters van de ware score leiden tot dezelfde rangorde van personen als de geobserveerde scores; daarom heeft men geen geschatte ware scores nodig. Anders wordt het als een score wordt gerelateerd aan een op voorhand gegeven criterium. Zo'n criterium is bijvoorbeeld een getal waarboven een score moet liggen om als voldoende aangemerkt te worden. Dan bestaat de mogelijkheid, door het gebruik van geschatte ware scores het aantal classificatiefouten te verminderen.

In veel boeken en artikelen over de klassieke testtheorie ziet men verwarring optreden tussen de begrippen standaardfout en standaardmeetfout. De standaardfout, die eigenlijk 'standaardfout van een schatting' (standard error of estimate) heet, is een maat voor de nauwkeurigheid van een schatter. Men kan de nauwkeurigheid van een schatter opvoeren door een grotere steekproef te trekken (hoofdstuk 2). De standaardmeetfout daarentegen is een kenmerk van een toets; het groter maken van een steekproef van aan de toets onderworpen personen heeft op de standaardmeetfout geen enkele invloed. Om de standaardmeetfout kleiner te maken moet men de betrouwbaarheid van de toets groter maken. Een van de middelen daartoe is, de toets met een aantal items te verlengen. Het verlengen van een toets wordt besproken in

paragraaf 3.6.2. De verwarring tussen de begrippen standaardfout en standaardmeetfout wordt wellicht verklaard door het feit dat de standaardmeetfout de rol speelt van standaardfout in (3.6).

3.6 Het schatten van de betrouwbaarheid en de standaardmeetfout

Er zijn diverse procedures ontwikkeld om de betrouwbaarheid en de standaardmeetfout van een toets te schatten. Men kan die grootheden immers niet precies bepalen omdat men in de praktijk alleen maar kan beschikken over een steekproef van personen uit de populatie $\mathcal{P}$. In de volgende paragrafen bespreken we methoden om de betrouwbaarheid en de standaardmeetfout te schatten uit parallelle metingen, uit twee afnames van de toets, uit toetsverlenging, en uit coëfficiënt alpha als een ondergrens van de betrouwbaarheid. In paragraaf 3.11 zullen we zien dat men ook de betrouwbaarheid kan schatten door middel van een variantie-analyse van itemscores.

3.6.1 Parallelle metingen

Een belangrijk begrip dat is toegevoegd aan de klassieke testtheorie is dat van de parallelle meting. Men beschikt niet alleen over de realisaties van de geobserveerde toetsscore X maar ook over die van een toetsscore X' die voldoet aan de volgende eigenschappen: $\mathcal{E}(X') = \mathcal{E}(X)$ en $\sigma^2_{X'} = \sigma^2_X$ in elke deelpopulatie van $\mathcal{P}$. Metingen die aan deze eigenschappen voldoen, noemt men parallelle metingen, of ook wel streng parallelle metingen. Beschouw nu de correlatie $\rho_{XX'}$ tussen parallelle metingen. De teller hiervan is gelijk aan:

$$\mathrm{Cov}(X, X') = \mathrm{Cov}(T + E, T + E') = \mathrm{Cov}(T, T) + \mathrm{Cov}(E, E') = \sigma^2_T + \mathrm{Cov}(E, E').$$

Nu wordt er verondersteld dat de bij beide metingen optredende meetfouten E en E' onder- ling onafhankelijk zijn; de meetfouten zijn niet gecorreleerd. Een correlatie ongelijk aan nul zou duiden op de aanwezigheid van een factor die beide metingen systematisch beïnvloedt. Bij parallelle metingen veronderstelt men dat zo'n factor er niet is. De meetfouten worden geacht experimenteel onafhankelijk te zijn. Experimentele onafhankelijkheid brengt met zich mee dat de meetouten niet gecorreleerd zijn. Er geldt dus: $\mathrm{Cov}(E, E') = 0$, en dus $\mathrm{Cov}(X, X') = \sigma^2_T$. De noemer van de correlatie tussen X en X' is gelijk aan: $\sigma_X \sigma_{X'} = \sigma_X \sigma_X = \sigma^2_X$. We zien hieruit dat de correlatie tussen parallelle metingen, $\rho_{XX'}$, gelijk is aan de betrouwbaarheid van

de meting X en ook aan die van de meting X'. Dit verklaart het gebruik van het symbool $\rho_{XX'}$ voor de betrouwbaarheid in veel boeken en artikelen over de klassieke testtheorie.

In de praktijk is het niet eenvoudig, parallelle metingen te construeren. Soms slaagt men er in metingen te maken die wel een paar, maar niet alle eigenschappen van parallelle metingen hebben. In tabel 3.1 zijn enige vormen van parallelliteit opgesomd, die afnemen in de strengheid van de eisen.

Tabel 3.1
Enige vormen van parallelliteit

Soort parallelliteit	Eigenschappen
Parallelliteit	$\mathscr{E}(X) = \mathscr{E}(X'),\ \sigma_X^2 = \sigma_{X'}^2$
Tau-equivalentie	$\mathscr{E}(X) = \mathscr{E}(X')$
Essentiële tau-equivalentie	$\mathscr{E}(X) = \mathscr{E}(X') + \kappa\ (\kappa \neq 0)$
Congenerieke parallelliteit	$T = \lambda T' + \kappa,\ (\lambda \neq 0)$

In deze tabel zijn κ en λ constanten die van de meetinstrumenten afhangen. De genoemde eigenschappen gelden in elke deelpopulatie van $\mathscr{P}$. Dat betekent onder meer dat voor elke persoon de ware scores op de parallelle toetsen aan elkaar gelijk zijn, en dus dat $\sigma^2(T) = \sigma^2(T')$. Uit tabel 3.1 ziet men dat men als eerste de veronderstelling laat vallen dat parallelle toetsen dezelfde geobserveerde variantie hebben en dus dezelfde foutenvariantie. Daarna verruimt men de relatie die tussen de ware scores van de beide toetsen bestaat: voor essentieel tau-equivalente metingen verschillen de ware scores een constante, terwijl voor congenerieke metingen de ware scores lineaire transformaties zijn van elkaar. Of aan de diverse vormen van parallelliteit is voldaan, kan men onderzoeken met methoden voor lineaire-structuurmodellen. Zulke methoden zijn beschreven in Bollen (1989).

In de praktijk zal men vaak moeite hebben, meetinstrumenten te maken die aan een van de genoemde definities van parallelliteit voldoen. Daarom heeft men, om de betrouwbaarheid en de standaardmeetfout van een meting X te schatten, methoden bedacht die geen gebruik maken van parallelle metingen. Een van die methoden bestaat eruit, de toets tweemaal af te nemen bij dezelfde personen. Andere methoden vereisen wel dat het mogelijk is het meetinstrument in stukken te verdelen. Bij toetsen die items bevatten, en ook als er diverse beoordelaars zijn, kan men spreken over onderdelen of deeltoetsen.

3.6.2 Test-hertestmethode

Als men niet kan beschikken over parallelvormen van een toets, kan men onder bepaalde omstandigheden dezelfde toets twee keer afnemen bij dezelfde personen. In feite beschouwt men de toets als parallel aan zichzelf. De procedure veronderstelt dat er geen leereffecten kunnen optreden tussen de twee toetsmomenten, en dat tussen die momenten in de populatie niet wezenlijk van karakter verandert. De betrouwbaarheid van de toets kan men dan eenvoudig schatten uit de correlatie tussen de twee verkregen toetsscores.

3.6.3 Toetsverlenging

Een van de methoden om de betrouwbaarheid te schatten, bestaat er uit het meetinstrument op de een of andere wijze in k parallelle delen te verdelen. Elk paar deeltoetsen heeft dezelfde correlatie ρ; deze correlatie is dan ook per definitie de betrouwbaarheid van elk der deeltoetsen. Deze betrouwbaarheid ρ wordt bekend verondersteld. In de praktijk kan dit het geval zijn als men een nieuwe toets wil samenstellen uit bestaande toetsen; een dergelijke samengestelde toets noemt men wel een verlengde toets. Als toetsscore op de verlengde toets kiest men de som van de scores op de deeltoetsen. Men kan dan het volgende afleiden. De geobserveerde variantie kan men als volgt schrijven:

$$\sigma_X^2 = \sigma^2\left(\sum_i^k X_i\right) = \sum_i^k \sigma_{X_i}^2 + \sum_{i\neq j} \mathrm{Cov}(X_i, X_j) = k\sigma_{X_i}^2 + \sum_{i\neq j} \sigma_{X_i}\sigma_{X_j}\rho =$$

$$= k\sigma_{X_i}^2 + k(k-1)\sigma_{X_i}^2\rho = k\sigma_{X_i}^2\left[1 + (k-1)\rho\right].$$

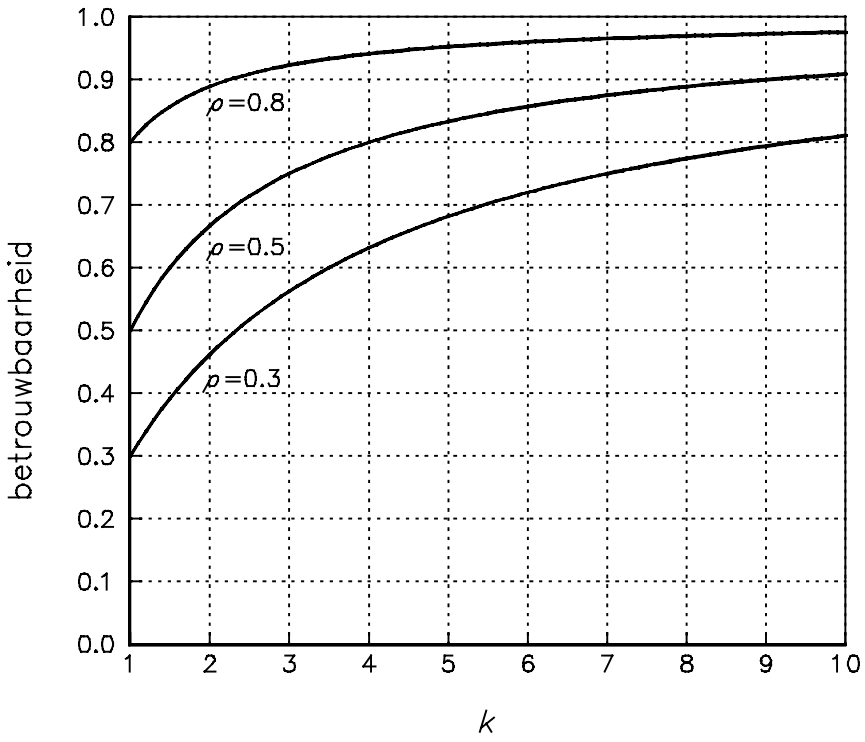
Evenzo kan men de ware variantie schrijven als:

$$\sigma_T^2 = \sigma^2\left(\sum_i^k T_i\right) = \sum_i^k \sigma_{T_i}^2 + \sum_{i\neq j} \mathrm{Cov}(T_i, T_j) = k\sigma_{T_i}^2 + k(k-1)\sigma_{T_i}^2 = k^2\sigma_{T_i}^2.$$

Als men deze twee uitdrukkingen substitueert in formule (3.4), verkrijgt men het volgende resultaat:

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} = \frac{k^2\sigma_{T_i}^2}{k\sigma_{X_i}^2\left[1 + (k-1)\rho\right]} = \frac{k\rho}{1 + (k-1)\rho}. \qquad (3.8)$$

Formule (3.8) is de Spearman-Brown-formule voor toetsverlenging (Brown, 1910; Spearman, 1910). Zij speelt een rol bij het samenstellen van toetsen uit gegeven deeltoetsen of items, vooral om te bepalen of men aan een toets in wording nog delen moet toevoegen om een bepaalde betrouwbaarheid te kunnen bewerkstelligen. In figuur 3.1 is voor een aantal waarden van ρ de betrouwbaarheid uitgezet tegen het aantal deeltoetsen k.



Figuur 3.1
Het verband tussen de lengte en de betrouwbaarheid van een toets

In de praktijk wordt de Spearman-Brown-formule voornamelijk gebruikt bij het construeren van toetsen. Een toets met k items blijkt een betrouwbaarheid ρ te hebben. Met behulp van de Spearman-Brown-formule kan men dan uitrekenen hoeveel maal men k items aan de toets moet toevoegen om een gewenste betrouwbaarheid $\rho' > \rho$ te bereiken.

3.6.4 Coëfficiënt alpha

De Spearman-Brown-formule veronderstelt dat men de betrouwbaarheid van de deeltoetsen kent. Aangezien dat in de praktijk dikwijls niet het geval is, kan men gebruik maken van de volgende ongelijkheid:

$$\rho_{XT}^2 \geq \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} \sigma_{X_i}^2}{\sigma_X^2}\right]. \qquad (3.9)$$

Het rechterlid van ongelijkheid (3.9) heet coëfficiënt alpha, of ook wel Cronbachs alpha (Cronbach, 1951). Merk op dat coëfficiënt alpha louter te schatten grootheden bevat. Met deze coëfficiënt is dus een ondergrens voor de betrouwbaarheid van een meetinstrument gegeven. De afleiding van coëfficiënt alpha bestaat uit een aantal stappen. In de eerste stap vormen we alle paren deeltoetsen, berekenen in elk paar de som van de ware varianties, en leiden voor de som van deze sommen een ongelijkheid af:

$$\sigma_{(T_i - T_j)}^2 = \sigma_{T_i}^2 + \sigma_{T_j}^2 - 2\,\mathrm{Cov}(T_i, T_j) \geq 0 \Rightarrow \sum_{i \neq j}\left[\sigma_{T_i}^2 + \sigma_{T_j}^2\right] \geq 2\sum_{i \neq j}\mathrm{Cov}(T_i, T_j).$$

De eerste ongelijkheid geldt omdat het linkerlid een variantie is, en dus nooit negatief kan zijn. In de tweede stap berekenen we opnieuw de som van sommen van ware varianties, maar nu met inbegrip van de oneigenlijke paren waarin elke deeltoets met zichzelf wordt gecombineerd. Voor de zo verkregen som leiden we weer een ongelijkheid af, waarbij de in de eerste stap afgeleide ongelijkheid wordt gebruikt:

$$\sum_i \sum_j \left[\sigma_{T_i}^2 + \sigma_{T_j}^2\right] = 2k\sum_i \sigma_{T_i}^2 = 2\sum_i \sigma_{T_i}^2 + \sum_{i \neq j}\left[\sigma_{T_i}^2 + \sigma_{T_j}^2\right] \geq$$

$$2\sum_i \sigma_{T_i}^2 + 2\sum_{i \neq j}\mathrm{Cov}(T_i, T_j) \Rightarrow (k-1)\sum_i \sigma_{T_i}^2 \geq \sum_{i \neq j}\mathrm{Cov}(T_i, T_j).$$

In de derde stap leiden we een eenvoudige ongelijkheid af voor de ware variantie:

$$\sigma_T^2 = \sigma^2\left(\sum_i T_i\right) = \sum_i \sigma_{T_i}^2 + \sum_{i \neq j}\mathrm{Cov}(T_i, T_j) \geq$$

$$\geq \frac{k}{k-1}\sum_{i \neq j}\mathrm{Cov}(T_i, T_j).$$

De som in het rechterlid van deze ongelijkheid kan als volgt worden herschreven:

$$\sum_{i \neq j}\mathrm{Cov}(T_i, T_j) = \sum_{i \neq j}\mathrm{Cov}(X_i, X_j) = \sigma_X^2 - \sum_i \sigma_{X_i}^2.$$

Als we alle ongelijkheden substitueren in formule (3.4), is het resultaat de volgende ongelijkheid:

$$\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_X} \geq \frac{k}{k-1}\left[1 - \frac{\sum_i \sigma^2_{X_i}}{\sigma^2_X}\right]. \qquad (3.10)$$

Als men coëfficiënt alpha beschouwt als een schatter van de betrouwbaarheid, kan men de standaardmeetfout schatten met: $\hat{\sigma}_E = \hat{\sigma}_X\sqrt{(1-\alpha)}$.

In het rechterlid van (3.10), dat gelijk is aan coëfficiënt alpha, ziet men de varianties optreden van de verschillende deeltoetsen. Er is niet verondersteld dat deze varianties aan elkaar gelijk zijn. In feite is het voldoende dat de deeltoetsen essentieel tau-equivalent zijn, als gedefinieerd in tabel 3.1.

Coëfficiënt alpha wordt wel een maat voor de interne consistentie van een toets genoemd. Men noemt een toets intern consistent als de items in de toets niet alle een correlatie van 0 met elkaar hebben. Men kan laten zien dat coëfficiënt alpha op de volgende manier kan worden geschreven:

$$\alpha = \frac{\bar{c}\langle X_i, X_j\rangle}{\sigma^2_{\bar{X}}}. \qquad (3.11)$$

In (3.11) is de teller, $\bar{c}\langle X_i, X_j\rangle$, gelijk aan het gemiddelde van de covarianties tussen alle paren itemscores: $\bar{c}(X_i, X_j) = [k(k-1)]^{-1}\sum_{i\neq j}\mathrm{Cov}(X_i, X_j)$. De noemer is gelijk aan de variantie van het gemiddelde van de itemscores: $\bar{X} = k^{-1}\sum_{i=1}^{k} X_i$. Als alle items onderling perfect correleren, zijn alle varianties van de itemscores aan elkaar gelijk, zijn de covarianties tussen de items gelijk aan deze varianties, en is de gemiddelde itemscore gelijk aan elk der itemscores. Uit (3.11) blijkt dat coëfficiënt alpha in dat geval gelijk is aan 1. Een enkele keer komt men in de literatuur de opvatting tegen dat een toets met een hoge interne consistentie, dus met een hoge waarde van coëfficiënt alpha, een enkele factor in de zin van de factoranalyse meet. Dat deze opvatting op een misverstand berust, is overtuigend aangetoond door Green en Lissitz (1977).

3.7 Toets- en itemanalyse

De toets- en itemanalyse is de praktische uitvoering van het schatten van de in de voorafgaande paragrafen beschreven grootheden. Aangezien in de praktijk toetsen

bestaan uit opgaven of items, worden ook kengetallen voor items berekend. Deze laatste grootheden spelen een belangrijke rol in het proces van toetsconstructie. Zij vormen niet alleen de bouwstenen van schattingsformules voor de betrouwbaarheid en de standaardmeetfout, maar zijn ook op zichzelf beschouwd van belang om eigenschappen van items te beschrijven. Doorgaans bepaalt men de kengetallen van items en toetsen in een proefafname: een concepttoets wordt aan een groep personen afgenomen, en op basis van de verkregen gegevens worden de grootheden van de items en de toets geschat. Zonodig worden er items herzien of wordt de samenstelling van de toets veranderd.

In deze paragraaf worden eerst de toets- en itemindices van een toets met meerkeuzevragen besproken. Daarna komen de indices van een toets met open vragen aan de orde voor zover deze niet besproken zijn bij de toets met meerkeuzevragen. In paragraaf 3.8 worden de betrouwbaarheid en de standaardmeetfout apart besproken. Omdat de toets- en itemindices veelal gebaseerd zijn op steekproeven, is paragraaf 3.9 gewijd aan standaardfouten van de geschatte toets- en itemindices. In paragraaf 3.10 tenslotte schenken we aandacht aan normen en richtlijnen voor diverse toets- en itemindices.

Aangezien er in een toets- en itemanalyse voortdurend sprake is van schattingen van grootheden op basis van de gegevens van een steekproef van personen, zal dikwijls de conventie worden gevolgd, de schatters aan te duiden met gewone letters. Zo zal een (schatter van de) variantie worden geschreven als s^2 en niet als $\hat{\sigma}^2$.

3.7.1 Toets- en itemindices bij toetsen met meerkeuzevragen

Toetsen met meerkeuzevragen bestaan uit vragen of items waarbij een persoon het goede antwoord moet kiezen uit verschillende alternatieven. We gaan er van uit dat elk goed beantwoord item 1 scorepunt oplevert en elk fout beantwoord item 0 scorepunten. De som van de itemscores vormt de toetsscore van een persoon. De toets- en itemindices worden besproken aan de hand van een toets die een tweekeuze-item en twee driekeuze-items bevat. De toets is door vier personen gemaakt. Dit is weliswaar geen realistische situatie maar het stelt de lezer in staat de indices na te rekenen. De itemantwoorden staan in tabel 3.2. In de kop van deze tabel zijn de goede antwoorden, samen wel de sleutel genoemd, vermeld. De itemantwoorden zijn met behulp van de sleutel omgezet in itemscores. Deze staan samen met de toetsscores in tabel 3.3.

Tabel 3.2

Antwoorden per persoon en per item (tussen haakjes de sleutel)

Tabel 3.3

Itemscores en toetsscores

persoon	item		
	1(B)	2(A)	3(C)
1	B	A	C
2	B	A	A
3	B	B	B
4	A	C	A

persoon	item			toetsscore
	1	2	3	
1	1	1	1	3
2	1	1	0	2
3	1	0	0	1
4	0	0	0	0
som	3	2	1	6

De resultaten van de toets- en itemanalyse van de gegevens uit tabel 3.3 staan in tabel 3.4. De indices uit deze tabel worden in de volgende deelparagraaf besproken.

Tabel 3.4

Resultaten toets- en itemanalyse van de toets met meerkeuzevragen

	p- en a-waarden			discriminatie-indices				r_{ir}- en r_{ar}-waarden		
item	A	B	C	s_i	r_{it}	r_{ir}	eff	A	B	C
1	0.25	0.75*		0.43	0.77	0.52	0.30	–0.52	0.52*	
2	0.50*	0.25	0.25	0.50	0.89	0.71	0.40	0.71*	0.00	–0.82
3	0.50	0.25	0.25*	0.43	0.77	0.52	0.30	–0.30	–0.17	0.52*

aantal personen : 4 gemiddelde p-waarde : 0.50
gemiddelde toetsscore : 1.50 betrouwbaarheid (KR-20) : 0.75
standaardafwijking : 1.12 standaardmeetfout : 0.56

3.7.2 Itemindices bij toetsen met meerkeuzevragen

In tabel 3.4 staan de waarden voor de moeilijkheid van een item en de aantrekkelijkheid van de afleiders onder de kop 'p- en a-waarden'. Bij elk alternatief is de fractie personen vermeld die het alternatief heeft gekozen. De fractie waarbij een ster (*) staat, hoort bij het goede antwoord en wordt de p-waarde van het item genoemd. De p-waarde wordt berekend door het aantal personen dat het item goed heeft, te delen door het aantal personen dat het item heeft gemaakt. De bij de afleiders of foute antwoorden vermelde fracties worden de a-waarden van het item genoemd en worden berekend door het aantal personen dat een afleider heeft gekozen te delen door het aantal personen dat het item heeft gemaakt. Bij item 2 in ons voorbeeld, een driekeuze-item, zien we bij de alternatieven A, B en C respectievelijk de waarden

0.50*, 0.25 en 0.25 staan. Dit betekent dat alternatief A het goede antwoord is met een p-waarde van 0.50. De a-waarden van de alternatieven B en C zijn beide gelijk aan 0.25.

Een p-waarde ligt per definitie tussen 0 en 1. Bij een p-waarde gelijk aan 0 hebben alle personen het item fout; bij een p-waarde gelijk aan 1 hebben alle personen het item goed. Het kan voorkomen dat een item een afleider heeft met een a-waarde die groter is dan de p-waarde. Dit kan er op wijzen dat een afleider niet fout is of dat het als goed bestempelde alternatief wellicht niet goed is. In het algemeen geeft een hoge a-waarde ons informatie over het item die in combinatie met andere informatie tot een definitief oordeel over de kwaliteit van het item moet leiden.

Onder het kopje 's_i' is de standaardafwijking van de items vermeld. De standaardafwijking van een item, s_i, wordt bij dichotome scores berekend als: $s_i = \sqrt{p\,q} = \sqrt{p(1-p)}$, waarin p de p-waarde van het item is en q gelijk is aan 1 - p. Wanneer alle personen een item goed dan wel fout hebben, is de standaardafwijking gelijk aan 0. De standaardafwijking is maximaal als p = 0.50, dus als de ene helft van de personen het item fout heeft en de andere helft het item goed. In dat geval is $s_i = \sqrt{0.5(1-0.5)} = 0.5$.

Omdat een item een onderdeel van een toets is, zijn er diverse indices ontwikkeld om de samenhang tussen een itemscore en de toetsscore weer te geven. Een index die veel gebruikt wordt is de r_{it}. De r_{it} is de produkt-moment-correlatie tussen de itemscore en de toetsscore. Deze correlatie wordt bij dichotoom gescoorde items wel puntbiseriële correlatie genoemd: het is de correlatie tussen een dichotome en een continu geachte variabele. Een produkt- moment-correlatie neemt waarden aan tussen +1 en -1. Een correlatie van +1 betekent dat er een perfect positief lineair verband bestaat tussen twee variabelen, in ons geval tussen de itemscore en de toetsscore. Dat de r_{it}-waarden in tabel 3.4 zo hoog zijn, heeft te maken met het feit dat de toets uit slechts drie items bestaat. Bij toetsen van veertig of meer items is een r_{it} van 0.50 al hoog (zie tabel 3.12).

De r_{it} wordt een discriminatie-index genoemd omdat zij aangeeft in hoeverre een item onderscheid maakt tussen personen met hoge toetsscores en personen met lage toetsscores. Een hoge r_{it} betekent dat veel personen met een hoge toetsscore het item goed hebben beantwoord en veel personen met een lage toetsscore het item fout hebben beantwoord. Later zullen we zien dat een hoge r_{it} ook betekent dat het item relatief veel bijdraagt aan de betrouwbaarheid van de toets (zie paragraaf 3.8.1).

Hiervoor zagen we dat de r_{it} een produkt-moment-correlatie is. Die kan met een van de algemene formules voor een correlatie berekend worden. Afgeleid kan worden dat voor dichotome scores de r_{it} van een item ook geschreven kan worden als:

$$r_{it} = \frac{\bar{x}_g - \bar{x}_f}{s_x} \sqrt{p(1-p)}\ , \tag{3.12}$$

waarin:
$\bar{x}_g$ = gemiddelde toetsscore van de personen die het item goed hebben,
$\bar{x}_f$ = gemiddelde toetsscore van de personen die het item fout hebben,
s_x = standaardafwijking van de toetsscores.

De teller in het deel voor het wortelteken in (3.12) maakt duidelijk waarom we de r_{it} een discriminatie-index noemen: hoe groter het verschil tussen $\bar{x}_g$ en $\bar{x}_f$, des te groter de r_{it}.

Naast de r_{it} is de r_{ir} een veel gebruikte discriminatie-index. De r_{ir} is een soortgelijke index als de r_{it}. Gaat het bij de r_{it} om de correlatie tussen itemscores en toetsscores, bij de r_{ir} gaat het om de correlatie tussen itemscores en restscores. De restscore van een persoon is gelijk aan zijn toetsscore minus de score op het desbetreffende item. Een persoon heeft dus evenzoveel restscores als er items zijn in de toets.

Zowel aan de r_{it} als aan de r_{ir} kleven bezwaren. De r_{it} geeft een geflatteerd beeld van de samenhang tussen de score op een item en de toetsscore, omdat de itemscore onderdeel is van de toetsscore. We correleren dus het item voor een deel met zichzelf. De r_{ir} ondervangt dit bezwaar, maar heeft als bezwaar dat de restscore waarmee een item gecorreleerd wordt, met het item varieert. De r_{ir}-waarden van eenzelfde toets zijn daardoor onderling niet te vergelijken. Als echter het aantal items in een toets veertig of meer is, zijn beide bezwaren van geen belang meer.

Nog een andere maat om het discriminerend vermogen van een item te karakteriseren is het effectieve gewicht dat te vinden is onder het kopje 'eff'. Onder het effectieve gewicht verstaan we de bijdrage van een item aan de spreiding van toetsscores. Hoe hoger het effectieve gewicht van een item is, des meer spreiding in de toetsscores toegeschreven kan worden aan het item. Het volgende kan worden afgeleid (Gulliksen, 1950; Ferguson & Takane, 1989):

$$\sum_{i=1}^{k} r_{it} s_i = s_x, \tag{3.13}$$

waarin k het aantal items is.

Het effectieve gewicht van item i is gedefinieerd als:

$$\frac{r_{it} \times s_i}{s_x}. \tag{3.14}$$

De teller in (3.14) wordt de itembetrouwbaarheidsindex genoemd en is een onderdeel van de formule om de betrouwbaarheid van de toets te schatten (zie paragraaf 3.8.1). Uit (3.14) volgt dat de som van de effectieve gewichten gelijk is aan 1. In ons voorbeeld van tabel 3.4 heeft item 2 een effectief gewicht van 0.40; dat betekent dat het item voor 40% bijdraagt aan de standaardafwijking van de toetsscores. Een andere interpretatie van het effectieve gewicht wordt gegeven door regressie-analyse. Als men de lineaire regressievergelijking van de itemscore op de toetsscore opstelt, blijkt de regressiecoëfficiënt gelijk te zijn aan het effectieve gewicht van het item.

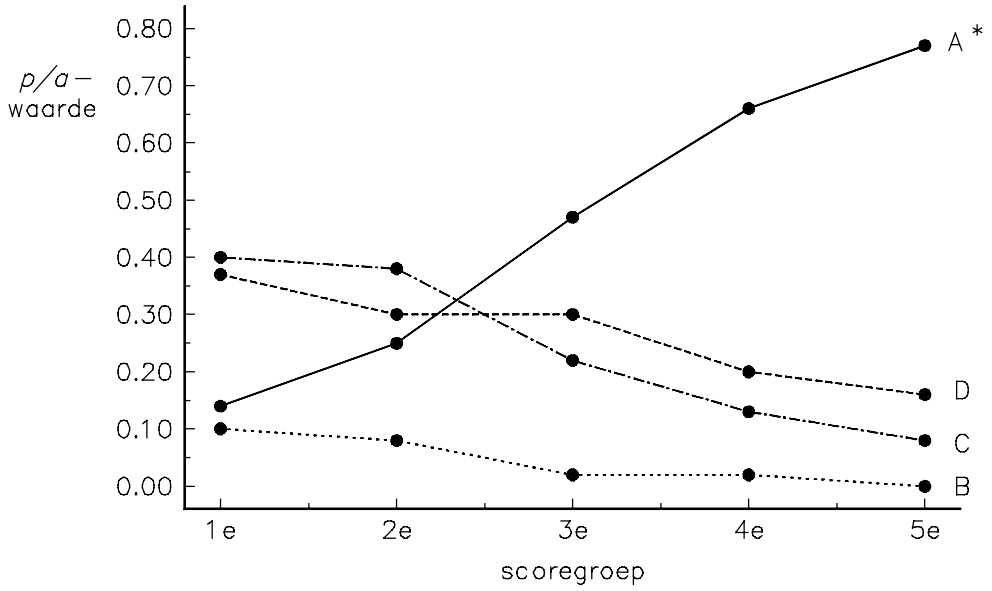
Bij een toets met meerkeuzevragen is het mogelijk, naast een discriminatie-index voor het goede antwoord discriminatie-indices voor de afleiders (foute antwoorden) te berekenen. In tabel 3.4 kunnen we zien dat er bij elk item r_{ar}-waarden zijn vermeld naast de r_{ir}-waarde. Per item zijn er uiteraard evenveel r_{ar}-waarden als er afleiders zijn. De r_{ar} wordt berekend door personen die het desbetreffende foute antwoord hebben gekozen een itemscore 1 en de anderen een itemscore 0 te geven. Vervolgens wordt de correlatie tussen het foute antwoord en de restscore berekend, waarbij de restscore per definitie dezelfde waarde heeft als bij de berekening van de r_{ir}. Omdat we toetsen met een hoge betrouwbaarheid nastreven, zijn items met positieve r_{ir}- en negatieve r_{ar}-waarden gewenst. Zulke waarden impliceren dat relatief veel personen met een hoge toetsscore het item goed hebben beantwoord en relatief veel personen met een lage toetsscore het item fout hebben beantwoord. Een positieve r_{ar} geeft aan dat relatief veel goede personen de desbetreffende afleider als het goede antwoord hebben aangemerkt. Soms kan dit een sleutelfout zijn: de verkeerde sleutel is per ongeluk opgegeven of bij nader inzien blijkt dat de afleider met de positieve r_{ar} het goede antwoord is.

Tabel 3.5

Per scoregroep de p- en a-waarden van een item

score	n	A*	B	C	D
0 - 18	123	0.14	0.10	0.40	0.37
19 - 22	124	0.25	0.08	0.38	0.30
23 - 29	124	0.47	0.02	0.22	0.30
30 - 35	124	0.66	0.02	0.13	0.20
36 - 47	124	0.77	0.00	0.08	0.16
0 - 47	619	0.46*	0.04	0.24	0.26
gem. score	26.0	30.8	18.8	21.0	23.5

Het discriminerend vermogen van een item kunnen we ook weergeven door de personen in een aantal scoregroepen op te delen en vervolgens per scoregroep de *p*- en *a*-waarden te berekenen. Als voorbeeld presenteren we in tabel 3.5 van een item de *p*- en *a*-waarden per scoregroep. In die tabel lezen we dat alternatief A het goede antwoord is met een *p*-waarde van 0.46. Van de afleiders is D het meest aantrekkelijk met een *a*-waarde van 0.26. Verder zien we dat de totale groep van 619 personen is opgesplitst in vijf bijna even grote scoregroepen. Bekijken we nu van het item de *p*-waarde per scoregroep, dan heeft het item in de minst vaardige groep, met scores tussen 0 en 18, een *p*-waarde van 0.14. De *p*-waarde van het item wordt groter met het vaardiger worden van de groep, en in de meest vaardige groep heeft het item een *p*-waarde van 0.77. Bij de afleiders is de tendens andersom; hoe vaardiger de groep, des te lager de *a*-waarde. Het item is dus een voorbeeld van een goed discriminerend item: de *p*-waarde van het item is in de groep van de beste personen veel hoger dan in de groep van de slechtste personen, en de *a*-waarden van het item zijn voor de slechtste personen hoger dan de *a*-waarden voor de beste personen. De *p*- en *a*-waarden uit tabel 3.5 zijn grafisch weergegeven in figuur 3.2. De keuze van het aantal scoregroepen is arbitrair. Om er echter voor te zorgen dat de standaardfout van een fractie niet te groot wordt, moet het aantal personen per scoregroep niet te klein zijn (zie tabel 3.8).



Figuur 3.2
Per scoregroep *p*- en *a*-waarden van het item uit tabel 3.5

3.7.3 Toetsindices bij toetsen met meerkeuzevragen

Behalve informatie over de drie afzonderlijke items uit de toets, bevat tabel 3.4 ook informatie die betrekking heeft op de toets als geheel. We kunnen in de tabel lezen dat vier personen, n = 4, de toets gemaakt hebben. Een maat voor de moeilijkheidsgraad van een toets is de gemiddelde toetsscore $\bar{x}$, die bij deze toets gelijk is aan 6/4=1.50. De standaardafwijking van de toetsscores, s_x, is een maat voor de spreiding van de toetsscores en kan als volgt berekend worden:

$$s_x = \left(\frac{\sum_{v=1}^{n} (x_v - \bar{x})^2}{n} \right)^{1/2}, \qquad (3.15)$$

waarin x_v de toetsscore is van persoon v.

De standaardafwijking kan volgens (3.13) ook verkregen worden door de itembetrouwbaar-
heidsindices te sommeren. Wanneer de standaardafwijking gelijk is aan 0, hebben alle personen dezelfde toetsscore. De standaardafwijking is maximaal wanneer de ene helft van de personen alle items goed heeft en de andere helft alle items fout.

De gemiddelde p-waarde, $\bar{p}$, is het gemiddelde van de p-waarden van de afzonderlijke items. Bij toetsen met meerkeuzevragen kan de gemiddelde p-waarde berekend worden hetzij door alle p-waarden op te tellen en de som te delen door het aantal items k, hetzij door de gemiddelde toetsscore te delen door het aantal items in de toets. In formulevorm:

$$\bar{p} = \frac{\sum_{i=1}^{k} p_i}{k} \text{ of } \bar{p} = \frac{\bar{x}}{k}. \qquad (3.16)$$

De toetsindices betrouwbaarheid en standaardmeetfout worden in paragraaf 3.8 besproken.

3.7.4 Toets- en itemindices bij toetsen met open vragen

Bij toetsen met open vragen moeten personen zelf het antwoord formuleren op de vragen die voorgelegd worden. Het is gebruikelijk dat er per vraag meer dan een

scorepunt behaald kan worden en dat de antwoorden door beoordelaars met behulp van een correctievoorschrift gescoord worden. In deze paragraaf gaan we er van uit dat beoordelaars geen factor zijn die de meetprocedure verstoren. In dat geval is er ook geen wezenlijk verschil tussen de analyse van een toets met open vragen en de analyse van een toets met meerkeuzevragen. Het enige verschil is dat er bij open vragen andere itemscores dan alleen maar 0 en 1 mogelijk zijn. Indien beoordelaars wel een storende factor zijn, dient er een analyse als beschreven in paragraaf 3.13 plaats te vinden.

In het voorbeeld in tabel 3.6 gaan we uit van vier open vragen die door zes personen beantwoord zijn. Op elke vraag kunnen maximaal twintig punten behaald worden.

Tabel 3.6

Itemscores en toetsscores

persoon	item				toetsscore
	1	2	3	4	
1	17	8	14	3	42
2	16	10	13	5	44
3	18	15	14	18	65
4	16	14	14	8	52
5	14	7	7	4	32
6	17	15	17	16	65
som	98	69	79	54	300

De resultaten van de toets- en itemanalyse staan in tabel 3.7. Aangezien de toets- en itemanalyse van open vragen voor een deel dezelfde indices bevat als de toets- en itemanalyse van meerkeuzevragen, komen hierna niet meer alle toets- en itemindices aan de orde. Alleen de voor open vragen specifieke indices worden besproken.

Tabel 3.7

Resultaten van de toets- en itemanalyse van de toets met open vragen

item	max. score	gem. score	p'	s_i	r_{it}	r_{ir}	eff
1	20.00	16.33	0.82	1.25	0.81	0.77	0.08
2	20.00	11.50	0.58	3.30	0.95	0.91	0.26
3	20.00	13.17	0.66	3.02	0.81	0.69	0.20
4	20.00	9.00	0.45	5.89	0.94	0.79	0.46

aantal personen : 6 gemiddelde p'-waarde : 0.63

gemiddelde toetsscore	: 50.00	betrouwbaarheid (alpha)	: 0.82
standaardafwijking	: 12.10	standaardmeetfout	: 5.12

3.7.5 Itemindices bij toetsen met open vragen

Bij een toets met open vragen kan het aantal te behalen scorepunten van vraag tot vraag variëren. Daarom is in tabel 3.7 een kolom met het opschrift 'max. score' opgenomen. In deze kolom staat het aantal punten dat op een item behaald kan worden. In het voorbeeld zijn bij alle items de maxima gelijk.

Een andere voor open vragen specifieke index staat in de kolom met opschrift 'gem. score'. In deze kolom staat de gemiddelde score die op elk van de items behaald is. Bij ongelijke maximale scores zijn de gemiddelde itemscores niet vergelijkbaar. Daarom wordt de p'-waarde berekend; deze staat in de kolom met het opschrift 'p''. De p'-waarde duidt de moeilijkheidsgraad van een item aan, en wordt berekend door de gemiddelde itemscore te delen door de maximale itemscore. Merk op dat we bij open vragen over de p'-waarde spreken en bij meerkeuzevragen over de p-waarde. De definitie van de twee grootheden is gelijk; het verschil in notatie heeft geen andere functie dan aan te geven om welke soort vraag het gaat.

3.7.6 Toetsindices bij toetsen met open vragen

Bij toetsen met open vragen worden dezelfde toetsindices berekend als bij toetsen met meerkeuzevragen. Om misverstanden te voorkomen, verdient de berekening van de gemiddelde p'-waarde enige toelichting. De gemiddelde p'-waarde wordt berekend door de gemiddelde toetsscore te delen door de maximaal te behalen toetsscore. In tegenstelling tot bij een toets met meerkeuzevragen mag de gemiddelde p'-waarde bij een toets met open vragen alleen maar op deze manier berekend worden en niet via de p'-waarden van de individuele vragen. Als men dat wel zou doen, zou men verschillen in maximaal te behalen itemscores veronachtzamen.

3.8 Betrouwbaarheid en standaardmeetfout

Bij de toets- en itemanalyse van de meerkeuzevragen is de KR-20 als betrouwbaarheidsmaat berekend en bij de toets- en itemanalyse van de open vragen coëfficiënt alpha. Hierna laten we zien dat de KR-20 een speciaal geval is van coëfficiënt alpha. In paragraaf 3.5 zijn twee manieren besproken om met behulp van de standaardmeetfout een intervalschatting voor de ware score te bepalen. Deze twee manieren worden in paragraaf 3.8.3 gebruikt om intervalschattingen te verkrijgen voor ware verschilscores.

3.8.1 Coëfficiënt alpha en de KR-20

Het is gebruikelijk, de betrouwbaarheid van een toets met coëfficiënt alpha te schatten. De formule voor coëfficiënt alpha is gegeven in het rechterlid van (3.9). Omdat bij dichotoom gescoorde vragen geldt dat $s_i^2 = p_i q_i$, kan coëfficiënt alpha voor dichotoom gescoorde items geschreven worden als:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} p_i q_i}{s_x^2}\right]. \qquad (3.17)$$

Formule (3.17) staat bekend als de KR-20 en is onafhankelijk van Cronbachs coëfficiënt alpha door Kuder en Richardson (1937) ontwikkeld. Vanwege (3.12) kan coëfficiënt alpha ook geformuleerd worden als:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} s_i^2}{\left(\sum_{i=1}^{k} r_{it} s_i\right)^2}\right]. \qquad (3.18)$$

Uit (3.18) laat zich het verband tussen de r_{it} en de betrouwbaarheid nog niet eenvoudig aflezen. Bij dichotoom gescoorde items liggen de itemvarianties in de praktijk tussen 0.21 en 0.25 ($0.3 < p < 0.7$). Indien we de itemvarianties nu als constant beschouwen voor alle items, kunnen we afleiden (Thorndike, 1982):

$$\alpha \approx \frac{k}{k-1}\left(1 - \frac{1}{k(\bar{r}_{it})^2}\right), \qquad (3.19)$$

waarin $\bar{r}_{it}$ het gemiddelde van de r_{it}-waarden is.

3.8.2 Verschilscores

In paragraaf 3.5 zijn schattingen van de ware score aan de orde geweest. Er is op gewezen dat het schatten van ware scores niet altijd nodig is. In de praktijk zou men willen weten of een toetsscore van 30 voor Kay en een toetsscore van 33 voor Wilko betekent dat de laatstgenoemde meer weet dan Kay. Daar kan men niet achter komen, omdat men de ware scores van Kay en Wilko niet kent. Wel kan men iets zeggen over het volgende probleem. Als men aselect twee personen uit de populatie trekt waarvan de waargenomen scores drie punten verschillen, kan men dan zeggen of dit verschil substantieel is? Statistisch gezien betekent dit dat we de nulhypothese willen toetsen dat de ware toetsscores van de twee aselect getrokken personen gelijk zijn. Noem deze ware scores τ_1 en τ_2, en de geobserveerde scores x_1 en x_2. Veronderstel dat de geobserveerde scores x_1 en x_2 normaal verdeeld zijn met verwachte waarden τ_1 respectievelijk τ_2, en beide met standaardafwijking σ_E. Dan is de verschilscore $x_1 - x_2$ normaal verdeeld met gemiddelde $\tau_1 - \tau_2$ en standaardafwijking $\sigma_E\sqrt{2}$. Naar analogie van (3.6) kunnen we een intervalschatting maken van het verschil $\delta = \tau_1 - \tau_2$. Dit interval bestaat uit alle waarden $\hat{\delta}$ waarvoor de volgende nulhypothese niet wordt verworpen:

$$H_0\text{: } (x_1 - x_2) - z \times \sigma_E\sqrt{2} \leq \hat{\delta} \leq (x_1 - x_2) + z \times \sigma_E\sqrt{2}\,.$$

Veronderstel dat de toets een standaardmeetfout σ_E heeft van 1, dan vindt men, bij een verschil van drie punten in geobserveerde scores, het 95%-betrouwbaarheidsinterval: $0.23 \leq \tau_1 - \tau_2 \leq 5.77$. Aangezien dit interval niet de waarde 0 bevat, zal men bij een waargenomen verschil van drie punten, de hypothese verwerpen dat de bijbehorende ware scores aan elkaar gelijk zijn.

Men kan ook een intervalschatting voor verschilscores bepalen op basis van de in paragraaf 3.5 genoemde Kelley-schatter. Men kan afleiden dat de verschilscore $\delta = \tau_1 - \tau_2$ een verwachte waarde heeft gelijk aan $\rho^2_{XT}(x_1 - x_2)$ en een standaardafwijking gelijk aan $(2\rho^2_{XT}\sigma^2_E)^{½}$. Voor een toets met een betrouwbaarheid van 0.80 en een standaardmeetfout van 1 is, bij een verschil in waargenomen scores van 3 punten, het 95%-betrouwbaarheidsinterval gelijk aan: $-0.08 \leq \tau_1 - \tau_2 \leq 4.88$. Nu zal men de nulhypothese van gelijke ware scores niet verwerpen. Merk op dat het laatst

gegeven betrouwbaarheidsinterval iets kleiner is dan het eerst gegeven interval: 4.96 tegenover 5.54.

3.9 Nauwkeurigheid van toets- en itemindices

Bij het berekenen van toets- en itemindices is het buitengewoon belangrijk dat men er zich rekenschap van geeft hoe nauwkeurig die indices geschat zijn. De statistiek geeft ons op deze vraag een antwoord omdat het mogelijk is betrouwbaarheidsintervallen te construeren. Zoals reeds eerder is aangegeven, is een betrouwbaarheidsinterval een stochastisch interval om een steekproefwaarde dat met een gegeven kans de te schatten populatiewaarde bevat. De p-waarde, de gemiddelde score, de r_{it}-waarde, de KR-20 en coëfficiënt alpha zijn allemaal voorbeelden van grootheden die gebaseerd zijn op steekproeven en daardoor behept met steekproeffouten. In de volgende paragrafen zullen we op deze steekproeffouten en op de constructie van betrouwbaarheidsintervallen ingaan.

3.9.1 Standaardfout van een p-waarde

De standaardfout s_p van een p-waarde wordt met de volgende formule berekend:

$$s_p = \left(\frac{p(1-p)}{n}\right)^{½}. \qquad (3.20)$$

In (3.20) staat n voor het aantal personen in de aselect getrokken steekproef. Nu zegt een vuistregel in de statistiek dat, indien $n > \{9 \times (1-p)/p\}$ bij $p \leq 0.50$ en $n > \{9 \times p/(1-p)\}$ bij $p \geq 0.50$, een p-waarde bij benadering normaal verdeeld is. Hiervan uitgaande, kunnen we een betrouwbaarheidsinterval construeren voor de werkelijke p-waarde. Veronderstel dat de geschatte p-waarde van een item 0.20 is en dat het item door 100 personen is gemaakt, dan is de bijbehorende standaardfout $\sqrt{0.2 \times 0.8/100} = 0.04$. We kunnen dan bijvoorbeeld de grenzen van het 95%-betrouwbaarheidsinterval berekenen. Uit de berekening volgt dat in 95% van de gevallen bij items met een geschatte p-waarde van 0.20 de werkelijke p-waarde tussen 0.12 en 0.28 zal liggen (0.12 = 0.20 - 1.96 × 0.04 en 0.28 = 0.20 +1.96 × 0.04). In tabel 3.8, die gebaseerd is op exacte berekeningen (De Jonge, 1963), kan men bij $p = 0.20$ en $n = 100$ aflezen dat de grenzen 0.13 en 0.29 zijn. De afwijkingen zijn minimaal.

Tabel 3.8

95%-betrouwbaarheidsintervallen voor fracties

steekproef -fractie p	aantal personen in de steekproef (n)									
	50		100		200		500		1000	
0.00	0.00	0.07	0.00	0.04	0.00	0.02	0.00	0.01	0.00	0.00
0.10	0.03	0.22	0.05	0.18	0.06	0.15	0.08	0.13	0.08	0.12
0.20	0.10	0.34	0.13	0.29	0.15	0.26	0.17	0.24	0.18	0.23
0.30	0.18	0.45	0.21	0.40	0.24	0.37	0.26	0.34	0.27	0.33
0.40	0.26	0.55	0.30	0.50	0.33	0.47	0.36	0.45	0.37	0.43
0.50	0.35	0.65	0.40	0.60	0.43	0.57	0.46	0.55	0.47	0.53
0.60	0.45	0.74	0.50	0.70	0.53	0.67	0.55	0.64	0.57	0.63
0.70	0.55	0.82	0.60	0.79	0.63	0.76	0.66	0.74	0.67	0.73
0.80	0.66	0.90	0.71	0.87	0.74	0.85	0.76	0.83	0.77	0.82
0.90	0.78	0.97	0.82	0.95	0.85	0.94	0.87	0.92	0.88	0.92
1.00	0.93	1.00	0.96	1.00	0.98	1.00	0.99	1.00	1.00	1.00

3.9.2 Standaardfout van een gemiddelde toetsscore en van een p'-waarde

De standaardfout $s_{\bar{x}}$ van de gemiddelde toetsscore $\bar{x}$ is gelijk aan:

$$s_{\bar{x}} = \frac{s_x}{\sqrt{n}}. \tag{3.21}$$

Neem als voorbeeld een toets die door 429 personen gemaakt is, en waarvan de gemiddelde toetsscore gelijk is aan 32.24 en de standaardafwijking van de toetsscores 6.29 is. De standaardfout bedraagt dan 0.30 en het 95%-betrouwbaarheidsinterval heeft de grenzen 31.64 en 32.84.

De standaardfout s_p van een p'-waarde is gelijk aan:

$$s_{p'} = \frac{s_i}{m\sqrt{n}}. \tag{3.22}$$

In (3.22) staat m voor de maximaal te behalen score op de vraag. Bij de toets met open vragen in tabel 3.7 heeft item 4 een p'-waarde van 0.45. We kunnen daarvan de standaardfout berekenen; deze bedraagt 0.12. Het 95%-betrouwbaarheidsinterval voor de werkelijke p'-waarde heeft de grenzen 0.14 en 0.76. Dit interval is groot omdat zo weinig personen het item gemaakt hebben.

3.9.3 Standaardfout van een r_{it}-waarde

De berekening van de standaardfout van een r_{it}-waarde is nogal gecompliceerd. In Iker en Perry (1960) staan benaderingsformules en tabellen voor de standaardfout.

Tabel 3.9

95%-betrouwbaarheidsintervallen voor r_{it}-waarden

r_{it}-waarde (steekproef)	aantal personen in de steekproef (n)							
	100		200		500		1000	
0.00	-0.20	0.20	-0.14	0.14	-0.08	0.08	-0.06	0.06
0.10	-0.10	0.30	-0.04	0.24	0.02	0.18	0.04	0.16
0.20	0.00	0.40	0.06	0.34	0.12	0.28	0.14	0.26
0.30	0.12	0.48	0.18	0.42	0.22	0.38	0.24	0.36
0.40	0.24	0.56	0.28	0.52	0.32	0.48	0.34	0.46
0.50	0.36	0.64	0.40	0.60	0.44	0.56	0.46	0.54
0.60	0.48	0.72	0.51	0.69	0.54	0.66	0.56	0.64

Tabel 3.9 is gebaseerd op Iker en Perry, en is van toepassing op p-waarden die tussen 0.20 en 0.80 liggen. In tabel 3.9 staan voor diverse waarden van de r_{it} en n de 95%-betrouwbaarheids- intervallen voor de werkelijke waarden van de r_{it} vermeld. Indien bijvoorbeeld bij een toets- en itemanalyse die gebaseerd is op 1000 personen, de r_{it}-waarde van een item 0.20 is, dan zijn de 95%-betrouwbaarheidsgrenzen van de werkelijke r_{it}-waarde 0.14 en 0.26.

3.9.4 Standaardfout van coëfficiënt alpha

Voor coëfficiënt alpha heeft Feldt (1965) de steekproefverdeling afgeleid waarop tabel 3.10 gebaseerd is. In deze tabel zijn bij diverse steekproefwaarden van coëfficiënt alpha de onder- en bovengrenzen vermeld van het 95%-betrouwbaarheidsinterval voor de werkelijke waarde van coëfficiënt alpha. De tabel mag alleen gebruikt worden indien een toets tien of meer vragen bevat. Als bijvoor- beeld de betrouwbaarheid van een toets die is afgenomen bij 500 personen gelijk is aan 0.70, dan loopt het 95%-betrouwbaarheidsinterval van 0.66 tot 0.74.

Tabel 3.10

95%-betrouwbaarheidsintervallen voor coëfficiënt alpha

α (steekproef)	aantal personen in de steekproef (n)							
	100		200		500		1000	
0.10	-0.17	0.33	-0.09	0.27	-0.02	0.21	0.02	0.18
0.20	-0.04	0.41	0.03	0.35	0.10	0.30	0.13	0.27
0.30	0.09	0.48	0.25	0.43	0.21	0.38	0.24	0.30
0.40	0.22	0.55	0.27	0.51	0.32	0.47	0.35	0.45
0.50	0.35	0.63	0.40	0.59	0.44	0.56	0.45	0.54
0.60	0.48	0.70	0.52	0.67	0.55	0.65	0.56	0.63
0.70	0.61	0.78	0.64	0.76	0.66	0.74	0.67	0.73
0.80	0.74	0.85	0.76	0.84	0.77	0.82	0.78	0.82
0.90	0.87	0.93	0.88	0.92	0.89	0.91	0.89	0.91

3.10 Normen voor toets- en itemindices

In de volgende paragrafen worden normen en richtlijnen voor toets- en itemindices geformuleerd. We moeten bedenken dat deze normen en richtlijnen opgesteld zijn met de gedachte dat we er naar moeten streven een toets met een zo hoog mogelijke betrouwbaarheid te construeren. Nogmaals dient er op gewezen te worden dat de indices bij kleine aantallen personen een relatief kleine precisie hebben, zodat voorzichtigheid geboden is bij de interpretatie van zulke indices.

3.10.1 Normen voor p- en p'-waarden

In de literatuur vinden we verschillende opvattingen over de optimale p-waarde van een item. Crocker en Algina (1986) stellen dat de optimale p-waarde halverwege de raadkans en 1.0 moet liggen. De veronderstelling hierbij is dat er geraden wordt als men niet weet wat het goede antwoord op een meerkeuze-item is. In formulevorm uitgedrukt: $p = 0.5 + 0.5/m$, waarin m het aantal alternatieven is en p de gewenste p-waarde. Naar aanleiding van een simulatie-onderzoek komt Lord (1952) tot een andere conclusie. De aanbevelingen van voornoemde auteurs over de optimale p-waarde van items met verschillende aantallen alternatieven staan in tabel 3.11.

De conclusie van een onderzoek van Feldt (1993) is, dat de optimale p-waarde tussen 0.57 en 0.67 moet liggen wanneer er geraden kan worden. Indien er geen reden is om aan te

Tabel 3.11

Optimale p-waarde bij items met 2-5 alternatieven

aantal alternatieven	optimale p-waarde (p=0.5+0.5/m)	optimale p-waarde (Lord)
2	0.75	0.85
3	0.67	0.77
4	0.63	0.74
5	0.60	0.70

nemen dat er geraden wordt, of als er niet geraden kan worden zoals bij open vragen, is de
optimale p-waarde gelijk aan 0.50. Het effect van de moeilijkheid van een item op de betrouwbaarheid blijkt echter verbazingwekkend klein te zijn, zelfs als de p-waarden variëren van 0.27 tot 0.79.

3.10.2 Normen voor r_{it}-waarden

Ook voor r_{it}-waarden vindt men in de literatuur geen absolute normen. Zoals bekend kan een produkt-moment-correlatie, dus ook een r_{it}-waarde, variëren tussen -1 en +1. Een r_{it}-waarde van 0.50 en hoger is echter in de praktijk bij toetsen met meer dan veertig items al erg hoog. Ebel en Frisbie (1986) komen tot de in tabel 3.12 vermelde normen voor de r_{it}-waarden.

Tabel 3.12

Normen voor r_{it}-waarden

r_{it}-waarde	itembeoordeling
0.40 en hoger	zeer goed
0.30 - 0.39	goed
0.20 - 0.29	twijfelachtig
0.19 en lager	slecht

Omdat de grootte van de r_{it} onder andere afhankelijk is van het aantal items in een toets, moet men strikt genomen bovenstaande normen alleen hanteren bij r_{it}-waarden die gecorrigeerd zijn voor toetslengte. De correctie kan uitgevoerd worden met een correctie-formule van Henrysson (1963). Vanwege het geringe effect kan de correctie achterwege blijven indien de items afkomstig zijn uit toetsen met veertig of meer items.

3.10.3 Normen voor de betrouwbaarheid

In de literatuur wordt 0.85 als vereiste ondergrens voor de betrouwbaarheid van een toets genoemd wanneer de vaardigheid van een groep personen op basis van slechts een enkele toets wordt bepaald. Wanneer de vaardigheid met meer toetsen of op verschillende momenten wordt getoetst zijn lagere ondergrenzen acceptabel, waarbij in de literatuur 0.65 wel als gewenste ondergrens wordt genoemd (Frisbie, 1988).

Een mogelijke norm voor de betrouwbaarheid zouden we kunnen ontlenen aan het percentage ten onrechte gezakte en ten onrechte geslaagde personen, ofwel het percentage niet-consistente beslissingen, bij een selectietoets (Dousma & Horsten, 1989). Met de ten onrechte gezakte en de ten onrechte geslaagde personen bedoelen we de personen waarvoor, indien ze een parallelle toets hadden afgelegd, de beslissing anders geweest had kunnen zijn. Het percentage niet-consistente beslissingen neemt toe als de betrouwbaarheid lager wordt en ook als het percentage gezakten stijgt, waarbij het percentage gezakten afhangt van de cesuur of grensscore. Tabel 3.13 laat de percentages niet-consistente beslissingen zien als functie van het percentage gezakten en van de betrouwbaarheid. Daarbij moet opgemerkt worden dat het gebruik van de tabel alleen zinvol is wanneer de toetsscores ongeveer normaal verdeeld zijn.

Tabel 3.13

Percentages niet-consistente beslissingen als functie van het percentage gezakten en de betrouwbaarheid

percentage gezakten	betrouwbaarheid						
	0.0	0.50	0.60	0.70	0.80	0.90	1.00
5	10	8	7	6	5	4	0
10	18	14	12	11	9	6	0
15	26	18	17	14	12	8	0
20	32	23	20	17	14	10	0
25	38	26	23	20	16	11	0
30	42	29	25	22	18	12	0
35	46	31	27	23	19	13	0
40	48	32	29	24	20	14	0
45	50	33	29	25	20	14	0
50	50	33	30	25	20	14	0

In tabel 3.13 kunnen we zien dat bij een toets met een betrouwbaarheid van 0.80 en met een percentage gezakten van 30, het percentage niet-consistente beslissingen gelijk aan 18 is. Dat wil dan zeggen dat 9% van de gezakten tot de geslaagden zou kunnen hebben behoord en 9% van de geslaagden tot de gezakten. Dus voor 18% van alle personen had de beslissing anders kunnen zijn.

3.11 Generaliseerbaarheidstheorie

De bespreking van de generaliseerbaarheidstheorie, (Cronbach, Gleser, Nanda & Rajaratnam, 1972), in dit hoofdstuk bestaat uit vier paragrafen. Het begrippenkader dat in de generaliseerbaarheidstheorie gehanteerd wordt en dat in belangrijke mate ontleend is aan de variantie-analytische literatuur, wordt in deze paragraaf besproken. In paragraaf 3.12 wordt de generaliseerbaarheidstheorie behandeld aan de hand van de analyse van de toets met meerkeuzevragen die in paragraaf 3.7 met de klassieke testtheorie geanalyseerd is. In paragraaf 3.13 wordt de generaliseerbaarheidstheorie verder toegelicht aan de hand van een analyse van een toets waarbij beoordelaars de antwoorden van personen op vragen beoordelen. In beide paragrafen wordt aandacht besteed aan verschillen tussen de klassieke testtheorie en generaliseerbaarheidstheorie. In paragraaf 3.14 komen kort een aantal andere aspecten van de generaliseerbaarheidstheorie aan de orde. Merk op dat de notatie die in de paragrafen 3.11 tot en met 3.14 gehanteerd wordt afwijkt van die uit voorgaande paragrafen. De reden hiervoor, is de notatie aan te laten sluiten bij de in de literatuur gebruikelijke notatie.

In de generaliseerbaarheidstheorie worden observaties of metingen beschreven in termen van de condities waaronder zij geobserveerd worden. Condities van een bepaalde soort worden aangeduid als 'facet'. De dertig meerkeuzevragen van een toets zijn volgens deze terminologie de dertig condities van het facet 'vragen'. En bij een toets bestaande uit tien open vragen waarbij de antwoorden door twee beoordelaars beoordeeld worden, spreken we over de tien condities van het facet 'vragen' en de twee condities van het facet 'beoordelaars'. Het door personen laten beantwoorden van vragen, kunnen we opvatten als een gestandaardiseerd experiment (Meerling, 1981). Een proefopzet waarin responsen of antwoorden van personen op (condities van het facet) vragen worden geobserveerd, wordt een een-facet-design genoemd. Een proefopzet waarin de observaties beoordelingen zijn van responsen van personen op (condities van het facet) vragen die beoordeeld worden door (condities van het facet) beoordelaars, wordt een twee-facet-design genoemd. Het aantal observaties dat per

persoon verkregen wordt, is afhankelijk van het design dat gebruikt wordt. Wanneer we aan tien personen een toets van dertig vragen voorleggen, een zogenaamd gekruist een-facet-design (personen × vragen), hebben we per persoon dertig observaties. Zouden we echter aan elke persoon drie andere vragen voorleggen, dan hebben we per persoon slechts drie observaties. Wanneer we aan tien personen een toets van tien vragen voorleggen en de responsen op de tien vragen laten beoordelen door twee beoordelaars, een zogenaamd gekruist twee-facet-design (personen × vragen × beoordelaars), krijgen we twintig observaties per persoon. Zouden we echter vijf vragen door de eerste beoordelaar en vijf andere vragen door de tweede beoordelaar laten beoordelen, dan krijgen we tien observaties per persoon.

Voor het bepalen van de rekenvaardigheid van personen, kunnen we antwoorden van personen op meerkeuzevragen observeren. De verzameling van alle denkbare observaties die naar onze mening acceptabel of geschikt zijn voor het geven van een oordeel over personen, wordt in de generaliseerbaarheidstheorie het universum genoemd. Uiteraard zouden we het bepalen van de rekenvaardigheid van personen willen baseren op de observaties of scores verkregen op alle vragen uit het universum, de universumscores. Om praktische redenen kunnen we de personen echter niet meer dan een steekproef van bijvoorbeeld dertig vragen uit het universum voorleggen. Het bepalen van de rekenvaardigheid baseren we op de scores die op de dertig vragen behaald worden, de geobserveerde scores. De nauwkeurigheid waarmee we menen te kunnen generaliseren van geobserveerde scores naar universumscores, dat wil zeggen de geobserveerde scores kunnen opvatten als universumscores, wordt 'generaliseerbaarheid' genoemd. Als maat voor de generaliseerbaarheid wordt de generaliseerbaarheidscoëfficiënt gebruikt. Deze coëfficiënt heeft een benedengrens van 0 en een bovengrens van 1.

In het geval van de meerkeuzevragen bestaat het universum alleen uit het facet vragen. Bestaat het universum niet uit meerkeuzevragen maar uit open vragen waarvan de antwoorden door beoordelaars beoordeeld moeten worden, dan kunnen we de beoordeling door alle in aanmerking komende beoordelaars laten verrichten. In dit geval bestaat het universum uit twee facetten: het facet 'open vragen' en het facet 'beoordelaars.' De universumscores zijn gelijk aan de scores die verkregen zouden zijn na het beoordelen van alle antwoorden op alle open vragen door alle beoordelaars. Aangezien we in de praktijk de beoordeling zullen moeten beperken tot een klein aantal beoordelaars, zijn de geobserveerde scores van de personen de scores verkregen na het beoordelen van de open vragen door dit kleine aantal beoordelaars.

De voorbeelden laten zien dat voor het generaliseren naar een universum een duidelijke beschrijving van het universum een voorwaarde is. Deze beschrijving bevat

in de eerste plaats de facetten waaruit het universum bestaat. In het eerste voorbeeld bestaat het universum alleen uit het facet 'vragen'. In het tweede voorbeeld bestaat het universum uit de facetten 'vragen' en 'beoordelaars'. In de tweede plaats moet een beschrijving van het universum uitsluitsel geven over de condities die binnen het universum vallen. Dit heeft te maken met het belangrijke onderscheid dat in de variantie-analyse aangeduid wordt met de termen 'random' en 'fixed'. In het eerste voorbeeld zijn de vragen uit de toets opgevat als een aselecte of random steekproef uit een zeer grote verzameling of 'oneindig universum' van vragen. In het tweede voorbeeld zijn vragen en beoordelaars opgevat als een random steekproef uit een oneindig universum van vragen en beoordelaars. In het voorbeeld van de meerkeuzevragen impliceert een random facet dat we vinden dat ook dertig andere vragen in aanmerking hadden kunnen komen om de rekenvaardigheid van personen te bepalen. Deze twee (of meer) toetsen van dertig vragen worden in de generaliseerbaarheidstheorie random parallelle toetsen genoemd. Voor het voorbeeld van de open vragen betekent een random facet 'open vragen' en een random facet 'beoordelaars' dat we vinden dat ook tien andere open vragen en twee andere beoordelaars in aanmerking hadden kunnen komen om de vaardigheid te bepalen. Zouden we in het tweede voorbeeld vinden dat slechts twee bepaalde beoordelaars in aanmerking komen, dan spreken we van een fixed facet 'beoordelaars'. Bij een fixed facet hebben we alle condities van een facet in ons design opgenomen en hoeven dan ook niet te generaliseren naar het universum. Later zullen we zien dat het onderscheid tussen random en fixed facetten consequenties voor de generaliseerbaarheid heeft.

3.12 Design met een facet

In een gekruist een-facet-design wordt de geobserveerde score van een persoon op een item, X_{pv}, uitgedrukt als een decompositie in vier componenten:

$$\begin{aligned} X_{pv} &= \mu && = \text{algemeen gemiddelde} \\ &+ \mu_p - \mu && = \text{persoonseffect} \\ &+ \mu_v - \mu && = \text{itemeffect} \\ &+ X_{pv} - \mu_p - \mu_v + \mu && = \text{residu} \end{aligned} \tag{3.23}$$

In (3.23) is de eerste component, het algemene gemiddelde, gedefinieerd als $\mu \equiv \mathcal{E}_p \mathcal{E}_v X_{pv}$, de gemiddelde score (= verwachting over personen en items) verkregen na het beantwoorden van alle items uit het universum door alle personen uit de

populatie. Het algemene gemiddelde geeft dezelfde constante bijdrage aan de geobserveerde score van alle personen.

De universumscore van een persoon is hier gedefinieerd als $\mu_p \equiv \mathscr{E}_v X_{pv}$, de gemiddelde score (= verwachting over items) van een persoon verkregen na het beantwoorden van alle items uit het universum van items. De tweede component, het persoonseffect $\mu_p - \mu$, is gelijk aan het verschil tussen de universumscore van een persoon en het algemene gemiddelde. Personen met een positief persoonseffect hebben een score die hoger is dan het algemene gemiddelde terwijl personen met een negatief persoonseffect een score hebben die lager is dan het algemene gemiddelde. Verschillen in vaardigheid tussen personen kunnen we weergeven als verschillen tussen hun persoonseffecten.

De moeilijkheidsgraad van een item is gedefinieerd als $\mu_v \equiv \mathscr{E}_p X_{pv}$, de gemiddelde score (= verwachting over personen) van een item na het beantwoorden van het item door alle personen uit de populatie. De derde component, het itemeffect $\mu_v - \mu$, is gelijk aan het verschil tussen de moeilijkheidsgraad van een item en het algemene gemiddelde. Een item met een positief itemeffect is gemakkelijker dan een item met een negatief itemeffect. Verschillen in moeilijkheidsgraad tussen items kunnen we weergeven als verschillen tussen hun itemeffecten.

De vierde component, de foutencomponent of het residu, is het verschil tussen X_{pv} en de eerste drie componenten. Zoals we in het voorbeeld van tabel 3.15 zullen zien, beschikken we bij het gekruiste een-facet-design maar over een enkele observatie voor elke combinatie van persoon en vraag. Dit betekent dat we het persoons- × itemeffect niet kunnen onderscheiden van andere foutenbronnen. Behalve het persoons- × itemeffect bevat het residu alle foutencomponenten die de geobserveerde score doen afwijken van de som van de eerste drie componenten.

Met uitzondering van het algemene gemiddelde, hebben de componenten in (3.23) een verdeling. Uit de wijze waarop de effecten in (3.23) gedefinieerd zijn, volgt dat hun gemiddelden gelijk zijn aan nul. De definitie van het gemiddelde van het persoonseffect bijvoorbeeld luidt $\mathscr{E}_p(\mu_p - \mu) = \mathscr{E}_p(\mu_p) - \mathscr{E}_p(\mu) = \mu - \mu = 0$. De drie componenten hebben ook elk een eigen variantie die we aanduiden met variantiecomponent. De variantiecomponenten voor respectievelijk personen, items en het residu zijn gedefinieerd als:

$$\sigma_p^2 = \mathscr{E}_p(\mu_p - \mu)^2, \tag{3.24}$$

$$\sigma_v^2 = \mathscr{E}_v(\mu_v - \mu)^2, \text{ en} \tag{3.25}$$

$$\sigma^2_{pv,e} = \mathcal{E}_p \mathcal{E}_v (X_{pv} - \mu_p - \mu_v + \mu)^2. \quad (3.26)$$

De notatie van de variantiecomponent voor het residu laat zien dat de component uit een variantiecomponent personen × vragen en een variantiecomponent voor de fouten (error) bestaat.

De variantie van de geobserveerde scores is gedefinieerd als

$$\sigma^2_X = \sigma^2_{(X_{pv})} = \mathcal{E}_p \mathcal{E}_v (X_{pv} - \mu)^2,$$

en deze totale variantie is gelijk aan de som van de drie variantiecomponenten, ofwel

$$\sigma^2_X = \sigma^2_p + \sigma^2_v + \sigma^2_{pv,e}. \quad (3.27)$$

3.12.1 Generaliseerbaarheidsstudie

Om schattingen van de variantiecomponenten van effecten te verkrijgen, dienen we een onderzoek, of wat wel genoemd wordt een generaliseerbaarheidsstudie of G-studie, uit te voeren. Het schatten gebeurt met behulp van procedures uit de variantie-analyse. We bespreken hieronder een gekruist design waarbij n_p personen en n_v items of vragen aselecte steekproeven zijn uit respectievelijk een populatie van personen en een universum van items. Tabel 3.14 bevat de variantie-analysetabel van dit gekruist random-effecten-design.

Tabel 3.14
Variantie-analysetabel van een gekruist design met twee random effecten

Effecten	Kwadraten-sommen	Vrijheids-graden	Gemiddelde kwadratensommen	Verwachte gemiddelde kwadratensommen
Personen *(p)*	SS_p	$df_p = n_p - 1$	$MS_p = SS_p/df_p$	$\mathcal{E}(MS_p) = \sigma^2_{pv,e} + n_v\sigma^2_p$
Items *(v)*	SS_v	$df_v = n_v - 1$	$MS_v = SS_v/df_v$	$\mathcal{E}(MS_v) = \sigma^2_{pv,e} + n_p\sigma^2_v$
Residu *(pv,e)*	$SS_{pv,e}$	$df_{pv,e} = (n_p - 1) \times (n_v - 1)$	$MS_{pv,e} = SS_{pv,e}/df_{pv,e}$	$\mathcal{E}(MS_{pv,e}) = \sigma^2_{pv,e}$

Schattingen van de variantiecomponenten krijgen we door het oplossen van vergelijkingen voor de verwachte gemiddelde kwadratensommen (expected mean squares). Daartoe worden de verwachte gemiddelde kwadratensommen gelijkgesteld aan de geobserveerde gemiddelde kwadratensommen (mean squares) en de exacte waarden van de variantiecomponenten vervangen door de geschatte waarden. Dit resulteert in de volgende vergelijkingen:

$$MS_{pv,e} = \hat{\sigma}^2_{pv,e},$$

$$MS_v = \hat{\sigma}^2_{pv,e} + n_p\hat{\sigma}^2_v, \text{ ofwel } \hat{\sigma}^2_v = (MS_v - MS_{pv,e})/n_p,$$

$$MS_p = \hat{\sigma}^2_{pv,e} + n_v\hat{\sigma}^2_p, \text{ ofwel } \hat{\sigma}^2_p = (MS_p - MS_{pv,e})/n_v.$$

Omdat de gemiddelde kwadratensom voor het residu gelijk is aan de schatting van de variantiecomponent voor het residu, $\hat{\sigma}^2_{pv,e} = MS_{pv,e}$, kunnen we de vergelijking voor de gemiddelde kwadratensom voor de items schrijven als $\hat{\sigma}^2_v = (MS_v - \hat{\sigma}^2_{pv,e})/n_p$. Door in deze vergelijking de gemiddelde kwadratensom van de items, berekend door het uitvoeren van een variantie-analyse, en de geschatte waarde voor de variantiecomponent van het residu in te vullen, verkrijgen we een schatting van de variantiecomponent voor items. Door herschrijven van de vergelijking voor de gemiddelde kwadratensom van de personen als $\hat{\sigma}^2_p = (MS_p - \hat{\sigma}^2_{pv,e})/n_v$, verkrijgen we op analoge wijze een schatting van de variantiecomponent voor personen.

In tabel 3.14 zien we, dat we om de drie variantiecomponenten te kunnen schatten, over de kwadratensommen (sums of squares) dienen te beschikken. Daartoe vervangen we de drie parameters μ, μ_p en μ_v in (3.14) door hun geobserveerde equivalenten, wat resulteert in de volgende decompositie:

$$X_{pv} = \bar{X} + (\bar{X}_p - \bar{X}) + (\bar{X}_v - \bar{X}) + (X_{pv} - \bar{X}_p - \bar{X}_v + \bar{X}). \qquad (3.28)$$

We illustreren de berekening van de kwadratensommen aan de hand van het voorbeeld in tabel 3.15. Deze tabel bevat de itemscores die vier personen op drie items behaald hebben. Daarnaast bevat de tabel de volgende statistische grootheden: de toetsgemiddelden, $\bar{X}_p$, van de vier personen, de itemgemiddelden, $\bar{X}_v$, van de drie items en het algemene gemiddelde, $\bar{X}$. Merk op dat het voorbeeld gelijk aan is aan het voorbeeld dat in paragraaf 3.7 bij de behandeling van de klassieke testtheorie besproken is. Voor de observaties en grootheden in deze tabel hebben we vergelijking (3.24) uitgeschreven in tabel 3.16.

De kwadratensom voor personen berekenen we door de getallen uit de kolom $(\bar{X}_p - \bar{X})$ van tabel 3.16 te kwadrateren en dan te sommeren.

Tabel 3.15

De itemscores van vier personen op drie items, de gemiddelde score per persoon en per item en het algemene gemiddelde

	Item			
Persoon	1	2	3	$\bar{X}_p$
1	1	1	1	1.00
2	1	1	0	.67
3	1	0	0	.33
4	0	0	0	.00
$\bar{X}_v$	.75	.50	.25	0.50 = $\bar{X}$

Op analoge wijze verkrijgen we de kwadratensom voor de items uit de kolom $(\bar{X}_v - \bar{X})$, en die voor het residu uit de kolom $(X_{pv} - \bar{X}_p - \bar{X}_v + \bar{X})$.

Tabel 3.16

Vergelijking (3.28) uitgeschreven voor de observaties en grootheden uit tabel 3.15

$X_{pv} =$	$\bar{X}$	$+(\bar{X}_p - \bar{X})$	$+ (\bar{X}_v - \bar{X})$	$+ (X_{pv} - \bar{X}_p - \bar{X}_v + \bar{X})$
$X_{11} = 1 =$	.500	+ .500	+ .250	— .250
$X_{12} = 1 =$	.500	+ .500	+ .000	+ .000
$X_{13} = 1 =$	.500	+ .500	— .250	+ .250
$X_{21} = 1 =$	.500	+ .167	+ .250	+ .083
$X_{22} = 1 =$	.500	+ .167	+ .000	+ .333
$X_{23} = 0 =$	.500	+ .167	— .250	— .417
$X_{31} = 1 =$	.500	— .167	+ .250	+ .417
$X_{32} = 0 =$	.500	— .167	+ .000	— .333
$X_{33} = 0 =$	.500	— .167	— .250	— .083
$X_{41} = 0 =$	.500	— .500	+ .250	— .250
$X_{42} = 0 =$	.500	— .500	+ .000	+ .000
$X_{43} = 0 =$	.500	— .500	— .250	+ .250

Voor de berekening van de totale kwadratensom brengen we in vergelijking (3.28) het algemene gemiddelde naar het linkerlid waardoor we in tabel 3.16 een nieuwe kolom, $(X_{pv} - \bar{X})$, krijgen. De getallen in deze kolom worden gekwadrateerd en daarna gesommeerd. De totale kwadratensom, SS_{tot}, is gelijk aan de som van de drie andere kwadratensommen en wordt geschreven als:

$$\Sigma_p \Sigma_v (X_{pv} - \bar{X})^2 = n_v \Sigma_p (\bar{X}_p - \bar{X})^2 + n_p \Sigma_v (\bar{X}_v - \bar{X})^2 + \Sigma_p \Sigma_v (X_{pv} - \bar{X}_p - \bar{X}_v + \bar{X})^2,$$

of:

$$\Sigma_p \Sigma_v (X_{pv} - \bar{X})^2 = SS_p + SS_v + SS_{pv,e}.$$

Tabel 3.17 bevat de resultaten van de generaliseerbaarheidsstudie voor de data uit tabel 3.15.
We laten het aan de lezer over de resultaten in tabel 3.17 na te rekenen. In de laatste kolom van de tabel staan de schattingen van de variantiecomponenten voor de drie effecten. Aangezien de grootte van de componenten afhangt van de scoreschaal die gebruikt wordt, geeft de absolute grootte van de variantiecomponenten ons geen bruikbare informatie.

Tabel 3.17
Resultaten generaliseerbaarheidsstudie voor data uit tabel 3.15

Effecten	Kwadraten-sommen	Vrijheids-graden	Gemiddelde kwadratensommen	Schattingen van variantiecomponenten
Personen *(p)*	1.667	3	0.555	$\hat{\sigma}^2_p$ = 0.139 (45.5%)
Items *(v)*	0.500	2	0.250	$\hat{\sigma}^2_v$ = 0.028 (9%)
Residu *(pv,e)*	0.833	6	0.139	$\hat{\sigma}^2_{pv,e}$ = 0.139 (45.5%)

Vandaar dat we voor elke component de procentuele bijdrage aan de totale variantie vermelden. In verband met de interpretatie van de variantiecomponenten willen we er met verwijzing naar de definities (3.24)-(3.27) nog eens benadrukken dat de variantie-componenten het resultaat zijn van de decompositie van de geschatte totale variantie van scores van afzonderlijke personen op afzonderlijke items. Dit betekent dus dat $\hat{\sigma}^2_v$ en $\hat{\sigma}^2_{pv,e}$ geen variantiecomponenten van gemiddelde of totaalscores zijn. Merk op dat we de items dichotoom gescoord hebben, zodat de variantiecomponenten in de tabel nooit groter kunnen zijn dan 0.25. De variantiecomponent voor de personen, de

geschatte universumscore-variantie, bedraagt bijna de helft van de totale variantie. De geschatte variantiecomponent voor de items is relatief klein. De geschatte variantiecomponent voor het residu is ook relatief groot. Deze variantiecomponent bestaat uit de interactiecomponent personen × vragen en andere foutenvariantie. Wanneer het residu louter uit de interactiecomponent zou bestaan, zou dit betekenen dat de rangorde van de personen niet voor alle items gelijk is. Dit zou in het voorbeeld het geval geweest zijn wanneer de eerste persoon het derde item fout en de vierde persoon het derde item goed beantwoord zou hebben.

3.12.2 Decisiestudie

Tot nu toe had de bespreking uitsluitend betrekking op de decompositie van een score van een persoon op een item uit het universum van items. Een persoon krijgt echter altijd een toets voorgelegd die uit een aantal items bestaat. Decisies of beslissingen over een persoon zijn dan ook altijd gebaseerd op de gemiddelde score of de totaalscore die behaald is op dat aantal items. In ons voorbeeld bestaat de toets uit drie random getrokken rekenitems uit het universum van rekenitems. Een andere toets met ook drie random getrokken items uit hetzelfde universum zouden we ook geschikt gevonden hebben voor het meten van de rekenvaardigheid. Dit betekent dat het universum waar in dit geval naar gegeneraliseerd wordt, het universum van random parallelle toetsen met drie items is.

Het lineaire model voor de decompositie van de gemiddelde score van een persoon op een toets met n_v items, aangeduid met X_{pV}, luidt:

$$X_{pV} = \mu + (\mu_p - \mu) + (\mu_V - \mu) + (X_{pV} - \mu_p - \mu_V + \mu). \qquad (3.29)$$

Vergelijking (3.29) is gelijk aan vergelijking (3.23) met dit verschil dat we in (3.29) de score, behaald op een enkel item, vervangen hebben door de gemiddelde score behaald op n_v items. In de notatie van (3.29) wordt een hoofdletter *V* gebruikt om aan te geven dat het de gemiddelde score van n_v items betreft. In (3.29) wordt de universumscore gedefinieerd als $\mu_p = \mathcal{E}_V X_{pV}$, de verwachte waarde van X_{pV} over random parallelle toetsen. De definities van de variantiecomponenten zijn gelijk aan die van (3.24), (3.25) en (3.26) met dien verstande dat *v* vervangen is door *V*. Het spreekt vanzelf dat door bij (3.24) de verwachting over *V* te nemen, de universumscorevariantie σ_p^2 niet verandert. De twee andere variantiecomponenten zijn: $\sigma_V^2 = \sigma_v^2 / n_v$ en $\sigma_{pV,e}^2 = \sigma_{pV,e}^2 / n_v$. Deze twee variantiecomponenten hebben betrekking op de populatie van personen en

het universum van random parallelle toetsen. De variantiecomponent $\sigma^2_V = \sigma^2_v / n_v$ moet geïnterpreteerd worden als de variantie van de verdeling van gemiddelde scores van random parallelle toetsen. De totale variantie, $\sigma^2_X = \sigma^2_{(XpV)}$ is gelijk aan $\sigma^2_X = \sigma^2_p + \sigma^2_V + \sigma^2_{pV,e}$. Wat het voorgaande betekent voor ons voorbeeld, hebben we samengevat in tabel 3.18.

In tabel 3.18 zien we hoe groot de variantiecomponenten die we in de generaliseerbaarheids-studie (G-studie) geschat hebben, in een zogenaamde decisiestudie (D-studie) worden wanneer de toets uit n_v items bestaat. Voor een gekruist een-facet-random-effect design zijn twee decisies of beslissingen van belang: de beslissing of we de toets voor het nemen van relatieve of absolute beslissingen zullen gebruiken en de beslissing uit hoeveel items we onze toets moeten laten bestaan.

Tabel 3.18

Resultaten decisiestudie voor data uit tabel 3.15

Effecten	Variantiecomponenten G-studie	Variantiecomponenten D-studie
Personen *(p)*	$\hat{\sigma}^2_p = 0.139$	$\hat{\sigma}^2_p = 0.139$
Items *(v)*	$\hat{\sigma}^2_v = 0.028$	$\hat{\sigma}^2_V = 0.028/3 = .009$
Residu *(pv,e)*	$\hat{\sigma}^2_{pv,e} = 0.139$	$\hat{\sigma}^2_{pV,e} = 0.139/3 = .046$

Het doel van een toets kan zijn, vast te stellen hoe de prestatie van een persoon zich verhoudt tot de prestaties van andere personen. Wanneer beslissingen over personen gebaseerd zijn op wat personen presteren in relatie tot andere personen, spreken we van relatieve beslissingen. De mate waarin we er met de toets in slagen personen van elkaar te onderscheiden, drukken we uit in een generaliseerbaarheidscoëfficiënt voor relatieve beslissingen. Voor het gekruiste één-facet-random-effect-design is de schatting van deze generaliseerbaarheidscoëfficiënt, een ratio van variantiecomponenten, gedefinieerd als:

$$\hat{\rho}^2 = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + \frac{\hat{\sigma}^2_{pv,e}}{n_v}}. \qquad (3.30)$$

De noemer van (3.30) bevat de universumscorevariantie $\hat{\sigma}^2_p$ en de foutenvariantie $\hat{\sigma}^2_{pv,e}/n_v$. Merk op dat de variantiecomponent $\hat{\sigma}^2_v/n_v$ niet als foutenvariantie in de noemer van (3.30) voorkomt. De reden hiervoor is dat verschillen in gemiddelde scores van random parallelle toetsen geen rol spelen wanneer we personen met elkaar willen

vergelijken. Wanneer we willen beslissen of Jan beter kan rekenen dan Piet, dan maakt het niet uit of we ze een toets met makkelijke of een toets met moeilijke items voorleggen. Brennan (1992, p. 16) laat formeel zien dat verschillen tussen scores van personen de voor beiden gelijke itemcomponent doet wegvallen.

We kunnen aan (3.30) zien dat we de coëfficiënt kunnen verhogen door de toets uit meer items laten bestaan waardoor de foutenvariantie kleiner zal worden. Omdat (3.30) een schatting van de generaliseerbaarheidscoëfficiënt na toetsverlenging geeft, wordt de formule ook wel de 'stepped-up generalizability coëfficiënt' genoemd. In hoofdstuk 11 laten we zien hoe (3.30) herschreven en gebruikt kan worden als de Spearman-Brown-formule voor toetsverlenging uit de klassieke testtheorie.

In tabel 3.18 zien we dat voor de toets met drie items de universumscorevariantie gelijk is aan .139, en de foutenvariantie aan .139/3 = .046. De generaliseerbaarheidscoëfficiënt is gelijk aan .139/{.139 + .046} = 0.75. De generaliseerbaarheidscoëfficiënt kan op twee manieren geïnterpreteerd worden. De eerste interpretatie is dat de coëfficiënt bij benadering gelijk is aan de verwachte waarde van de gekwadrateerde correlatie tussen geobserveerde en universumscores. Daarnaast kan de coëfficiënt geïnterpreteerd worden als de correlatie tussen de scores van twee random parallelle toetsen, elk bestaande uit n_v items.

Met behulp van de gemiddelde kwadratensommen kunnen we (3.30) ook uitdrukken als:

$$\hat{\rho}^2 = \frac{MS_p - MS_{pv,e}}{MS_p}. \qquad (3.31)$$

Bewezen kan worden dat in het geval van dichotome scores (3.31) gelijk is aan de KR-20 en in het geval van polytome scores aan Cronbachs coëfficiënt alpha (Sirotnik, 1970).

Het doel van de toets kan ook zijn, vast te stellen of personen in staat zijn een bepaalde prestatie te leveren, bijvoorbeeld tachtig procent van de items uit het universum goed te beantwoorden. In deze situatie zijn we niet geïnteresseerd in wat een persoon presteert in vergelijking met andere personen, maar in het absolute prestatieniveau van de persoon. Beslissingen die gebaseerd zijn op het absolute prestatieniveau van een persoon worden absolute beslissingen genoemd. In dit geval spelen verschillen in toetsen wel degelijk een rol bij de beslissing of personen aan het gewenste prestatieniveau voldoen. Wanneer een toets namelijk uit makkelijke items bestaat, kan eerder aan het prestatieniveau voldaan worden dan wanneer de toets uit moeilijke items bestaat. Dit betekent dat wanneer met een toets absolute beslissingen over personen genomen worden, $\hat{\sigma}_v^2/n_v$ bijdraagt aan de foutenvariantie.

De schatting van de generaliseerbaarheidscoëfficiënt voor absolute beslissingen is gedefinieerd als:

$$\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_v^2}{n_v} + \frac{\hat{\sigma}_{pv,e}^2}{n_v}}. \qquad (3.32)$$

Door de variantiecomponenten uit tabel 3.18 in (3.32) in te vullen, verkrijgen we de generaliseerbaarheidscoëfficiënt voor de toets uit ons voorbeeld. De coëfficiënt is gelijk aan .139/{.139 + .028/3 + .139/3} = 0.72. Merk op dat de coëfficiënten voor relatieve en absolute beslissingen slechts weinig verschillen. Dit verschil wordt uiteraard nog kleiner als we de toets verlengen.

Het onderscheid tussen relatieve en absolute beslissingen wijst op een belangrijk verschil tussen de generaliseerbaarheidstheorie en de klassieke testtheorie. De assumptie van parallelle toetsen in de klassieke testtheorie impliceert namelijk dat de gemiddelde toetsscores gelijk zijn wat betekent dat $\hat{\sigma}_v^2/n_v$ per definitie gelijk is aan nul. Dit sluit aan op de praktijk dat met de klassieke testtheorie doorgaans alleen relatieve beslissingen maar geen absolute beslissingen over personen genomen worden.

3.13 Design met twee facetten

Hiervoor hebben we de verschillende fasen van de analyse van een-facet-design besproken. Aangezien de analyse van een twee-facet-design op vergelijkbare wijze verloopt, kan de bespreking van de diverse fasen relatief kort zijn. Een voorbeeld van een gekruist twee-facet- design is een design waarbij de antwoorden op vragen van personen beoordeeld worden door beoordelaars. In een gekruist twee-facet-design wordt de geobserveerde score van een persoon *p* op een item *v*, toegekend door een beoordelaar *b*, X_{pvb}, uitgedrukt als een decompositie van de score in zeven componenten:

$$\begin{aligned}
X_{pvb} &= \mu && \text{(algemene gemiddelde)}\\
&+ \mu_p - \mu && \text{(persoonseffect)}\\
&+ \mu_v - \mu && \text{(itemeffect)}\\
&+ \mu_b - \mu && \text{(beoordelaarseffect)}\\
&+ \mu_{pv} - \mu_p - \mu_v + \mu && \text{(persoons- × itemeffect)}\\
&+ \mu_{pb} - \mu_p - \mu_b + \mu && \text{(persoons- × beoordelaarseffect)}\\
&+ \mu_{vb} - \mu_v - \mu_b + \mu && \text{(item- × beoordelaarseffect)}\\
&+ X_{pvb} - \mu_{pv} - \mu_{pb} - \mu_{vb} + \mu_p + \mu_v + \mu_b - \mu. && \text{(residu)}
\end{aligned} \qquad (3.33)$$

In (3.33) is het algemene gemiddelde gedefinieerd als $\mu = \mathcal{E}_p \mathcal{E}_v \mathcal{E}_b X_{pvb}$, de gemiddelde score (= verwachting over personen, vragen en beoordelaars) na beoordeling van alle antwoorden van alle personen uit de populatie op alle vragen uit het universum door alle beoordelaars uit het universum van beoordelaars. De universumscore van een persoon is gedefinieerd als $\mu_p = \mathcal{E}_v \mathcal{E}_b X_{pvb}$, de gemiddelde score (= verwachting over items en beoordelaars) van een persoon na beoordeling van de antwoorden op alle vragen uit het universum door alle beoordelaars uit het universum. De strengheid van een beoordelaar is gedefinieerd als $\mu_b = \mathcal{E}_p \mathcal{E}_v X_{pvb}$, de gemiddelde score (= verwachting over personen en items) van een beoordelaar na beoordeling van de antwoorden op alle vragen uit het universum door alle personen uit de populatie. De parameter μ_{pv} is gedefinieerd als $\mu_{pv} = \mathcal{E}_b X_{pvb}$, de gemiddelde score (= verwachting over beoordelaars) van een persoon op een vraag na beoordeling van het antwoord door alle beoordelaars uit het universum. De definities van de parameters μ_v, μ_{pb} en μ_{vb} zijn respectievelijk $\mu_v = \mathcal{E}_p \mathcal{E}_b X_{pvb}$, $\mu_{pb} = \mathcal{E}_v X_{pvb}$ en $\mu_{vb} = \mathcal{E}_p X_{pvb}$. De definities van de variantiecomponenten voor personen, vragen en beoordelaars zijn respectievelijk $\sigma_p^2 = \mathcal{E}_p(\mu_p - \mu)^2$, $\sigma_b^2 = \mathcal{E}_b(\mu_b - \mu)^2$ en $\sigma_v^2 = \mathcal{E}_v(\mu_v - \mu)^2$. Voor wat betreft de overige variantiecomponenten volstaan we met het geven van de definitie voor het persoons-× itemeffect: $\sigma_{pv}^2 = \mathcal{E}_p \mathcal{E}_v(\mu_{pv} - \mu_p - \mu_v + \mu)^2$.

De totale variantie is gelijk aan:

$$\sigma_X^2 = \sigma_p^2 + \sigma_v^2 + \sigma_b^2 + \sigma_{pv}^2 + \sigma_{pb}^2 + \sigma_{vb}^2 + \sigma_{pvb,e}^2. \qquad (3.34)$$

In het twee-facet-design met slechts een observatie voor elke combinatie van persoon, vraag en beoordelaar, bestaat de variantiecomponent voor het residu, $\sigma_{pvb,e}^2$, uit de niet te scheiden variantiecomponenten voor de interactie personen × vragen × beoordelaars en voor de fouten. Daarnaast worden er in (3.34) nog vijf andere variantiecomponenten voor mogelijke foutenbronnen onderscheiden: de twee variantiecomponenten voor de twee hoofdeffecten en de drie variantiecomponenten voor de drie eerste-orde-interactie-effecten.

De mogelijkheid om door toepassing van designs met meer facetten verschillende foutenbronnen te onderscheiden, is het belangrijkste verschil tussen de generaliseerbaarheids-theorie en de klassieke testtheorie. In voorgaande paragrafen zagen we dat in de klassieke testtheorie geen onderscheid gemaakt wordt tussen de verschillende storende factoren die de toetsscore van een persoon beïnvloeden en dat alle foutenbronnen door een enkele variantie-component gerepresenteerd worden.

3.13.1 Generaliseerbaarheidsstudie

De tabellen 3.19 en 3.20 bevatten alle informatie die nodig is om een generaliseerbaarheidsstudie uit te voeren. Tabel 3.19 geeft de variantie-analysetabel van een gekruist twee-facet-design met drie random effecten. In tabel 3.20 staat hoe men de kwadratensommen kan berekenen en hoe de zeven variantiecomponenten geschat kunnen worden.

Aan de hand van het voorbeeld, ontleend aan Thorndike (1982, p. 161), in tabel 3.21 laten we zien hoe de berekening van de kwadratensommen verloopt. Daartoe dienen we de zeven parameters in (3.33) te vervangen door hun geobserveerde equivalenten. Dit resulteert in de volgende decompositie:

$$X_{pvb} = \bar{X} + (\bar{X}_p - \bar{X}) + (\bar{X}_v - \bar{X}) + (\bar{X}_b - \bar{X}) + \bar{X}_{pv\sim} + \bar{X}_{pb\sim} + \bar{X}_{vb\sim} + X_{pvb\sim} \qquad (3.35)$$

In (3.35) staat $\bar{X}_{pv\sim}$ als afkorting voor $\bar{X}_{pv} - \bar{X}_p - \bar{X}_v + X$. De betekenis van afkortingen voor de andere interactietermen staat in tabel 3.20.

Tabel 3.19

Variantie-analysetabel van een gekruist design met drie random effecten en schattingen van variantiecomponenten

Effecten	Kwadraten-sommen	Vrijheidsgraden	Gemiddelde kwadratensommen	
Personen (*p*)	SS_p	$df_p = n_p - 1$	$MS_p = SS_p/df_p$	$\hat{\sigma}^2_p = (MS_p$
Items (*v*)	SS_v	$df_v = n_v - 1$	$MS_v = SS_v/df_v$	$\hat{\sigma}^2_v = (MS_v$
Beoordelaars (*b*)	SS_b	$df_b = n_b - 1$	$MS_b = SS_b/df_b$	$\hat{\sigma}^2_b = (MS_b$
Personen x items (*pv*)	SS_{pv}	$df_{pv} = (n_p - 1)(n_v - 1)$	$MS_{pv} = SS_{pv}/df_{pv}$	$\hat{\sigma}^2_{pv} = (MS_{pv}$
Personen x beoordelaars (*pb*)	SS_{pb}	$df_{pb} = (n_p - 1)(n_b - 1)$	$MS_{pb} = SS_{pb}/df_{pb}$	$\hat{\sigma}^2_{pb} = (MS_{pb}$
Items x beoordelaars (*vb*)	SS_{vb}	$df_{vb} = (n_v - 1)(n_b - 1)$	$MS_{vb} = SS_{vb}/df_{vb}$	$\hat{\sigma}^2_{vb} = (MS_{vb}$
Residu (*pvb,e*)	$SS_{pvb,e}$	$df_{pvb,e} = (n_p - 1)(n_v - 1)(n_b - 1)$	$MS_{pvb,e} = SS_{pvb,e}/df_{pvb,e}$	$\hat{\sigma}^2_{pvb,e} = MS_{pvb,e}$

Tabel 3.20

Definities van kwadratensommen en schattingen van variantiecomponenten

SS_p	$= n_v n_b \sum_p (\bar{X}_p - \bar{X})^2$	$= MS_{pvb,e}$
SS_v	$= n_p n_b \sum_v (\bar{X}_v - \bar{X})^2$	$(MS_{vb} - MS_{pvb,e})/n_p$
SS_b	$= n_p n_v \sum_b (\bar{X}_b - \bar{X})^2$	$(MS_{pb} - MS_{pvb,e})/n_v$

$$SS_{pv} = n_b \sum_p \sum_v (\bar{X}_{pv\sim})^2 = n_b \sum_p \sum_v (\bar{X}_{pv} - \bar{X}_p - \bar{X}_v + \bar{X})^2 \qquad \hat{\sigma}^2_{pv} = (MS_{pv} - MS_{pvb,e})/n_b$$

$$SS_{pb} = n_v \sum_p \sum_b (\bar{X}_{pb\sim})^2 = n_v \sum_p \sum_b (\bar{X}_{pb} - \bar{X}_p - \bar{X}_b + \bar{X})^2 \qquad \hat{\sigma}^2_b = (MS_b - MS_{vb} - MS_{pb} + MS_{pvb,e})/(n_p n_v)$$

$$SS_{vb} = n_p \sum_v \sum_b (\bar{X}_{vb\sim})^2 = n_p \sum_v \sum_b (\bar{X}_{vb} - \bar{X}_v - \bar{X}_b + \bar{X})^2 \qquad \hat{\sigma}^2_v = (MS_v - MS_{vb} - MS_{pv} + MS_{pvb,e})/(n_p n_b)$$

$$SS_{pvb,e} = \sum_p \sum_v \sum_b (X_{pvb\sim})^2 = \sum_p \sum_v \sum_b (X_{pvb} - \bar{X}_{pv} - \bar{X}_{pb} - \bar{X}_{vb} + \bar{X}_p + \bar{X}_v + \bar{X}_b - \bar{X})^2 \qquad \hat{\sigma}^2_p = (MS_p - MS_{pb} - MS_{pv} + MS_{pvb,e})/(n_v n_b)$$

$$SS_{tot} = \sum_p \sum_v \sum_b (X_{pvb} - \bar{X})^2$$

Tabel 3.21

De itemscores van zes personen op vier items en twee beoordelaars, per beoordelaar de gemiddelde score per item en per persoon, de gemiddelde score per beoordelaar, de gemiddelde score van elke persoon en het algemene gemiddelde

	Beoordelaar 1					Beoordelaar 2					
Pers.	Item: 1	2	3	4	Gem.	Item: 1	2	3	4	Gem.	$\bar{X}_p$
1	9	6	6	2	5.75	8	2	8	1	4.75	5.25
2	9	5	4	0	4.50	7	5	9	5	6.50	5.50
3	8	9	5	8	7.50	10	6	9	10	8.75	8.13
4	7	6	5	4	5.40	9	8	9	4	7.70	6.50
5	7	3	2	3	3.75	7	4	5	1	4.25	4.00
6	10	8	7	7	8.00	7	7	10	9	8.25	8.13
Gem.	8.33	6.17	4.83	4.00	5.83	8.00	5.33	8.33	5.00	6.67	$\bar{X}$ = 6.25

Tabel 3.21 bevat de itemscores die twee beoordelaars aan de antwoorden op vier items aan zes personen toegekend hebben. Voor persoon 1 uit deze tabel hebben we (3.35) uitgeschreven in tabel 3.22.

Tabel 3.22

Vergelijking (3.35) uitgeschreven voor persoon 1 uit tabel 3.21

X_{pvb} =	$\bar{X}$	+ $(\bar{X}_p - \bar{X})$	+ $(\bar{X}_v - \bar{X})$	+ $(\bar{X}_b - \bar{X})$	+ $\bar{X}_{pv}$~	+ $\bar{X}_{pb}$~	+ $\bar{X}_{vb}$~	+ X_{pvb}~
X_{111} = 9 =	6.25	— 1.00	+ 1.92	— 0.42	+ 1.33	+ 0.92	+ 0.58	— 0.58
X_{112} = 8 =	6.25	— 1.00	+ 1.92	+ 0.42	+ 1.33	— 0.92	— 0.58	+ 0.58
X_{121} = 6 =	6.25	— 1.00	— 0.50	— 0.42	— 0.75	+ 0.92	+ 0.83	+ 0.67
X_{122} = 2 =	6.25	— 1.00	— 0.50	+ 0.42	— 0.75	— 0.92	— 0.83	— 0.67
X_{131} = 6 =	6.25	— 1.00	+ 0.33	— 0.42	+ 1.42	+ 0.92	— 1.33	— 0.17
X_{132} = 8 =	6.25	— 1.00	+ 0.33	+ 0.42	+ 1.42	— 0.92	+ 1.33	+ 0.17
X_{141} = 2 =	6.25	— 1.00	— 1.75	— 0.42	— 2.00	+ 0.92	— 0.08	+ 0.08
X_{142} = 1 =	6.25	— 1.00	— 1.75	+ 0.42	— 2.00	— 0.92	+ 0.08	— 0.08

Voor het berekenen van de kwadratensommen moeten we vergelijking (3.35) ook nog uitschrijven voor de vijf andere personen, wat een uitbreiding betekent van tabel 3.22 met de decomposities van veertig itemscores. De zeven kwadratensommen worden verkregen door de getallen in de desbetreffende kolommen van tabel 3.22 te kwadrateren en te sommeren. Beschikken we over de kwadratensommen, dan kunnen we schattingen van de variantie-componenten eenvoudig berekenen met behulp van tabel 3.20. Wellicht ten overvloede merken we op dat de standaardfouten van variantiecomponenten bij kleine aantallen personen en condities zeer groot zijn (Brennan, 1992, p. 104). De steekproef uit de populatie moet uit minstens honderd personen bestaan teneinde acceptabele standaardfouten te verkrijgen (Smith, 1978). De resultaten van de generaliseerbaarheidsstudie voor het voorbeeld staan vermeld in tabel 3.23.

Tabel 3.23

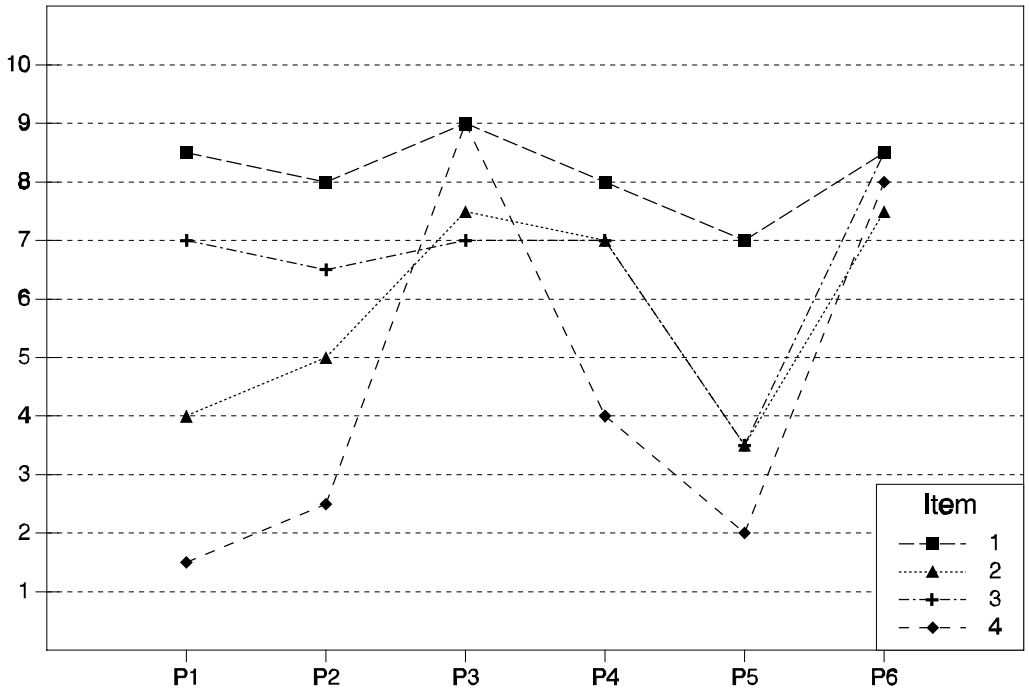
Resultaten generaliseerbaarheidsstudie voor data uit tabel 3.21

Effecten	Kwadraten-sommen	Vrijheids-graden	Gemiddelde kwadraten-sommen	Schattingen van variantie-componenten	
Personen *(p)*	109.75	5	21.95	$\hat{\sigma}^2_p$	= 2.16 (28%)
Items *(v)*	85.17	3	28.39	$\hat{\sigma}^2_v$	= 1.26 (15%)
Beoordelaars *(b)*	8.33	1	8.33	$\hat{\sigma}^2_b$	= –0.15 (0%)

Personen × items *(pv)*	59.08	15	3.94	$\hat{\sigma}^2_{pv}$ = 0.98 (12%)
Personen × beoordelaars *(pb)*	13.42	5	2.68	$\hat{\sigma}^2_{pb}$ = 0.18 (2%)
Items × beoordelaars *(vb)*	33.83	3	11.28	$\hat{\sigma}^2_{vb}$ = 1.55 (19%)
Residu *(pvb,e)*	29.42	15	1.96	$\hat{\sigma}^2_{pvb,e}$ = 1.96 (24%)

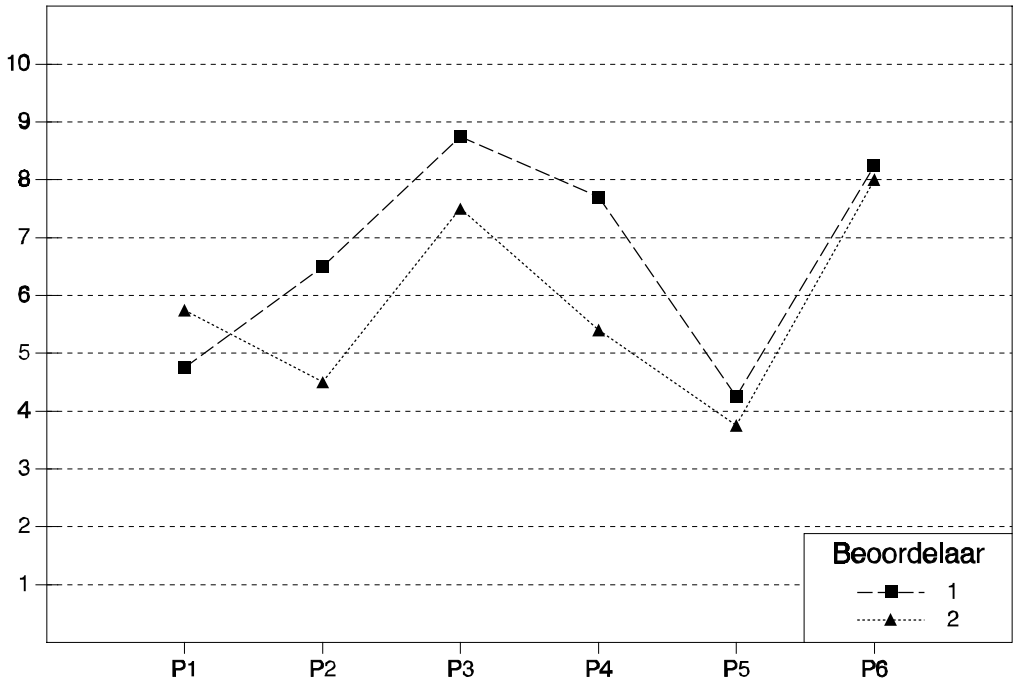
De laatste kolom van tabel 3.23 bevat de schattingen van de variantiecomponenten en hun procentuele bijdrage aan de totale variantie. We zien dat de variantiecomponent van de beoordelaars negatief is. Hoewel in theorie variantiecomponenten niet negatief kunnen zijn, kunnen schattingen van variantiecomponenten wel negatief zijn. Negatieve schattingen hebben veelal twee mogelijke oorzaken. Relatief grote negatieve componenten zijn meestal het gevolg van het gebruik van het verkeerde model. Een relatief grote negatieve component van beoordelaars had er in ons voorbeeld op kunnen wijzen dat het lineaire model in (3.33) niet het juiste model was om de data te analyseren. Relatief kleine negatieve componenten zijn meestal het gevolg van het gebruik van een te kleine steekproef. Dit laatste is waarschijnlijk de oorzaak van de negatieve component in ons voorbeeld. Aangezien negatieve componenten niet mogelijk zijn, worden negatieve schattingen vervangen door nul. Merk op dat er andere schattingsmethoden voor variantiecomponenten zijn die niet leiden tot negatieve schattingen. Een daarvan is de restrictieve grootste-aannemelijkheidschattingsmethode. De relatief grote bijdrage van de variantiecomponent voor de items is met name het gevolg van het grote verschil in moeilijkheidsgraad tussen item 1 en item 4. De gemiddelde itemscore van item 1 is 8.17, terwijl die van item 4 gelijk is aan 4.50.

De bijdrage van de interactiecomponent personen × items is veel groter dan die van de interactiecomponent personen × beoordelaars. Interactie tussen personen en items betekent dat personen niet consistent antwoorden op de verschillende items. Interactie tussen personen en beoordelaars houdt in dat personen niet consistent beoordeeld worden door verschillende beoordelaars. In figuur 3.3. hebben we de interactie personen × items grafisch gepresenteerd.



Figuur 3.3 Interactie personen × items

In figuur 3.3 is voor elk item een lijn getrokken die de gemiddelde itemscores, $\bar{X}_{pv}$, van personen, P1-P6, met elkaar verbindt. We zien dat de vier lijnen elkaar bij verschillende personen kruisen, wat betekent dat het niet dezelfde persoon is die de hoogste of laagste score op elk item behaalt. Lijnen die elkaar kruisen wijzen er op dat er sprake is van interactie. Merk op dat in tabel 3.22 de berekening van de variantiecomponent voor de interactie tussen personen en items gebaseerd is op $\bar{X}_{pv}\sim = \bar{X}_{pv} - \bar{X}_{p} - \bar{X}_{v} + \bar{X}$. We hadden de interactie tussen personen en items ook met behulp van $X_{pv}\sim$ in plaats van $\bar{X}_{pv}$ kunnen afbeelden. Wanneer de vier lijnen parallel lopen is, de kwadratensom personen × items, en dus ook de variantiecomponent, gelijk aan nul.



Figuur 3.4 Interactie personen × beoordelaars

Om mogelijke interactie tussen personen en beoordelaars te onderzoeken, is in figuur 3.4 voor elk item een lijn getrokken die de gemiddelde beoordelaarsscores, $\bar{X}_{pb}$, van personen met elkaar verbindt. We zien dat de twee lijnen elkaar bij de eerste persoon kruisen maar bij de andere vijf personen nagenoeg parallel lopen. Dit betekent dat de twee beoordelaars de eerste persoon niet, maar de vijf andere personen wel op dezelfde wijze onderscheiden. De variantiecomponent voor de interactie tussen personen en beoordelaars blijkt dan ook gering te zijn.

De interactie items × beoordelaars is de grootste eerste-orde-interactie, met name veroorzaakt door de derde vraag. Die vraag heeft van de eerste beoordelaar een lage beoordeling, gemiddelde score 4.83, en van de tweede beoordelaar een hoge beoordeling, gemiddelde score 8.33, ontvangen.

3.13.2 Decisiestudie

In ons voorbeeld bestaat de toets uit vier random getrokken items uit het universum van items en twee random getrokken beoordelaars uit het universum van beoordelaars die de antwoorden op de items beoordelen. Een andere toets met vier random getrokken items en twee random getrokken beoordelaars zou ook acceptabel geweest

zijn. Het universum waar in dit geval naar gegeneraliseerd wordt, is het universum van random parallelle toetsen met vier items en twee beoordelaars.

De schatting van de generaliseerbaarheidscoëfficiënt voor relatieve beslissingen is voor het gekruiste twee-facet-random-effect-design gedefinieerd als:

$$\hat{\rho}^2 = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + \frac{\hat{\sigma}^2_{pv}}{n_v} + \frac{\hat{\sigma}^2_{pb}}{n_b} + \frac{\hat{\sigma}^2_{pvb,e}}{n_v n_b}}. \qquad (3.36)$$

Naast de universumscorevariantie, bevat de noemer van (3.36) drie variantiecomponenten die interacties met personen betreffen. Hiervoor zagen we dat een relatief grote variantie- component voor de interactie tussen personen en items inhoudt dat bijvoorbeeld Jan niet op ieder item meer presteert dan Piet. Het maakt voor het nemen van relatieve beslissingen dan ook wel degelijk uit welke items aan welke personen voorgelegd worden. Een bepaald item wordt namelijk door Jan als gemakkelijk en door Piet als moeilijk opgevat, terwijl bij een ander item het omgekeerde het geval is. De variantiecomponent voor de interactie tussen personen en items dient dan ook beschouwd te worden als foutenvariantie. Ook de variantiecomponent voor de interactie tussen personen en beoordelaars, dat wil zeggen dat het van de beoordelaar afhangt of Jan beter is dan Piet, dient als foutenvariantie beschouwd te worden. De variantiecomponent voor het residu is per definitie foutenvariantie. Voor de toets uit ons voorbeeld is de generaliseerbaarheidscoëfficiënt gelijk aan: 2.16/{2.16 + 0.99/4 + 0.18/2 + 1.96/8} = .79.

De schatting van de generaliseerbaarheidscoëfficiënt voor absolute beslissingen is voor het gekruiste twee-facet-random-effect design gedefinieerd als:

$$\hat{\rho}^2 = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + \frac{\hat{\sigma}^2_v}{n_v} + \frac{\hat{\sigma}^2_b}{n_b} + \frac{\hat{\sigma}^2_{pv}}{n_v} + \frac{\hat{\sigma}^2_{pb}}{n_b} + \frac{\hat{\sigma}^2_{pvb,e}}{n_v n_b}}. \qquad (3.37)$$

Bij het nemen van absolute beslissingen maakt het niet alleen uit of er makkelijke of moeilijke vragen aan de personen voorgelegd worden, maar ook of die vragen door milde of strenge beoordelaars beoordeeld worden. Vandaar dat in (3.37) naast de variantiecomponenten voor de drie interacties ook de variantiecomponenten voor de items en voor de beoordelaars beschouwd worden als foutenvariantie. De generaliseerbaarheidscoëfficiënt voor absolute beslissingen is gelijk aan 2.16/{2.16 + 1.26/4 + 0.0/2 + 0.99/4 + 0.18/2 + 1.96/8} = .71 voor de toets uit ons voorbeeld.

3.14 Andere aspecten van de generaliseerbaarheidstheorie

Formule (3.36) laat zien dat we de generaliseerbaarheidscoëfficiënt kunnen verhogen door de toets te verlengen, wat neerkomt op het vergroten van het aantal items of het aantal beoordelaars. Voor het realiseren van dezelfde generaliseerbaarheidscoëfficiënt hebben we meer condities nodig van een facet met een relatief grote variantiecomponent die bijdraagt aan de foutenvariantie, dan condities van een facet met een relatief kleine variantiecomponent. We verwijzen naar hoofdstuk 11 voor een bespreking van toetsverlenging bij designs met meer facetten.

De generaliseerbaarheidscoëfficiënt kan ook verhoogd worden door een random facet op te vatten als een fixed facet. Dat een facet fixed is, wil zeggen dat een toets alle condities van een facet bevat. Beschouwen we in ons voorbeeld de items als fixed facet, dan generaliseren we niet meer naar het universum van random parallelle toetsen met vier items en twee beoordelaars, maar naar het universum van random parallelle toetsen met twee beoordelaars. Het spreekt vanzelf dat door het beperken van het universum waar naar gegeneraliseerd wordt, de beslissingen over personen nauwkeuriger kunnen zijn. Voor een bespreking van designs met fixed facets verwijzen we naar Shavelson en Webb (1991, pp. 65-82).

De bespreking in voorgaande paragrafen heeft zich beperkt tot gekruiste designs met een enkel facet en met twee facetten. Binnen de generaliseerbaarheidstheorie kunnen echter ook designs met meer dan twee facetten geanalyseerd worden. Daarnaast kunnen ook zogenaamde genestelde designs geanalyseerd worden. Ons voorbeeld met twee facetten zou een genesteld design zijn wanneer de eerste en de tweede vraag door de eerste beoordelaar beoordeeld worden en de derde en vierde vraag door de tweede beoordelaar. In dat geval zeggen we dat de vragen genesteld zijn binnen de beoordelaars. Genestelde designs komen vooral voor bij niet-experimenteel onderzoek (Feldt & Brennan, 1989). In het algemeen heeft het gebruik van gekruiste designs de voorkeur, omdat het met de resultaten van de generaliseerbaarheidsstudie van gekruiste designs mogelijk is na te gaan hoe de resultaten voor een genesteld design geweest zouden zijn. Het omgekeerde is niet het geval.

In de voorbeelden die tot nu toe besproken zijn, hadden de beslissingen steeds betrekking op personen. In veel onderzoek, met name onderzoek op het gebied van het onderwijs, zijn we echter niet of niet uitsluitend geïnteresseerd in (verschillen tussen) personen maar ook in klassen, leerdoelen of andere meetobjecten. Om aan te geven dat elk facet uit een design het meetobject kan zijn, introduceerden Cardinet, Tourneur en Allal (1981) het zogenaamde symmetrieprincipe. Uitgaande van dat principe laten zij

zien hoe binnen het kader van de generaliseerbaarheidstheorie een grote verscheidenheid aan onderzoeksvragen beantwoord kan worden.

De meest gebruikte schatting van de universumscore van een persoon is de geobserveerde gemiddelde score van een persoon. In Cronbach e.a. (1972) worden echter ook varianten van Kelley's formule (zie paragraaf 3.5) voor schattingen van universumscores besproken. Hoe schattingen van universumscores verkregen kunnen worden met behulp van lineaire predictiefuncties wordt beschreven door Jarjoura (1983).

Tenslotte dient opgemerkt te worden dat met de generaliseerbaarheidstheorie niet alleen univariate maar ook multivariate modellen, dat wil zeggen modellen waarbij de personen een aantal universumscores hebben, geanalyseerd kunnen worden. Voor een bespreking van modellen uit de multivariate generaliseerbaarheidstheorie verwijzen we naar Cronbach e.a. (1972), Shavelson en Webb (1981) en Brennan (1992).

4

Itemresponstheorie

Het belangrijkste concept in de klassieke testtheorie is de betrouwbaarheid: daarmee wordt aangegeven in welke mate geobserveerde verschillen in toetsscores werkelijke verschillen tussen personen weerspiegelen. De definitie van de betrouwbaarheid steunt op de opsplits- baarheid van de variantie van de toetsscores X (zie hoofdstuk 3):

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2, \tag{4.1}$$

of de variantie van de toetsscore, de totale variantie, is de som van de variantie van de ware scores plus de variantie van de meetfout. De betrouwbaarheid is dan per definitie de ver- houding tussen de variantie van de ware score en de totale variantie:

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} = \rho_{XX'}. \tag{4.2}$$

Het rechterlid van (4.2) geeft aan hoe die betrouwbaarheid kan worden vastgesteld, namelijk als de correlatie tussen X en een parallelvorm X'. Indien we (4.2) wat nader onderzoeken dan duiken er twee problemen op waarvoor niet zo snel een oplossing gevonden is.

Het eerste probleem betreft het gebruik van spreidingsmaten, zoals de variantie, die altijd naar een verdeling of een populatie verwijzen. Hoewel dit in (4.2) niet uitdrukkelijk gezegd wordt, is de referentie naar een of andere populatie impliciet aanwezig, en dit impliceert weer dat de betrouwbaarheid van een toets een eigenschap is die niet alleen de toets karakteriseert, maar de toets in de populatie. Het niet expliciteren van die betrekkelijkheid, wat in de praktijk nogal eens voorkomt, dekt het

probleem misschien toe, maar lost het zeker niet op. Een mededeling zoals "de betrouwbaarheid van X is 0.8" is dus zinloos als men er zich niet van verzekert dat spreker en aangesprokene dezelfde populatie in gedachten hebben.

Het tweede probleem is dat de ware score T toetsspecifiek is: de intuïtieve betekenis van de ware score is de gemiddelde score die een persoon behaalt wanneer de toets X een zeer groot aantal keer onder dezelfde omstandigheden wordt afgenomen. Het is daarbij irrelevant of dit al dan niet praktisch realiseerbaar is. De belangrijke vraag is echter of het kennen of schatten van deze ware toetsscore op zichzelf een belangrijke aangelegenheid is. In theoretisch onderzoek en in toepassingen zal men toch eerder tot het standpunt neigen dat een toetsscore iets dient te onthullen over een meer abstracte entiteit, een vaardigheid, een geschiktheid of een attitude, waarbij de items die men in de toets gebruikt in principe zouden kunnen worden vervangen door andere items. De belangrijke vraag is dus of de ware toetsscore, die samenhangt met een specifieke toets, iets kan zeggen over een meer abstracte, onderliggende vaardigheid. Dit resulteert in een aantal vragen waarop de klassieke testtheorie geen afdoend antwoord kan bieden.

Een toets bestaat uit een aantal onderdelen of items. Hoe kan een toetsconstructeur weten of het zinvol is bepaalde items samen in dezelfde toets op te nemen? Immers, als de toetsscore een indicator is van de mate waarin een theoretisch concept aanwezig is of beheerst wordt, dient elk item dat in de toets wordt opgenomen relevant te zijn voor dit concept, dat wil zeggen de toets moet homogeen zijn met betrekking tot dit concept. Nu is het natuurlijk niet zo dat professioneel gemaakte toetsen een willekeurig allegaartje van items zijn. De toetsconstructeur gebruikt wel degelijk theoretische kennis om tot een verantwoorde keuze van items te komen. Het belangrijke punt is echter dat de klassieke testtheorie, als statistische theorie, geen middelen aanbiedt aan de hand waarvan duidelijk kan beslist worden of deze homogeniteit in conceptuele relevantie al dan niet bereikt is. Het beste wat de klassieke theorie kan bieden is een index van interne consistentie, de KR-20 bijvoorbeeld, maar zulke indices hebben een dubbelzinnige betekenis. Indien ze hoog zijn, waarbij de vraag wat hoog is een nieuw probleem oproept, dan wijst dit op homogeniteit en grote betrouwbaarheid. Echter, indien de KR-20 laag is, wijst dit op een gebrek aan homogeniteit of betrouwbaarheid of beide, en uit de waarde van de KR-20 valt niet af te leiden wat er nu precies het geval is.

De tweede vraag betreft de scoringsregel. In de klassieke testtheorie wordt de toetsscore bij dichotome items meestal gedefinieerd als het aantal items juist, ook wel aangeduid als ruwe somscore. Hoewel deze definitie voor de hand liggend kan lijken, is ze in principe willekeurig. Er zijn andere scoreregels denkbaar die in bepaalde omstandigheden veel zinvoller kunnen zijn. De klassieke benadering bevat echter geen

theorie waaruit de superioriteit van de gewone somscoreregel of welke regel dan ook volgt.

De derde vraag, die binnen de klassieke testtheorie in principe onoplosbaar is, is de volgende. Een steekproef van kinderen, aselect getrokken uit een goed gedefinieerde populatie, wordt op tijdstip t_1 gemeten met een toets X_1 en op tijdstip t_2 met een toets X_2, waarbij het de bedoeling is te schatten of de gemiddelde vaardigheid in de populatie veranderd is in het interval $(t_1 - t_2)$. Indien X_1 niet identiek is aan X_2 treedt er een dubbel probleem op. Indien het gemiddelde op X_2 groter is dan het gemiddelde op X_1 zou het verschil te wijten kunnen zijn aan het feit dat X_2 gemakkelijker is dan X_1, of aan het feit dat de gemiddelde vaardigheid inderdaad is toegenomen, of aan beide. Om de verklaring van een gemakkelijker toets uit te sluiten dienen dus speciale maatregelen genomen te worden, bijvoorbeeld het afnemen van toets X_2 op tijdstip t_1 bij een onafhankelijke steekproef uit dezelfde populatie, zodanig dat X_1 en X_2 kunnen geëquivaleerd worden (zie hoofdstuk 8). Equivaleren is echter een puur technische ingreep, en is zeker geen oplossing voor het tweede, veel fundamenteler probleem: hoe kan gegarandeerd worden dat X_1 en X_2 inderdaad hetzelfde concept meten. Indien men op dit probleem geen afdoende antwoord kan geven staat men weerloos tegen de aantijging dat bovengenoemde vergelijking het vergelijken is van appels met peren, en dus zinloos.

In de moderne testtheorie wordt aan de eerdergenoemde twee problemen van de klassieke testtheorie, te weten de populatie-afhankelijkheid en de toetsspecificiteit van de score, tegemoet gekomen. De theorie wordt ontwikkeld zonder enige referentie aan een of andere populatie, hoewel we verderop zullen zien dat in sommige omstandigheden dit populatiebegrip weer zal opduiken. Bovendien staat in die theorie niet de toetsscore centraal, maar het item en het antwoord op het item. Dit verklaart meteen ook de naam van deze theorie: itemresponstheorie (IRT). Hiervoor hebben we gezegd dat de ware score T van een persoon in principe observeerbaar is door de scores van een groot aantal toetsafnames te middelen. De IRT hanteert een begrip dat men losjes zou kunnen omschrijven als de te meten vaardigheid, dat in principe niet observeerbaar is. Om deze principiële onobserveerbaarheid aan te duiden gebruikt men de term latent, en het begrip vaardigheid wordt soms vervangen door de meer neutrale term trek. Een equivalente doch verouderde benaming voor IRT is dan ook latente-trektheorie (in het Engels: latent trait theory).

Een IRT is een geheel van uitspraken over de samenhang tussen de latente trek en het antwoordgedrag op een verzameling items. De conceptuele homogeniteit waarover hierboven werd gesproken is niets anders dan deze samenhang. In de mate dat deze samenhang duidelijk gedefinieerd is, weten we ook wat precies met homogeniteit wordt

bedoeld. In paragraaf 4.1 wordt een algemene inleiding van deze theorie gegeven aan de hand van één speciaal geval, het Raschmodel.

De uitspraken in zo'n theorie zijn meestal niet heel specifiek: de voorspellingen over het gedrag hangen af van kenmerken van de items en van de personen. Deze kenmerken worden meestal gekwantificeerd als kengetallen of parameters, en de waarden van deze parameters zijn in de regel niet bekend. Een belangrijk probleem in de IRT is dan ook het schatten van deze parameters en het geven van een aanduiding van de nauwkeurigheid waarmee deze parameters kunnen worden geschat. De schattingsproblematiek wordt behandeld in paragraaf 4.2.

Een theorie is alleen die naam waardig indien ze gefalsificeerd kan worden. In paragraaf 4.3 worden methoden besproken waarmee kan worden nagegaan of de predicties over het gedrag die uit de theorie volgen wel met de werkelijkheid overeenkomen. Deze methoden steunen sterk op de statistische theorie, en nemen meestal de vorm aan van formele statistische toetsen waarbij het gehanteerde model de status van nulhypothese krijgt.

Paragraaf 4.4 bevat een technische uiteenzetting van de werkwijze bij parameterschattingen en modeltoetsen indien de data verzameld zijn in een onvolledig design.

Men kan zich natuurlijk gaan afvragen waar de meetprocedure zelf blijft. De bedoeling van het meten is het toekennen van een getal aan een persoon op zodanige manier dat de grootte van het getal ook de mate van zijn vaardigheid uitdrukt. Het is kenmerkend voor de literatuur in IRT dat de eerste en meeste aandacht gaat naar het zorgvuldig opbouwen en toetsen van de theorie, en dat de meetprocedures zelf veel minder aandacht krijgen. Niettemin is de meetprocedure zelf belangrijk en een aantal subtiele problemen in verband hiermee verdienen meer aandacht dan ze doorgaans in de literatuur krijgen. Dit is het onderwerp van paragraaf 4.5.

4.1 Begrippen en algemene theorie

Centraal in de IRT staat het begrip latente variabele. Hoewel er verschillende opvattingen zijn over de status van deze variabele, zullen we ons hier beperken tot één geval, namelijk waar het domein van de latente variabele de reële as is. Elke persoon in een populatie kan afgebeeld worden als een punt van de reële as, of wat equivalent hiermee is, aan elke persoon kan een getal worden toegevoegd dat een uitdrukking is van de mate waarin die persoon over de vaardigheid beschikt. Aan die latente variabele geen inhoud toegeschreven, het is dus een abstracte variabele, die we verder dan ook

met het algemeen symbool θ zullen aanduiden. De getalswaarde die aan persoon *v* is toegekend duiden we aan als θ_v.

Merk op dat de waarde van θ niet begrensd is: $-\infty < \theta < \infty$. Om iets te kunnen zeggen over de θ-waarde van een persoon veronderstelt men dat de antwoorden op bepaalde items enige indicatie geven over de vaardigheid. Bijvoorbeeld door een uitspraak als: "een correct antwoord op dit item duidt op een grotere vaardigheid dan een fout antwoord". Met zo'n vage uitspraak kan natuurlijk niet veel gedaan worden. In de IRT staat het expliciet maken van het verband tussen de latente variabele θ en de itemantwoorden dan ook centraal.

Eerst een definitie. Met X_i duiden we het antwoord aan op item *i*, en voorlopig gaan we ervan uit dat X_i dichotoom is, met waarden toegekend volgens onderstaande regel:

$$X_i = \begin{cases} 1 \text{ indien het antwoord op item } i \text{ correct is,} \\ \\ 0 \text{ indien het antwoord op item } i \text{ fout is.} \end{cases}$$

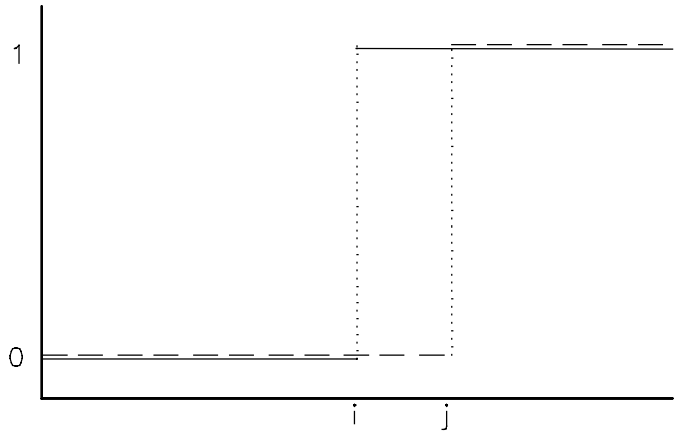
Centraal in de IRT is de aanname dat het antwoord op een item nooit volledig vastligt, hoe groot of hoe klein de vaardigheid van de persoon die het item beantwoordt ook is. Daarom wordt met kansen gewerkt, en de variabele X_i is een toevalsvariabele. De itemresponsfunctie drukt uit hoe groot de kans is dat het item juist wordt beantwoord als functie van de vaardigheid. Deze functie wordt aangeduid met het symbool $f_i(\theta)$. Dus,

$$f_i(\theta) = P(X_i=1|\theta) \qquad (4.3)$$

of, de itemresponsfunctie is de conditionele kans op een juist antwoord gegeven de waarde van θ. Formule (4.3) is nog geen theorie; zij is eigenlijk niets meer dan een conventie over de notatie. We schrijven kortheidshalve het linkerlid op, als we het rechterlid bedoelen. Om een echte theorie te maken zullen we de functie moeten specificeren, dat wil zeggen we moeten het verloop ervan beschrijven en er de eigenschappen van vastleggen. Omdat we later mathematische manipulaties met die functie zullen moeten uitvoeren, zullen we eisen dat ze niet te gek is en dat ze geloofwaardig is. Voor een goed begrip van de theorie beginnen we echter met een niet-geloofwaardige functie, die als volgt geconstrueerd wordt. Voor een item *i* veronderstelt men dat er een bepaalde hoeveelheid vaardigheid nodig is om een correct antwoord te produceren. Iemand die over minder vaardigheid beschikt zal nooit een correct antwoord geven, de kans op een correct antwoord is 0, terwijl iemand met meer

vaardigheid het item altijd juist beantwoordt, dat wil zeggen met kans 1. De grafiek van de itemresponsfunctie is weergegeven in figuur 4.1. Merk op dat de grafiek van de functie een sprong maakt op de plaats *i*. In dezelfde figuur is ook de plaats aangegeven voor een moeilijker item *j*. Dit item is moeilijker dan item *i*, omdat de minimale vaardigheid vereist voor een correct antwoord op item *j* groter is dan voor item *i*.

Deze theorie ziet er misschien aantrekkelijk uit, want ze impliceert het principe: wie een moeilijk item *(j)* juist beantwoordt, geeft ook een juist antwoord op een gemakkelijker item *(i)*. Een verzameling items, waarbij bovenstaande uitspraak geldig is voor alle paren wordt een Guttman-schaal genoemd, naar een van de grondleggers van de moderne testtheorie (Guttman, 1950). Deze theorie is echter niet erg geloofwaardig, omdat het in de praktijk bijna nooit voorkomt dat er in de steekproef niemand is die dit principe schendt. Eén inbreuk op dit principe is voldoende om de theorie te verwerpen. Uit inspectie van figuur 4.1 konden we eigenlijk al dit soort moeilijkheden verwachten. Omdat de kans op een juist antwoord altijd precies 0 of 1 is, leggen we de waarde van X_i volledig vast als we θ kennen, en in de praktijk kunnen we daarvoor gestraft worden. Dergelijke modellen noemt men deterministisch. In de IRT werkt men meestal met itemresponsfuncties die nooit exact de waarde 0 of 1 aannemen. Een andere eigenschap die de functies in figuur 4.1 onrealistisch maken is de sprong op een bepaald punt van 0 naar 1: de functies zijn discontinu.



Figuur 4.1
Itemresponsfunctie in een deterministisch model

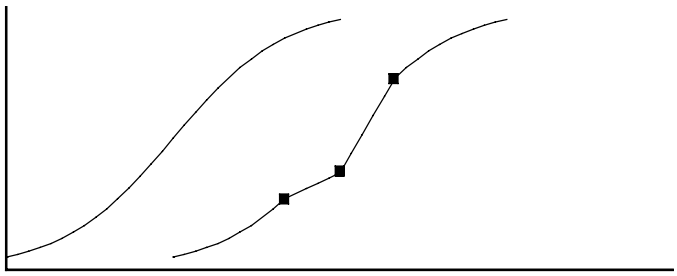
Wat we dan wel weer als een realistische eigenschap kunnen beschouwen, is dat de functies in figuur 4.1 nooit dalen: de kans op een juist antwoord wordt nooit kleiner als de vaardigheid toeneemt. We gaan deze eigenschap aanscherpen door te eisen dat de functie overal stijgend moet zijn, dat wil zeggen dat ze niet constant mag blijven in een bepaald gebied.

Samengevat stellen we de volgende eisen aan de itemresponsfunctie:

(1) $0 < f_i(\theta) < 1$;

(2) de functie is continu: de grafiek moet getekend kunnen worden zonder de pen op te tillen;

(3) de functie is strikt stijgend.



Figuur 4.2
Een 'vloeiende' en een 'hoekige' itemresponsfunctie

Figuur 4.2 toont twee grafieken die aan deze drie eisen voldoen. Een eigenschap die de twee grafieken onderscheidt is de 'hoekigheid'. Functies die dit soort hoekigheid vertonen zijn wiskundig meestal niet elegant om mee te werken. Daarom sluiten we hoekige functies uit door een vierde eis:

(4) de functie moet een vloeiend verloop hebben, of exacter uitgedrukt: de functie moet overal differentieerbaar zijn.

Hoewel de vier gestelde eisen een groot aantal functies uitsluiten, blijven er nog heel veel functies over die aan alle gestelde eisen voldoen. Door één specifieke functie te kiezen perkt men de theorie verder in tot één speciaal geval. Zo'n speciaal geval noemt men een IRT-model. Een specifieke keuze baseert men op een veelheid aan argumenten. Op deze argumenten gaan we hier niet verder in, tenzij door op te merken dat mathematische hanteerbaarheid vaak een belangrijke overweging is.

In de rest van het hoofdstuk beperken we ons tot een eenvoudig IRT-model dat in de literatuur veel aandacht heeft gekregen. Het werd in 1960 voorgesteld door de Deense statisticus G. Rasch (Rasch, 1960, 1980). Meer ingewikkelde modellen worden in hoofdstuk 5 besproken.

4.1.1 Het Raschmodel

In het Raschmodel is de itemresponsfunctie een logistische functie. De logistische functie van een argument *y* wordt gedefinieerd als

$$f(y) = \frac{\exp(y)}{1+\exp(y)}. \tag{4.4}$$

In het Raschmodel is het argument van de logistische functie het verschil $(\theta - \beta_i)$, waarbij β_i een kengetal is dat item *i* karakteriseert. Vervangen we nu in het rechterlid van (4.4) het argument *y* door dit verschil, dan krijgen we

$$f_i(\theta) = \frac{\exp(\theta - \beta_i)}{1+\exp(\theta - \beta_i)}. \tag{4.5}$$

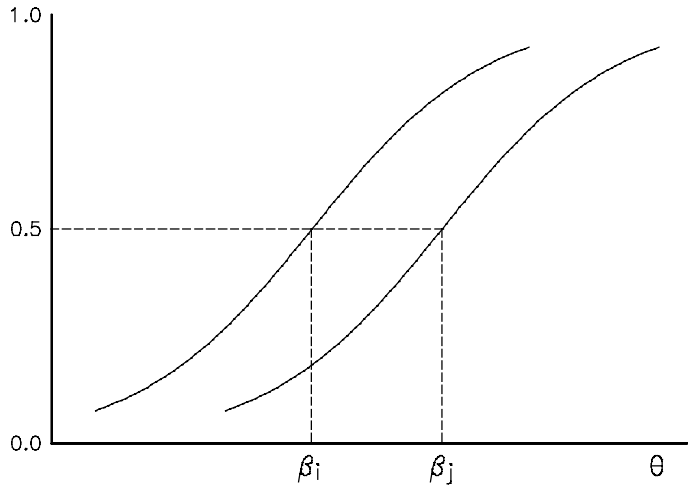
Het zal duidelijk zijn dat door de waarde van β_i te veranderen een andere functie ontstaat. Omdat we nu nog niets willen zeggen over de precieze waarde van β_i, definieert (4.5) in feite een hele familie van functies die allemaal aan de logistische functieregel voldoen. We doen een eenvoudig functieonderzoek van (4.4). Het is gemakkelijk na te gaan dat de logistische functie $f(y)$ altijd tussen 0 en 1 ligt: de teller is steeds positief en de noemer is groter dan de teller. Bovendien geldt dat $f(0) = 0.5$. Dus geldt dat

$$f_i(\beta_i) = 0.5 \tag{4.6}$$

Het is bovendien eenvoudig na te gaan dat de volgende twee limieten gelden:

$$\begin{aligned} &\lim_{\theta\to\infty} f_i(\theta) = 1, \\ &\lim_{\theta\to-\infty} f_i(\theta) = 0. \end{aligned} \tag{4.7}$$

In figuur 4.3 staan twee itemresponsfuncties afgebeeld. Twee punten van commentaar op bovenstaand functie onderzoek. Formule (4.6) betekent dat, indien de vaardigheid precies gelijk is aan het getal β_i, de kans op een juist antwoord precies 0.5 is. Omgekeerd kunnen we β_i interpreteren als de hoeveelheid vaardigheid die nodig is om een kans te hebben van 0.5 op een juist antwoord. In figuur 4.3 zien we dat meer vaardigheid vereist is om die kans te halen bij item *j* dan bij item *i*. Het is dus gerechtvaardigd om te zeggen dat β_i de moeilijkheid uitdrukt van item *i*. De parameter β_i wordt daarom vaak de moeilijkheids- parameter van het item genoemd. Omdat er in het Raschmodel met elk item slechts een parameter gemoeid is, wordt β_i ook vaak kortweg de itemparameter genoemd.



Figuur 4.3
Twee itemresponsfuncties in het Raschmodel

Het tweede commentaar heeft betrekking op (4.7). Voor zeer kleine waarden van θ is de kans bijna 0 dat een correct antwoord wordt gegeven. Dit betekent dat het Raschmodel eigenlijk ongeschikt is voor items waarvan het juiste antwoord door raden tot stand komt. Dit betekent dat extra voorzichtigheid geboden is wanneer het Raschmodel wordt toegepast bij meerkeuze-items: iemand die helemaal niets weet over het gevraagde onderwerp heeft een substantiële kans op een juist antwoord als hij gaat raden.

Een inspectie van figuur 4.3 laat zien dat de twee curven een identieke vorm hebben; ze zijn alleen verschoven ten opzichte van elkaar. Dit betekent ook dat ze elkaar nooit kruisen. Daaruit volgt dat $f_i(\theta) > f_j(\theta)$ voor elke waarde van θ. In woorden: wat ook de waarde van θ is, de kans om item *i* juist te maken is steeds groter dan de kans om item *j* juist te maken.

4.1.2 Lokale stochastische onafhankelijkheid

Formule (4.5) beschrijft het gedrag van iemand met vaardigheid θ op één item. Dit is echter niet voldoende om het Raschmodel te karakteriseren. Er moet ook nog iets gezegd worden over het gedrag, indien meer items moeten worden beantwoord. Stel dat we over vier items beschikken die precies even moeilijk zijn, en we leggen die items voor aan twee personen waarvan we weten dat ze dezelfde θ-waarde hebben. Na het beantwoorden van de eerste drie items stellen we vast dat de eerste persoon drie juiste antwoorden heeft gegeven en de tweede persoon drie onjuiste. Is het dan niet redelijk

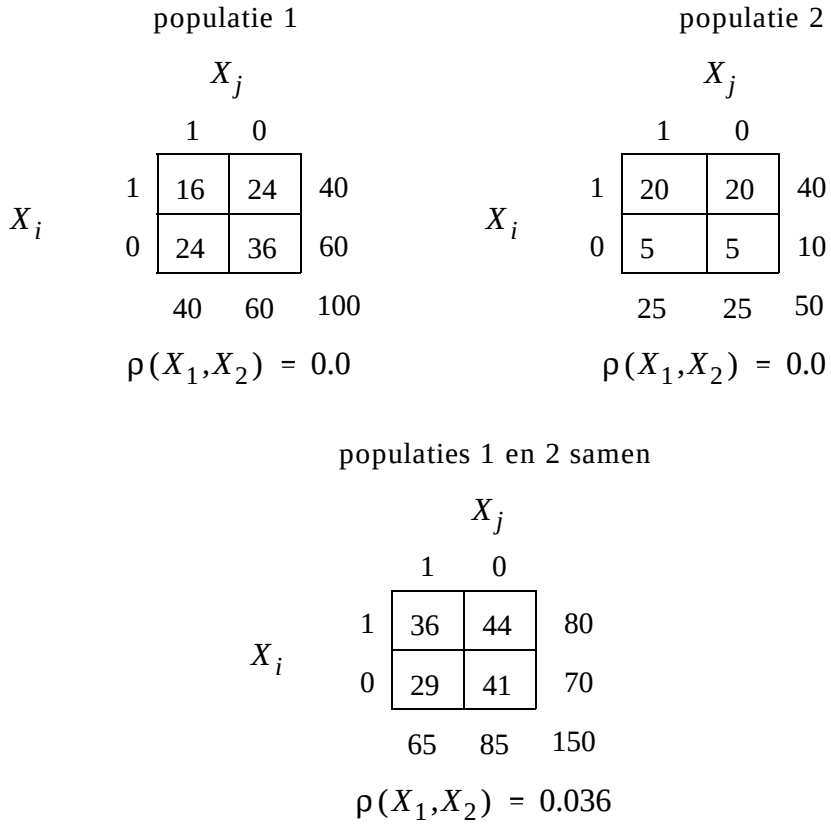
te veronderstellen dat de eerste persoon een grotere kans heeft om het vierde item juist te maken dan de tweede persoon? De eerste persoon heeft immers er blijk van gegeven vaardiger te zijn dan de tweede, gezien zijn drie juiste antwoorden. Het antwoord luidt: neen. Immers, als we aannemen dat het Raschmodel geldig is, dan hangt de kans op een juist antwoord alleen af van de vaardigheid en de moeilijkheid van het item, en in de beschreven situatie gaat het om items met dezelfde moeilijkheid en om personen met dezelfde vaardigheid. Dus moeten die kansen gelijk zijn. Kennis van antwoorden op andere items kan die kans niet veranderen. Deze redenering volgt niet automatisch uit (4.5); ze wordt toegevoegd als een onafhankelijk principe of axioma, namelijk het axioma der lokale stochastische onafhankelijkheid. Dit principe kan op verschil- lende equivalente manieren in formulevorm worden uitgedrukt. We geven twee belangrijke formules. De antwoordvariabelen X_i en X_j zijn lokaal stochastisch onafhankelijk (van elkaar) indien

$$P(X_i=1 \mid \theta \text{ en } X_j=1) = P(X_i=1 \mid \theta) = f_i(\theta), \tag{4.8}$$

of

$$P(X_i=1 \text{ en } X_j=1 \mid \theta) = P(X_i=1 \mid \theta)\, P(X_j=1 \mid \theta) = f_i(\theta)\, f_j(\theta). \tag{4.9}$$

Let wel (4.8) en (4.9) zijn niet twee verschillende voorwaarden; ze zijn equivalent en betekenen dus precies hetzelfde. De beperking 'lokaal' wijst erop dat X_i en X_j alleen onafhankelijk zijn bij gelijke θ. Daaruit volgt niet dat X_i en X_j onafhankelijk zijn van elkaar. Dus uit lokale stochastische onafhankelijkheid volgt niet dat $P(X_i=1 \text{ en } X_j=1) = P(X_i=1) \times P(X_j=1)$. Immers, indien dit waar zou zijn, dan zou de correlatie tussen de antwoorden op item *i* en item *j* nul bedragen, iets wat in het algemeen niet waar is als die items dezelfde vaardigheid meten. Het principe van de lokale stochastische onafhankelijkheid impliceert wel dat de correlatie tussen X_i en X_j nul is in alle populaties waar θ constant is. Dit geeft ons meteen een aardige manier om de correlatie tussen items te verklaren: als in een populatie de correlatie tussen item *i* en *j* niet nul is, dan komt dat doordat de vaardigheid in die populatie niet constant is. Door de invloed van de vaardigheid te controleren, dat wil zeggen door de vaardigheid constant te houden verdwijnt de correlatie. We illustreren dit aan de hand van een voorbeeld. In figuur 4.4 is duidelijk te zien dat de variabelen X_i en X_j niet corre- leren in populatie 1 noch in populatie 2. Voegen we de twee populaties echter samen, dan wordt de correlatie positief.



Figuur 4.4
Een voorbeeld van lokale stochastische onafhankelijkheid

Het axioma van de lokale stochastische onafhankelijkheid is zeer belangrijk in de IRT, maar het is erg moeilijk om te controleren of eraan voldaan is. We kunnen namelijk niet te werk gaan op de manier zoals weergegeven in figuur 4.4. Dit zou vereisen dat we de totale steekproef zouden kunnen opdelen in groepjes personen die dezelfde θ-waarde hebben. Doch θ kennen we niet, dus is deze benadering onmogelijk. Voor de toetsconstructeur is het belang- rijk het axioma niet te schenden door items te maken die functioneel afhankelijk zijn van elkaar, waar een juist antwoord op een item een juist antwoord op een ander item veronder- stelt.

4.2 Het schatten van de parameters in het Raschmodel

4.2.1 Grootste-aannemelijkheidsschatters: een voorbeeld

Door het Raschmodel als model voor het beantwoorden van de items aan te nemen zijn we natuurlijk nog niet klaar met het werk. Om (4.4) uit te rekenen moeten we een getalswaarde invullen voor θ en voor β_i en die getallen kennen we niet. θ en β_i worden parameters ge- noemd en men gebruikt de observaties om schattingen te maken van de parameters.

Er zijn verschillende manieren om parameters te schatten. Hier wordt er één besproken, namelijk de grootste-aannemelijkheidsmethode. In het Engels: maximum likelihood, afgekort als ML. De ML-methode wordt verreweg het meest gebruikt in de IRT-literatuur; ze heeft bepaalde theoretische voordelen waarop later uitvoerig wordt teruggekomen. We leggen de methode uit aan de hand van een voorbeeld. Een onzuiver muntstuk wordt vijf maal opgegooid, waarbij de uitkomst munt als een succes beschouwd wordt en de uitkomst kruis als een mislukking. We definiëren weer toevalsvariabelen X_i als

$$X_i = \begin{cases} 1 \text{ indien munt bij de } i\text{-de beurt,} \\ \\ 0 \text{ indien kruis bij de } i\text{-de beurt, } (i = 1, \ldots, 5). \end{cases}$$

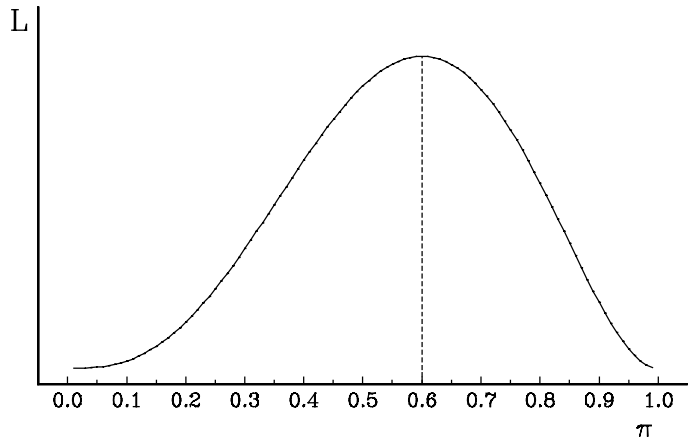
Het model is zeer simpel. Het zegt dat de kans op succes bij opgooien gelijk is aan π, waarbij π een getal is tussen 0 en 1. Wij willen de uitkomst van ons kleine experimentje gebruiken om π te schatten. Stel dat we de volgende uitkomst waarnemen: (1 0 1 1 0). De kans op die uitkomst is

$$\begin{aligned} P(X_1=1, X_2=0, X_3=1, X_4=1, X_5=0; \pi) &= \pi(1-\pi)\pi\pi(1-\pi) \\ &= \pi^3(1-\pi)^2. \end{aligned} \tag{4.10}$$

Formule (4.10) kunnen we op twee manieren bekijken. We kunnen de uitkomst van het experiment als argument van de functie P bekijken en voor alle mogelijke uitkomsten van het experiment een uitdrukking vinden die analoog is aan het rechterlid van (4.10). Dan vinden we een aantal uitdrukkingen waarin π verschijnt als een vast, hoewel nog onbekend, getal. Daarom staat π na de ';' in het linkerlid van (4.10). We kunnen (4.10) echter ook bekijken als een functie van π, waarbij we de uitkomst van ons experiment beschouwen als een gegeven. Voor elke waarde van π die we dan invullen, krijgen we als uitkomst hoe waar- schijnlijk onze observaties zijn, als π die waarde aanneemt. De functie (4.10) zo bekeken noemt men de aannemelijkheidsfunctie (Engels: likelihood function) en die wordt gegeven door

$$L(\pi; (1\,0\,1\,1\,0)) = P((1\,0\,1\,1\,0); \pi). \tag{4.11}$$

De grafiek van het rechterlid van (4.11) is weergegeven in figuur 4.5.



Figuur 4.5
Aannemelijkheidsfunctie voor de observatie (1 0 1 1 0)

De ML-schatting van π is die waarde van π waarvoor de aannemelijkheidsfunctie zo groot mogelijk wordt, dat wil zeggen die waarde waarvoor de gegeven observaties de grootste waarschijnlijkheid hebben. In het voorbeeld is dit 0.6 zoals makkelijk uit figuur 4.5 kan worden afgelezen. Natuurlijk zal men niet steeds een grafiek van de aannemelijkheidsfunctie maken om de schatting te bepalen. Men gebruikt een standaardtechniek, die hier even kort wordt besproken.

Aan de manier waarop (4.10) is opgesteld kan men duidelijk zien dat de volgorde waarin successen en mislukkingen zich voordoen tijdens het experiment niet belangrijk is voor de aannemelijkheidsfunctie; alleen het aantal successen en mislukkingen telt. Indien er *n* keer wordt opgegooid en er zijn *s* successen, dan zijn er *n-s* mislukkingen. Stellen we de uitkomsten van een experiment voor door $\boldsymbol{x} = (x_1,\ldots,x_n)$ dan krijgen we als algemene uitdrukking voor de aannemelijkheidsfunctie

$$L(\pi;\boldsymbol{x}) = \pi^s(1-\pi)^{n-s}, \tag{4.12}$$

waarin $s = \sum_{i=1}^{n} x_i$. Om het maximum van (4.12) te zoeken kiest men gewoonlijk een andere
functie waarvan men weet dat ze monotoon is met de aannemelijkheidsfunctie. De functie die meestal wordt gebruikt is de logaritme van de aannemelijkheidsfunctie:

$$\ln L(\pi;\boldsymbol{x}) = s\ln\pi + (n-s)\ln(1-\pi). \tag{4.13}$$

Een standaardmanier om een maximum van een functie te zoeken is, de eerste afgeleide van die functie te bepalen, die afgeleide gelijk te stellen aan nul en de aldus ontstane vergelijking op te lossen naar de onbekende parameter. Deze vergelijking wordt schattingsvergelijking of aannemelijkheidsvergelijking genoemd. De eerste afgeleide van (4.13) is

$$\frac{d \ln L(\pi;\boldsymbol{x})}{d\pi} = \frac{s}{\pi} - \frac{n-s}{1-\pi}. \tag{4.14}$$

Gelijkstellen van (4.14) aan 0 geeft als oplossing

$$\hat{\pi} = \frac{s}{n}. \tag{4.15}$$

Het rechterlid van (4.15) is een functie van de gegevens. We zien dus dat we een algemene oplossing krijgen voor het muntexperiment: de grootste-aannemelijkheids-schatter is het aantal successen gedeeld door het aantal keren opgooien. De functie s/n wordt de schatter genoemd. De waarde die die functie aanneemt in een concreet geval wordt de schatting genoemd. In het voorbeeld is de schatting van π dus gelijk aan 0.6. Het dakje boven het parametersymbool wordt gebruikt om aan te geven dat het hier niet gaat om de echte waarde van π, maar om een schatter of schatting. De schatter is een functie van het aantal successen, en dit aantal is een toevalsvariabele; dus is de schatter ook een toevalsvariabele, en de schatting zelf zal van experiment tot experiment verschillen.

Omdat we meestal niet een zeer groot aantal experimenten uitvoeren maar slechts één, blijven we met de vraag zitten of de schatting die we in een concreet geval voor π krijgen wel een goede schatting is. Bovendien is er nog een ander probleem: de oplossing (4.15) garandeert ons alleen dat de eerste afgeleide van (4.14) 0 is indien $\pi = s/n$, doch daaruit volgt niet automatisch dat dit punt met een maximum overeenkomt. Daartoe moeten we hogere afgeleiden van (4.14) onderzoeken. Indien de tweede afgeleide negatief is op het punt waar de eerste afgeleide nul wordt weten we dat we te doen hebben met een maximum. De tweede afgeleide van de log-aannemelijkheidsfunctie is gegeven door

$$\frac{d^2 \ln L(\pi;\boldsymbol{x})}{d\pi^2} = -\frac{s}{\pi^2} - \frac{n-s}{(1-\pi)^2}, \tag{4.16}$$

en deze functie is negatief voor alle waarden van π in het interval (0,1). (De gevallen waar $\pi = 0$ en $\pi = 1$ laten we buiten beschouwing.) De oplossing (4.15) komt dus overeen met een maximum van de aannemelijkheidsfunctie.

De tweede afgeleide kunnen we ook gebruiken om iets te zeggen over de nauwkeurigheid van de ML-schatter van π. In de theoretische statistiek zijn belangrijke resultaten bekend over de statistische eigenschappen van ML-schatters. Hoewel deze resultaten niet altijd geldig zijn, zijn ze wel bruikbaar voor de modellen die in dit boek worden behandeld. Bovendien staan deze resultaten bekend als 'asymptotische' resultaten, dit wil zeggen dat ze strikt genomen alleen geldig zijn als $n \to \infty$. In de praktijk kunnen ze echter goed gebruikt worden als de steekproef niet al te klein is. Het belangrijkste resultaat luidt:

> De ML-schatter is asymptotisch normaal verdeeld met gemiddelde de werkelijke parameter π van het model en als variantie één gedeeld door de informatiefunctie. (Zie bijvoorbeeld Kendall & Stuart, 1973.)

De informatiefunctie $I(\pi)$ met betrekking tot de parameter π is gedefinieerd als

$$I(\pi) = -\mathscr{E}\left[\frac{d^2 \ln L(\pi;\boldsymbol{x})}{d\pi^2}\right], \tag{4.17}$$

waarbij de verwachte waarde genomen dient te worden over alle mogelijke steekproeven (met vaste n). In het voorbeeld met het muntstuk geeft dit

$$\begin{aligned} I(\pi) &= -\mathscr{E}\left[\frac{d^2 \ln L(\pi;\boldsymbol{x})}{d\pi^2}\right] \\ &= \frac{\mathscr{E}(s)}{\pi^2} + \frac{n - \mathscr{E}(s)}{(1-\pi)^2} \\ &= \frac{n\pi}{\pi^2} + \frac{n(1-\pi)}{(1-\pi)^2} = \frac{n}{\pi(1-\pi)}. \end{aligned} \tag{4.18}$$

Uit (4.18) en het bovengenoemde resultaat volgt onmiddellijk dat de schatter $\hat{\pi} = s/n$ asymptotisch normaal verdeeld is met gemiddelde π en variantie $\pi(1-\pi)/n$, een resultaat dat in elke cursus statistiek gepresenteerd wordt. Om de variantie uit te rekenen moeten we echter de waarde van π kennen. Omdat die niet bekend is, vult men daarvoor de ML-schatting in van π. Dit geeft dus als resultaat

$$\sigma^2(\hat{\pi}) \approx \frac{1}{I(\hat{\pi})} = \frac{\hat{\pi}(1-\hat{\pi})}{n}. \tag{4.19}$$

Het teken '$\approx$' geeft aan dat de gelijkheid slechts asymptotisch geldt; de echte standaardfout bij een eindige steekproef is in de regel groter dan door (4.19) is

aangegeven. De standaardfout (verder afgekort als *SE*, van het Engelse standard error), dit is de vierkantswortel uit (4.19), kan gebruikt worden om bijvoorbeeld betrouwbaarheidsintervallen voor de parameter te berekenen. Passen we (4.19) toe op het voorbeeld, dan vinden we $\sigma^2(\hat{\pi}) \approx .24/5 = .048$. Het 95%-betrouwbaarheidsinterval is dus gegeven door $(\hat{\pi} - 1.96 \times \sqrt{0.48}\,, \hat{\pi} + 1.96 \times \sqrt{0.48}) = (0.17,1.03)$. Dit grote betrouwbaarheidsinterval, dat zich hier uitstrekt buiten het toegestane bereik van de parameter, is te wijten aan de uiterst kleine steekproef, die ons niet veel informatie over de parameter oplevert. Hadden we 50 keer opgegooid met het muntstuk, dan hadden we bij 30 successen een variantie gekregen van .0048, en een standaardfout die $10^{½} = 3.16$ zo klein was, en dus ook een betrouwbaar- heidsinterval dat 3.16 kleiner is: (0.46,0.74).

In de literatuur wordt nog een andere manier gebruikt om een schatting van de standaardfout te verkrijgen. In plaats van de verwachte waarde te nemen van minus de tweede afgeleide van de log-aannemelijkheidsfunctie, neemt men gewoon minus de tweede afgeleide van de log-aannemelijkheidsfunctie zelf. Deze functie, geëvalueerd op de ML-schatting, wordt de geobserveerde-informatiefunctie genoemd. Het symbool dat hiervoor gebruikt wordt is *J*. Uit (4.15) volgt dat $s = n\hat{\pi}$. Dus krijgen we, door invullen in (4.16)

$$J(\hat{\pi}) = \frac{n\hat{\pi}}{\hat{\pi}^2} + \frac{n - n\hat{\pi}}{(1-\hat{\pi})^2} = \frac{n}{\hat{\pi}(1-\hat{\pi})}. \qquad (4.20)$$

Het feit dat we voor de informatiefunctie, geëvalueerd op de ML-schatter, en voor de geobserveerde informatiefunctie hetzelfde resultaat krijgen is niet toevallig en heeft te maken met een speciale eigenschap van de log-aannemelijkheidsfunctie. Het is niet moeilijk na te gaan dat de log-aannemelijkheidsfunctie geschreven kan worden als

$$\ln L(\pi; \boldsymbol{x}) = s \ln \frac{\pi}{1-\pi} + n \ln (1-\pi). \qquad (4.21)$$

De eerste term in het rechterlid van (4.21) is een produkt van twee factoren: de eerste factor is een functie van de gegevens (*s*) en de tweede factor is een functie van de parameter. De
tweede term is alleen een functie van de parameter π (*n* dient beschouwd te worden als een constante). Dit is een iets gespecialiseerde vorm van een meer algemene vorm van de log-aannemelijkheidsfunctie. Indien men een model beschouwt met meer dan één parameter, bijvoorbeeld *k*, waarbij de parameters verzameld zijn in de *k*-vector π, en men kan de log-aannemelijkheidsfunctie schrijven als

$$\ln L(\pi;\boldsymbol{x}) = \sum_{i=1}^{k} A_i(\boldsymbol{x})\, B_i(\pi) + C(\pi) + D(\boldsymbol{x}), \tag{4.22}$$

waarin A_i en D functies zijn van de gegevens maar niet van de parameters, en B_i en C functies zijn van de parameters maar niet van de gegevens, dan zegt men dat de log-aannemelijkheidsfunctie (of het model) behoort tot de exponentiële familie. Formule (4.21) is gemakkelijk te herkennen als een speciaal geval van (4.22), met $k = 1$, $A_1 = s$, $B_1 = \ln[\pi/(1-\pi)]$, $C = n\ln(1-\pi)$ en $D = 0$. De exponentiële familie heeft veel prettige eigenschappen, en één ervan is dat de informatiefunctie, geëvalueerd op de ML-schatter, en de geobserveerde informatiefunctie gelijk zijn aan elkaar.

Tenslotte nog een opmerking over de functies A_i in (4.22). Deze functies worden de minimaal voldoende steekproefgrootheden, in het Engels: minimal sufficient statistics, genoemd voor de functies $B_i(\pi)$. Dat een steekproefgrootheid voldoende is om de parameter te schatten, betekent dat we van de observaties niet méér gebruiken dan door deze grootheid wordt aangegeven. Bij het muntstuk experiment is het aantal successen voldoende om de parameter π te schatten; de precieze afwisseling van successen en mislukkingen levert geen bijkomende informatie over de parameter. Op de term 'minimaal' dienen we echter nog even in te gaan. Stel dat de k-de functie $B_k(\pi)$ in (4.22) kan geschreven worden als een lineaire combinatie van de $k-1$ andere functies $B_i(\pi)$, dat wil zeggen dat er getallen $\alpha_1,\ldots,\alpha_{k-1}$ bestaan zodat

$$\begin{aligned} B_k(\pi) &= \alpha_1 B_1(\pi) + \ldots + \alpha_{k-1} B_{k-1}(\pi) \\ &= \sum_{i=1}^{k-1} \alpha_i B_i(\pi), \end{aligned} \tag{4.23}$$

dan kan (4.22) geschreven worden als

$$\begin{aligned} \ln L(\pi;\boldsymbol{x}) &= \sum_{i=1}^{k-1} A_i(\boldsymbol{x})\, B_i(\pi) + A_k(\boldsymbol{x}) \sum_{i=1}^{k-1} \alpha_i B_i(\pi) + C(\pi) + D(\boldsymbol{x}) \\ &= \sum_{i=1}^{k-1} [A_i(\boldsymbol{x}) + \alpha_i A_k(\boldsymbol{x})]\, B_i(\pi) + C(\pi) + D(\boldsymbol{x}). \end{aligned} \tag{4.24}$$

Doch de factor tussen [ ] in het rechterlid van (4.24) is geen functie van de parameters, en dus is (4.24) een log-aannemelijkheidsfunctie uit de exponentiële familie, maar nu met $k-1$ parameters. Op analoge manier kan men soms het aantal parameters verminderen door aan te tonen dat een functie $A_i(\boldsymbol{x})$ lineair afhankelijk is van de

andere *A*-functies. Als we spreken over het aantal parameters in een model, dan zullen we altijd het aantal bedoelen waarvoor een verdere restrictie als gegeven in (4.23) niet meer mogelijk is. Deze parameters worden ook wel aangeduid als vrije parameters.

4.2.2 JML-schatting in het Raschmodel

In het Raschmodel kunnen we proberen op een soortgelijke manier te werk te gaan als in de vorige paragraaf. De principes blijven dezelfde, er is alleen een complicatie omdat we nu niet één parameter moeten schatten, maar verschillende tegelijkertijd. Nemen we een toets bestaande uit *k* items af aan *n* personen, dan moeten we *n* θ-parameters schatten en *k* itemparameters. De J in JML staat voor 'joint'. Men gebruikt deze aanduiding niet om aan te geven dat er meer parameters geschat moeten worden, maar om aan te geven dat de twee soorten parameters, persoonsparameters en itemparameters, tegelijkertijd geschat worden. Om de aannemelijkheidsfunctie op te stellen moeten we de notatie iets uitbreiden. De toevalsvariabele X_{vi} verwijst naar het antwoord van persoon *v* op item *i*. De waarden die die toevalsvariabele kan aannemen, 0 of 1, zullen we in het algemeen aanduiden met $\boldsymbol{x}_{vi}$. Willen we verwijzen naar de antwoorden van persoon *v*, dan wordt dit aangeduid met $\boldsymbol{x}_v$, en willen we verwijzen naar alle antwoorden van alle personen in de steekproef dan wordt dit aangeduid met $\boldsymbol{X}$.

Beschouw eerst als voorbeeld een steekproef van een persoon *v*, met $\theta = \theta_v$, en een toets van k=3 items. Veronderstel dat we de antwoorden (1,0,1) hebben geobserveerd. Gebruik makend van het principe van de lokale stochastische onafhankelijkheid en van formule (4.3), kan de aannemelijkheidsfunctie voor dit antwoordpatroon geschreven worden als

$$L(\beta_1,\beta_2,\beta_3,\theta_v;(1\,0\,1)) = f_1(\theta_v)(1-f_2(\theta_v))f_3(\theta_v). \qquad (4.25)$$

Merk op dat bovenstaand produkt bestaat uit k=3 factoren, dat met een juist antwoord op item *i* een factor $f_i(\theta_v)$ overeenkomt, en met een verkeerd antwoord een factor $(1-f_i(\theta_v))$. Om een algemene formule te verkrijgen, wordt het produkt in (4.25) uitgebreid tot 2*k* factoren, twee per item. Het produkt van die twee factoren heeft de gedaante

$$[f_i(\theta_v)]^{x_{vi}}[1-f_i(\theta_v)]^{1-x_{vi}}.$$

Indien $x_{vi} = 1$ is dit produkt gelijk aan $f_i(\theta_v)$, en indien $x_{vi} = 0$, is het produkt gelijk aan $(1 - f_i(\theta_v))$. Duiden we nu met β de vector $(\beta_1,...,\beta_k)$ aan, dan krijgen we als directe veralgemening van (4.25):

$$L(\beta,\theta_v;\boldsymbol{x}_v) = \prod_{i=1}^{k} [f_i(\theta_v)]^{x_{vi}} [1-f_i(\theta_v)]^{1-x_{vi}}. \tag{4.26}$$

Veralgemenen we dit nu tot een steekproef van *n* personen. Elke persoon levert een aannemelijkheidsfunctie op van de gedaante (4.26). De aannemelijkheidsfunctie voor alle gegevens samen is het produkt van de aannemelijkheidsfunctie voor alle antwoordpatronen afzonderlijk. Dit is waar indien de antwoorden van de personen onafhankelijk zijn van elkaar. Let wel, de reden is niet de lokale stochastische onafhankelijkheid, want we kunnen er niet van uitgaan dat alle personen de zelfde θ-waarde hebben. Onafhankelijkheid betekent hier dat de antwoorden van de ene persoon geen informatie bevatten over de antwoorden van een andere persoon. Dit soort onafhankelijkheid wordt in de testtheorie experimentele onafhankelijkheid genoemd. Duiden we de vector $(\theta_1,...,\theta_n)$ aan met θ, dan vinden we

$$L(\beta,\theta;\boldsymbol{X}) = \prod_{v=1}^{n}\prod_{i=1}^{k} [f_i(\theta_v)]^{x_{vi}} [1-f_i(\theta_v)]^{1-x_{vi}}. \tag{4.27}$$

Substitueren we nu (4.5) in (4.27), en nemen we de logaritme, dan vinden we

$$\ln L(\beta,\theta;\boldsymbol{X}) = \sum_{v=1}^{n} s_v\theta_v + \sum_{i=1}^{k} t_i(-\beta_i) - \sum_{v=1}^{n}\sum_{i=1}^{k} \ln[1 + \exp(\theta_v - \beta_i)], \tag{4.28}$$

waarin

$$s_v = \sum_{i=1}^{k} x_{vi}, \qquad t_i = \sum_{v=1}^{n} x_{vi}.$$

Het is makkelijk in te zien dat (4.28) een log-aannemelijkheidsfunctie uit de exponentiële familie is, met s_v, $v = 1,...,n$ en t_i, $i = 1,...,k$, de voldoende steekproef-grootheden voor respectievelijk θ_v, $v = 1,...,n$, en $(-\beta_i)$, $i = 1,...,k$. De laatste term in (4.28) komt overeen met de functie *C* in (4.22). Er geldt echter:

$$\sum_v s_v = \sum_i t_i,$$

dat wil zeggen dat er een lineaire restrictie op de grootheden s_v en t_i ligt. Er zijn dus niet $k + n$ maar hoogstens $k + n - 1$ vrije parameters; meer parameters kunnen dus ook

niet ge- schat worden. Dit betekent dat het Raschmodel in zijn algemeenheid niet schatbaar is, of zoals men het ook uitdrukt: het model is niet geïdentificeerd. Dit valt reeds af te leiden uit de itemresponsfunctie (4.5). Stel dat we van alle personen θ_v en van alle items β_i kennen. Een andere, doch evenwaardige oplossing bestaat erin aan elke persoon v het getal $\theta_v^* = \theta_v + c$ en aan elk item het getal $\beta_i^* = \beta_i + c$ toe te kennen, waarbij c een willekeurige constante is. Dan geldt natuurlijk dat $\theta_v^* - \beta_i^* = \theta_v - \beta_i$, en dus blijft de itemresponsfunctie onveranderd welke waarde we ook aan c geven. Willen we zinvol over de parameters kunnen spreken dan moeten we de waarde van c vastleggen, of met ander woorden, we moeten het nulpunt van de schaal vastleggen. Dit kunnen we doen door bijvoorbeeld één van de parameters (bijvoorbeeld β_1) gelijk te stellen aan nul. Doch in dat geval zijn er nog maar $k - 1$ vrije itemparameters over, hetgeen in overeenstemming is met de bovenvermelde lineaire restrictie. Het kiezen van het nulpunt noemt men normaliseren. De meest gebruikte normalisatie is het nulpunt zo te kiezen dat $\sum_{i=1}^{k} \beta_i = 0$.

Om het maximum van (4.28) te vinden, kan men een generalisatie van de techniek toepassen die in paragraaf 4.2.1 werd besproken. Op het maximum van een functie van meerdere parameters moeten alle partiële afgeleiden gelijk zijn aan nul. De partiële afgeleide van een functie naar een parameter is de afgeleide van de functie naar die parameter, waarbij alle andere parameters als constante worden beschouwd. We hoeven deze exercitie echter niet uit te voeren omdat we gebruik kunnen maken van een resultaat dat geldig is in de exponentiële familie. Dit resultaat luidt:

> In een exponentieel familie model zijn de aannemelijkheidsvergelijkingen gegeven door de voldoende steekproefgrootheden gelijk te stellen aan hun verwachte waarde (Andersen, 1980).

Dit geeft dus voor de θ-parameters:

$$\begin{aligned} s_v &= \mathscr{E}(S_v) = \mathscr{E}\left[\sum_i X_{vi}\right] = \sum_i \mathscr{E}(X_{vi}) \\ &= \sum_i [1 \times P(X_{vi}=1 \mid \theta_v) + 0 \times P(X_{vi}=0 \mid \theta_v)] \\ &= \sum_i f_i(\theta_v), \qquad (v = 1, \ldots, n), \end{aligned} \tag{4.29}$$

waarin S_v de toevalsvariabele 'score van persoon v' aanduidt met als realisatie de geobser- veerde score s_v. Zij T_i de toevalsvariabele 'aantal juiste antwoorden gegeven op item i', dan worden de schattingsvergelijkingen voor de β-parameters gegeven door

$$t_i = \mathscr{E}(T_i) = \sum_v f_i(\theta_v), \qquad (i = 2, \ldots, k). \tag{4.30}$$

In (4.30) is geen vergelijking opgenomen voor i=1. Dit betekent dat β_1 niet beschouwd wordt als een parameter die geschat moet worden, maar als een bekende constante. De waarde die we aan β_1 geven is in principe willekeurig; wij zullen echter aannemen dat $\beta_1 = 0$. Merk op dat (4.29) en (4.30) een stelsel van vergelijkingen vormen in $k+n-1$ onbekenden. Dit stelsel kan niet expliciet worden opgelost, de oplossing wordt gezocht met een iteratieve procedure, waarbij in elke iteratie aan de parameters waarden worden toegekend die de oplossing steeds dichter benaderen. Op de technische aspecten van deze oplossingsmethode gaan we hier niet in.

Er zijn echter twee problemen verbonden met het stelsel gevormd door (4.29) en (4.30). Het eerste is gemakkelijk duidelijk te maken. Stel dat er een persoon v is in de steekproef die alle items juist heeft beantwoord. Dan geldt dat het linkerlid in (4.29) gelijk is aan k. Het rechterlid bestaat uit k termen die alle strikt kleiner zijn dan 1, dus hun som is kleiner dan k, welke waarden men ook voor de parameters invult. Een analoog probleem krijgt men wanneer $s_v = 0$. Bij de vergelijkingen (4.30) geldt hetzelfde argument indien $t_i = n$ of $t_i = 0$. In deze gevallen bestaat er dus geen schatter van de parameter.

Het tweede probleem is van theoretische aard en heeft betrekking op een eigenschap van schatters die men consistentie noemt (Kendall & Stuart, 1973). Ruwweg betekent consistentie dat, hoe meer informatie men verzamelt over een parameter door de steekproef steeds groter te maken, des te nauwkeuriger de schatting moet zijn en in de limiet, bij $n \to \infty$ is de kans dat men de parameter juist schat gelijk aan 1. In het geval van het Raschmodel treedt er echter een complicatie op: om meer informatie te verzamelen over itemparameters dient men de toets steeds bij nieuwe personen af te nemen, doch elke persoon die men aan de steekproef toevoegt brengt zijn eigen onbekende θ-parameter mee. Dit wil zeggen dat de omvang van het probleem, het aantal te schatten parameters, even snel groeit als de steekproef. Het gevolg hiervan is dat de JML-schatters van de itemparameters niet consistent zijn. Bovendien gelden de asymptotische resultaten over de standaardfout, die in paragraaf 4.2.1. werden besproken, hier niet automatisch. Dit maakt de JML-schattingsmethode oninteressant. Als men echt in de itemparameters is geïnteresseerd, dan is het veel handiger naar een schattingsmethode te zoeken waarbij men geen last meer heeft van het steeds groeiende aantal θ-parameters. Deze parameters, waar men in eerste instantie niet zo in geïnteresseerd is, maar die toch in het model aanwezig zijn worden in de literatuur aangeduid met de term ’nuisance parameters’. De andere parameters waarin men wel is geïnteresseerd worden structurele parameters genoemd.

In de literatuur zijn verschillende methodes bekend om de 'nuisance parameters' kwijt te raken. In de twee volgende subparagrafen worden twee van deze methodes besproken.

4.2.3 CML-schatting in het Raschmodel

Het is nuttig om even het volgende gedachtenexperiment uit te voeren. De itemresponsfunctie is een conditionele kans om een juist antwoord te geven op een item. Stel nu dat we er in zouden slagen een grote steekproef samen te stellen van personen die allemaal dezelfde θ-waarde hebben, zeg θ_m. Indien aan al die personen hetzelfde item *i* zou worden voorgelegd, dan zal een proportie $p_i(\theta_m)$ het item juist beantwoorden. Deze proportie is een schatting van de conditionele kans $f_i(\theta_m)$ en uit (4.5) volgt dat, als we deze schatter invullen en de logaritme nemen,

$$\hat{\beta}_i = \theta_m - \ln \frac{p_i(\theta_m)}{1 - p_i(\theta_m)}.$$

Passen we deze methode toe op twee items, *i* en *j*, bij dezelfde steekproef, dan volgt uit het bovenstaande direct dat

$$\hat{\beta}_i - \hat{\beta}_j = \ln \frac{p_j(\theta_m)[1 - p_i(\theta_m)]}{p_i(\theta_m)[1 - p_j(\theta_m)]}. \tag{4.31}$$

Dit wil zeggen dat we een schatting krijgen van het verschil tussen twee itemparameters die onafhankelijk is van de θ-parameter, want de proportie $p_i(\theta_m)$ is een direct geobserveerde grootheid. Het bezwaar tegen deze methode is echter dat ze principieel niet uitgevoerd kan worden, omdat de θ-waarde van een persoon niet observeerbaar is; dat wil zeggen dat we geen groep van personen met allemaal dezelfde θ kunnen vormen. Wat echter wel uitvoerbaar is, is het indelen in groepen van personen met dezelfde toetsscore *s*. We bekijken eerst een voorbeeld.

Veronderstel dat $k = 3$ en beschouw het antwoordpatroon (1 0 1). De score *s* van dit ant- woordpatroon is 2. Nu zijn er exact drie mogelijke antwoordpatronen met score 2, namelijk (1 0 1), (1 1 0) en (0 1 1). Conditioneren op score 2 betekent dat we reeds weten dat een van die drie antwoordpatronen is opgetreden, en nu willen we weten wat

de kans is dat (1 0 1) is opgetreden, als alleen die drie mogelijk zijn. De formule hiervoor is

$$P(1\,0\,1|s{=}2,\theta) = \frac{P(1\,0\,1|\theta)}{P(1\,0\,1|\theta) + P(1\,1\,0|\theta) + P(0\,1\,1|\theta)}. \tag{4.32}$$

Bekijken we nu even twee equivalente formules voor het Raschmodel:

$$P(X_i{=}1|\theta) = f_i(\theta) = \frac{\exp(\theta-\beta_i)}{1+\exp(\theta-\beta_i)}, \tag{4.33}$$

en

$$P(X_i{=}0|\theta) = 1 - f_i(\theta) = \frac{1}{1+\exp(\theta-\beta_i)}. \tag{4.34}$$

Als we de aannemelijkheidsfunctie opstellen moeten we produkten nemen van uitdrukkingen met de gedaante (4.33) voor juiste antwoorden of (4.34) voor foute antwoorden. Merk op dat de noemers van (4.33) en (4.34) identiek zijn. De noemer van het produkt is dus onafhankelijk van het specifieke antwoordpatroon. Stel deze noemer voor door het symbool *K*. Beschouw nu de kans op het antwoordpatroon (1 0 1):

$$P(1\,0\,1|\theta) = \frac{\exp(\theta)\exp(-\beta_1)\exp(\theta)\exp(-\beta_3)}{K} = \frac{\exp(2\theta)\exp(-\beta_1-\beta_3)}{K}. \tag{4.35}$$

In de teller van (4.35) komt 2θ voor in de exponent. Het is duidelijk dat die 2 daar staat, omdat het over een antwoordpatroon gaat met precies 2 juiste antwoorden. Doch dit is ook het geval voor de antwoordpatronen (1 1 0) en (0 1 1). Dan is het niet moeilijk in te zien dat

$$\begin{aligned} &P(1\,0\,1\,|s=2,\theta) \\ &= \frac{\dfrac{\exp(2\theta)\exp(-\beta_1-\beta_3)}{K}}{\dfrac{\exp(2\theta)\exp(-\beta_1-\beta_3)}{K} + \dfrac{\exp(2\theta)\exp(-\beta_1-\beta_2)}{K} + \dfrac{\exp(2\theta)\exp(-\beta_2-\beta_3)}{K}} \\ &= \frac{\exp(-\beta_1-\beta_3)}{\exp(-\beta_1-\beta_3)+\exp(-\beta_1-\beta_2)+\exp(-\beta_2-\beta_3)}. \end{aligned} \tag{4.36}$$

Het belangrijke aspect van (4.36) is dat het rechterlid onafhankelijk is van θ en alleen nog een functie van de itemparameters. Bij de vereenvoudiging van (4.36), dat wil zeggen de overgang van het tweede lid naar het derde lid, merken we dat niet alleen de noemers K verdwijnen, maar ook de uitdrukking 2θ. Dit kon alleen maar door ervoor te zorgen dat θ telkens met hetzelfde getal, 2, werd vermenigvuldigd. Maar 2 is precies de score die met de drie beschouwde antwoordpatronen is geassocieerd. De 'truc' om θ te laten verdwijnen werkt dus alleen maar als we conditioneren op de score.

De uitdrukking (4.36), maar nu beschouwd als een functie van de β-parameters, noemen we de conditionele aannemelijkheidsfunctie voor het patroon (1 0 1). Om een algemene formule op te stellen voor de conditionele aannemelijkheid is het handig over te gaan op een andere parametrisering. Definieer

$$\varepsilon_i = \exp(-\beta_i), \quad (i=1,\ldots,k). \tag{4.37}$$

Met deze parameters kan (4.36) geschreven worden als

$$P(1\,0\,1\,|s=2,\theta) = \frac{\varepsilon_1\varepsilon_3}{\varepsilon_1\varepsilon_3+\varepsilon_1\varepsilon_2+\varepsilon_2\varepsilon_3} = \frac{\prod_{i=1}^{k}\varepsilon_i^{x_i}}{\varepsilon_1\varepsilon_3+\varepsilon_1\varepsilon_2+\varepsilon_2\varepsilon_3}. \tag{4.38}$$

De noemer in het rechterlid van (4.38) heeft een merkwaardige structuur: het is een som van drie termen, en elke term is een produkt van twee parameters. De indices van de parameters in elke term kan men opvatten als een aanduiding van de items die men juist moet hebben om een score van 2 te behalen. Er zijn drie termen omdat men slechts op drie verschillende manieren een score van 2 kan behalen. In het algemeen, bij k items en een score s $(s=0,1,\ldots,k)$, zijn er $(k!)/[s!(k-s)!]$ manieren om een score s te behalen. De noemer in de overeenkomstige formule voor de conditionele aannemelijkheid zal dus uit even zo veel termen bestaan, en elke term bestaat uit een produkt van s ε-parameters, waarvan de indices aangeven welke items juist werden beantwoord om de score s te behalen. De noemer is dus een functie van de ε-parameters, en deze functie draagt de naam 'symmetrische basisfunctie'. Voor elke score is er een andere functie; de aanduiding van de score wordt de 'orde'van de functie genoemd. Definiëren we $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_k)$, dan worden de symmetrische basisfuncties van de orde s aangeduid als $\gamma_s(\boldsymbol{\varepsilon})$. Hun definitie is

$$
\begin{aligned}
\gamma_0(\boldsymbol{\varepsilon}) &= 1,\\
\gamma_1(\boldsymbol{\varepsilon}) &= \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_k,\\
\gamma_2(\boldsymbol{\varepsilon}) &= \varepsilon_1\varepsilon_2 + \varepsilon_1\varepsilon_3 + \ldots + \varepsilon_1\varepsilon_k + \varepsilon_2\varepsilon_3 + \ldots + \varepsilon_{k-1}\varepsilon_k,\\
&\vdots\\
\gamma_k(\boldsymbol{\varepsilon}) &= \varepsilon_1\varepsilon_2\ldots\varepsilon_k,
\end{aligned}
\tag{4.39}
$$

$\gamma_s(\boldsymbol{\varepsilon}) = 0$ indien $s < 0$ *of* $s > k$.

De conditionele aannemelijkheidsfunctie, gegeven dat de score gelijk is aan s kunnen we nu dus algemeen schrijven als

$$
L(\boldsymbol{\varepsilon}; \boldsymbol{x} \mid s) = \frac{\prod_{i=1}^{k} \varepsilon_i^{x_i}}{\gamma_s(\boldsymbol{\varepsilon})}. \tag{4.40}
$$

De conditionele aannemelijkheidsfunctie voor alle geobserveerde antwoordpatronen samen is het produkt van soortgelijke uitdrukkingen:

$$
L(\boldsymbol{\varepsilon}; \boldsymbol{X} \mid \boldsymbol{s}) = \frac{\prod_{v=1}^{n} \prod_{i=1}^{k} \varepsilon_i^{x_{vi}}}{\prod_{v=1}^{n} \gamma_{s_v}(\boldsymbol{\varepsilon})}, \tag{4.41}
$$

waarin $\boldsymbol{s} = (s_1, \ldots, s_n)$.

Om de schattingsvergelijkingen op te stellen, hebben we de partiële afgeleiden nodig van de γ-functies naar de ε-parameters. Neem als voorbeeld

$$
\gamma_3(\varepsilon_1,\varepsilon_2,\varepsilon_3,\varepsilon_4) = \varepsilon_1\varepsilon_2\varepsilon_3 + \varepsilon_1\varepsilon_2\varepsilon_4 + \varepsilon_1\varepsilon_3\varepsilon_4 + \varepsilon_2\varepsilon_3\varepsilon_4
$$

en beschouw de partiële afgeleide naar ε_2. Van de term in de uitdrukking hierboven die ε_2 niet bevat is de partiële afgeleide nul, en van de termen die ε_2 wel bevatten is de partiële afgeleide het produkt van de andere ε-parameters. Dus

$$
\frac{\partial \gamma_3(\varepsilon_1,\varepsilon_2,\varepsilon_3,\varepsilon_4)}{\partial \varepsilon_2} = \varepsilon_1\varepsilon_3 + \varepsilon_1\varepsilon_4 + \varepsilon_3\varepsilon_4,
$$

doch dit is eveneens een symmetrische basisfunctie, maar nu van orde 2 en van de parameters (ε_1, ε_3, ε_4). De parameter waarnaar gedifferentieerd wordt, is uit het rijtje weggevallen. In het algemeen krijgen we dan ook de uitdrukking

$$
\frac{\partial \gamma_s(\boldsymbol{\varepsilon})}{\partial \varepsilon_i} = \gamma_{s-1}^{(i)}(\boldsymbol{\varepsilon}), \tag{4.42}
$$

waarbij de *(i)* in superscript aanduidt dat ε_i niet meer tot het argument van de γ-functie behoort.

De logaritme van (4.41) is

$$\ln L(\boldsymbol{\varepsilon}; \boldsymbol{x} \mid \boldsymbol{s}) = \sum_i t_i \ln \varepsilon_i - \sum_v \ln \gamma_{s_v}(\boldsymbol{\varepsilon}), \tag{4.43}$$

waarin weer duidelijk de structuur van de exponentiële familie tot uiting komt: de grootheden t_i zijn de voldoende steekproefgrootheden voor de parameters $\ln(\varepsilon_i)$. Dus ook de conditionele verdeling van X gegeven $\boldsymbol{s}$ behoort tot deze familie. Stellen we de partiële afgeleiden van (4.43) naar ε_i gelijk aan 0, dan krijgen we als schattingsvergelijkingen

$$t_i = \sum_v \frac{\varepsilon_i \gamma^{(i)}_{s_v - 1}(\boldsymbol{\varepsilon})}{\gamma_{s_v}(\boldsymbol{\varepsilon})}, \qquad (i = 2, \ldots, k). \tag{4.44}$$

Gebruik makend van een reeds eerder vermelde eigenschap van de exponentiële familie, kunnen we echter ook schrijven dat

$$t_i = \mathscr{E}(T_i \mid \boldsymbol{s}) = \sum_v \pi_{i \mid s_v}, \qquad (i = 2, \ldots, k), \tag{4.45}$$

waarin $\pi_{i|s}$ de kans is op een juist antwoord gegeven dat de toetsscore gelijk is aan s. Het rechterlid van (4.44) is dus gelijk aan het rechterlid van (4.45), en deze gelijkheid geldt, ongeacht welke scores in de steekproef zijn geobserveerd. Daarom moet de gelijkheid ook term per term gelden, en we krijgen het belangrijke resultaat

$$\pi_{i/s} = \frac{\varepsilon_i \gamma^{(i)}_{s-1}(\boldsymbol{\varepsilon})}{\gamma_s(\boldsymbol{\varepsilon})}. \tag{4.46}$$

De oplossing van het stelsel (4.44) moet successief benaderd worden. Het zoeken van de oplossing is rekenintensief omdat veelvuldig de γ-functies moeten worden berekend. Een bijkomend probleem hierbij is dat bij het berekenen van die γ-functies, althans indien men er bepaalde algoritmen voor gebruikt, de resultaten zeer onnauwkeurig kunnen worden als gevolg van afrondingen. Om deze onnauwkeurigheden te vermijden, dient men algoritmen te gebruiken die nog meer tijd vergen. Deze omstandigheid brengt sommige auteurs er toe CML als schattingsmethode af te raden of zelfs af te wijzen (bijvoorbeeld Wainer & Mislevy, 1990, p. 80). Er is echter aangetoond dat met

een bepaalde berekeningsmethode van de symmetrische basisfuncties zeer nauwkeurige resultaten verkregen worden: bij k=5000 zijn slechts de laatste vier cijfers van het resultaat aangetast door afrondingsfouten (Verhelst, Glas & Van der Sluis, 1984). In gewone praktijktoepassingen waarbij k zelden groter is dan 100 is het verlies in de regel niet groter dan twee decimalen. In het computerprogramma OPLM (Verhelst, Glas & Verstralen, 1993) waar deze nauwkeurige methode is geïmplementeerd wordt gerekend met een nauwkeurigheid van ongeveer 14 decimalen, zodat van de berekende γ-functies de eerste 12 cijfers zeker correct zijn. Bovendien zijn de moderne computers zo snel dat het oplossen van (4.44) voor k=100 maar enkele minuten duurt. Praktische bezwaren tegen het gebruik van de CML-methode kunnen dus als volkomen achterhaald worden beschouwd. Voor technische details over het berekenen van de γ-functies en het oplossen van (4.44), zie Fischer (1974, hoofdstuk 14), Verhelst, Glas en van der Sluis (1984), Verhelst en Veldhuijzen (1991) en Verhelst, Glas en Verstralen (1993).

Met betrekking tot de statistische nauwkeurigheid van de schatters, moet het begrip informatie dat in paragraaf 4.2.1. werd besproken, uitgebreid worden tot het geval van meer parameters, waar men spreekt van een informatiematrix. Bij een model met k parameters is de informatiematrix een $k \times k$ symmetrische matrix $I(\beta)$, waarvan de cel *(i,j)* gegeven is door minus de verwachte waarde van de tweede partiële afgeleide van de log-aannemelijk- heidsfunctie naar de i-de en de j-de parameter. Voor de conditionele aannemelijkheidsfunctie (4.41) is dit dus

$$I_{ij}(\beta) = -\mathscr{E}\left[\frac{\partial^2 \ln L(\beta;\boldsymbol{X}\,|\,\boldsymbol{s})}{\partial\beta_i\,\partial\beta_j}\right]. \tag{4.47}$$

Toegepast op het Raschmodel geeft dit

$$I_{ij}(\beta) = \begin{cases} \sum_v [\pi_{i|s_v}(1-\pi_{i|s_v})] & \text{indien } i=j, \\ \sum_v [\pi_{ij|s_v} - \pi_{i|s_v}\pi_{j|s_v}] & \text{indien } i\neq j, \end{cases} \tag{4.48}$$

waarin

$$\pi_{ij|s_v} = P(X_{vi}=1, X_{vj}=1\,|\,s_v) = \frac{\varepsilon_i\varepsilon_j\gamma^{(i,j)}_{s_v-2}(\boldsymbol{\varepsilon})}{\gamma_{s_v}(\boldsymbol{\varepsilon})}. \tag{4.49}$$

In (4.49) betekent *(i,j)* in superscript dat zowel ε_i als ε_j uit de argumentvector $\boldsymbol{\varepsilon}$ zijn weggelaten. De afleiding van (4.48) gebeurt geheel analoog aan de afleiding van (4.44). Details hierover zijn te vinden in Fischer (1974, p. 235 e.v.). De multivariate versie van het resultaat dat in 4.2.1. vermeld werd, luidt dan:

> De schatters van de *k*-1 vrije parameters zijn asymptotisch normaal verdeeld met als gemiddelde de werkelijke waarden van de parameters en de inverse van de informatiematrix als variantie-covariantie-matrix.

Net als in het univariate geval worden de itemparameters in (4.48) vervangen door hun ML-schattingen. De standaardfout (*SE*) van de itemparameterschatters is dan gegeven door de vierkantswortel van de elementen op de hoofddiagonaal van de inverse van $I(\boldsymbol{\beta})$.

In verband met de standaardfouten dient men zich te hoeden voor een veel voorkomende fout. Meestal wordt bij het rapporteren van de schattingen van de itemparameters, een standaardfout vermeld bij elk item. Dit betekent dat men een standaardfout krijgt voor *k* parameters, terwijl het model slechts *k*-1 vrije itemparameters heeft. Het antwoord op deze schijnbare paradox is dat de standaardfouten afhankelijk zijn van de gekozen normalisatie. Indien men bijvoorbeeld kiest $\beta_1 = 0$, dan is β_1 een constante en heeft per definitie een standaardfout van 0. De andere schattingen zullen een standaardfout opleveren die verschilt van 0. Gaan we nu over op een andere normalisatie, bijvoorbeeld met $\beta_2 = 0$, dan vinden we de nieuwe schattingen door van de eerste de oorspronkelijke schatting van β_2 af te trekken. Duiden we de nieuwe schattingen aan met $\hat{\tau}$, dan zijn de nieuwe schattingen en hun varianties gegeven in tabel 4.1

Tabel 4.1

Effecten van de normalisatie op schattingen en hun variantie

item	schatting bij $\beta_1 = 0$	schatting bij $\beta_2 = 0$	variantie bij $\beta_2 = 0$
1	0	$\hat{\tau}_1 = -\hat{\beta}_2$	$\text{var}(\hat{\tau}_1) = \text{var}(\hat{\beta}_2)$
2	$\hat{\beta}_2$	0	0
i (>2)	$\hat{\beta}_i$	$\hat{\tau}_i = \hat{\beta}_i - \hat{\beta}_2$	$\text{var}(\hat{\tau}_i) = \text{var}(\hat{\beta}_i) + \text{var}(\hat{\beta}_2) - 2\,\text{cov}(\hat{\beta}_i,\hat{\beta}_2)$

Bij de veel gebruikte normalisatie waarbij de som van de schattingen gelijk is aan nul, beschouwt men *k* functies van de oorspronkelijke *k*-1 vrije parameters. Stel dat weerom de oorspronkelijke normalisatie gekozen was met $\beta_1 = 0$, dan zijn de *k* functies $\hat{\delta}_i$ waarvoor geldt dat $\sum_{i=1}^{k} \hat{\delta}_i = 0$ gegeven door

$$\hat{\delta}_i = \hat{\beta}_i - \frac{1}{k}\sum_{j=1}^{k}\hat{\beta}_j \tag{4.50}$$

en hun variantie is

$$\begin{aligned}\mathrm{var}(\hat{\delta}_i) &= \frac{(k-1)^2}{k^2}\mathrm{var}(\hat{\beta}_i) + \frac{1}{k^2}\sum_{j\neq i}\mathrm{var}(\hat{\beta}_j) \\ &\quad - \frac{2(k-1)}{k^2}\sum_{j\neq i}\mathrm{cov}(\hat{\beta}_i,\hat{\beta}_j) + \frac{1}{k^2}\sum_{j\neq i}\sum_{\substack{m\neq i\\ m\neq j}}\mathrm{cov}(\hat{\beta}_j,\hat{\beta}_m),\end{aligned} \tag{4.51}$$

waarbij $\mathrm{var}(\hat{\beta}_1) = \mathrm{cov}(\hat{\beta}_1,\hat{\beta}_i) = 0,\ (i\neq 1)$.

Het is instructief de CML-methode nog eens op een andere manier te bekijken. Voor een antwoordpatroon $\boldsymbol{x}$ met score s geldt

$$L(\beta,\theta;\boldsymbol{x},s) = P(\boldsymbol{x}|s)P(s|\theta). \tag{4.52}$$

De eerste factor in het rechterlid van (4.52) is de conditionele aannemelijkheidsfunctie gegeven door (4.40) en is onafhankelijk van θ. De tweede factor is de som van de kansen voor alle antwoordpatronen die score s opleveren, en is dus gegeven door

$$P(s|\theta) = \frac{\gamma_s(\boldsymbol{\varepsilon})\exp(s\theta)}{\prod_{i=1}^{k}[1+\varepsilon_i\exp(\theta)]}. \tag{4.53}$$

Deze kans is overduidelijk afhankelijk van θ maar ook van de itemparameters. Bij toepassing van CML wordt alleen de eerste factor in (4.52) gebruikt; de tweede factor wordt 'weggegooid'. Het lijkt er dus op dat door die tweede factor niet mee te nemen, informatie over de itemparameters, die in de score bevat is, wordt verwaarloosd, waardoor minder nauwkeurige schattingen van de itemparameters verkregen worden. Andersen (1970) heeft echter aangetoond dat dit niet zo is. De CML-methode gebruikt dus alle informatie over de itemparameters die in de gegevens aanwezig is.

Tot hiertoe is nog niets gezegd over de manier waarop de getoetste personen uit de populatie getrokken dienen te worden. Dit is met opzet gebeurd. Er is niet stilzwijgend verondersteld dat de steekproef een aselecte trekking moet zijn uit de populatie. Integendeel, door gebruik te maken van de CML-methode maakt het in principe niets uit hoe de steekproef uit de populatie is getrokken. Immers de CML-methode wordt gebruikt om iets te kunnen zeggen over de itemparameters en niet over de populatie

van personen. Bij de derde schat- tingsmethode, die in de volgende subparagraaf wordt besproken, hebben we dit voordeel niet. Dit voordeel van de CML-methode wordt vaak steekproefonafhankelijkheid genoemd. Als hierboven gezegd werd dat het 'in principe' niets uitmaakt hoe de steekproef wordt getrokken, wordt daarmee bedoeld dat CML niet in alle omstandigheden goed werkt. Als we bijvoorbeeld de gegevens inspecteren voor de analyse, en we gooien alle personen die item twee fout hadden uit de steekproef, dan zal de CML-methode geen consistente schatters van de itemparameters opleveren. Wanneer het precies wel en niet goed gaat, wordt gedetailleerd uiteengezet in hoofdstuk 6. Een tweede kanttekening die bij de notie van steekproefonafhankelijkheid gemaakt moet worden betreft de nauwkeurigheid van de parameterschattingen. Twee steekproeven van dezelfde omvang leveren niet noodzakelijkerwijze even nauwkeurige schattingen van de parameters op. In paragraaf 4.2.5 wordt hierop teruggekomen.

4.2.4 MML-schatting in het Raschmodel

Een tweede methode om de individuele θ-parameters kwijt te raken bestaat eruit ze een andere status te geven. De status van de θ-waarden is het standpunt van waaruit men de gegevens beschouwt. Tot nog toe hebben we eigenlijk impliciet aangenomen dat, als Jan en Piet tot de steekproef behoren, we ter zelfder tijd geïnteresseerd zijn in de waarde van de itemparameters en in de θ-waarde van Jan en Piet en van alle andere personen die tot de steekproef behoren. Een ander standpunt is dat het ons eigenlijk niet kan schelen wie er in de steekproef zit, omdat we alleen maar geïnteresseerd zijn in de itemparameters. Dit impliceert dat we de steekproef als een aselecte steekproef uit een of andere populatie beschouwen, en dat we de gedragingen van die toevallige steekproef willen gebruiken om de itemparameters te schatten. Dit standpunt biedt de mogelijkheid om θ kwijt te raken op de volgende manier.

Veronderstel dat θ slechts drie verschillende waarden kan aannemen in de populatie, namelijk -1, 0 en 1, en veronderstel dat deze waarden in de populatie voorkomen met een proportie van respectievelijk .25, .35 en .40. We beschouwen nu de kans dat we het ant-woordpatroon $\boldsymbol{x}$ = (1 0 1) observeren bij aselecte trekking van een persoon uit de populatie. Deze kans is gegeven door

$$P(\boldsymbol{x}) = 0.25 \times P(\boldsymbol{x}|\theta = -1) + 0.35 \times P(\boldsymbol{x}|\theta = 0) + 0.40 \times P(\boldsymbol{x}|\theta = 1).$$

Dat wil zeggen, als we θ niet kennen, kunnen we alle conditionele kansen $P(\boldsymbol{x}|\theta)$ als het ware gaan middelen door te vermenigvuldigen met de kans dat die θ optreedt, en die gewogen conditionele kansen op te tellen. Het resultaat noemt men marginale kans. Vandaar de eerste M in MML. Laten we dit nu veralgemenen tot de situatie waarin het aantal verschillende waarden dat θ kan aannemen gelijk is aan W:

$$P(\boldsymbol{x}) = \sum_{j=1}^{W} P(\boldsymbol{x}|\theta_j)\, P(\theta_j). \tag{4.54}$$

Het gebruik van (4.54) zonder meer is niet erg aantrekkelijk, omdat we dan een waarde voor W moeten kennen, de verschillende waarden die θ kan aannemen en de kansen $P(\theta_j)$. Als we die niet kennen, moeten we ze ook uit de data schatten, zodat er naast de itemparameters nog eens $2W$ parameters bijkomen: W waarden van θ, W-1 vrije kansen $P(\theta_j)$ en W zelf. Boven- dien is W discreet, en kan bijgevolg niet geschat worden met de standaardmethodes die in paragraaf 4.2.1 zijn uiteengezet. Het gebruik van het rechterlid van (4.54) als aannemelijk- heidsfunctie brengt dan ook enkele moeilijke problemen met zich mee. Voor enkele interessante resultaten bij deze benadering, zie De Leeuw en Verhelst (1986), Follman (1988) en Lindsay, Clifford en Grego (1991).

Hoe paradoxaal het ook klinkt, het probleem wordt veel eenvoudiger door θ oneindig veel waarden te laten aannemen, en nog sterker: door θ continu te laten worden, en een bepaalde regel te veronderstellen waaruit de 'kans' op een bepaalde θ uit θ zelf bepaald kan worden. We mogen bij continue variabelen niet meer spreken van kans; men spreekt van dichtheid. Die dichtheid duiden we aan met het functiesymbool g. We kennen een heel populaire dichtheid, namelijk die van de normale verdeling:

$$g(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right], \tag{4.55}$$

waarin $\pi = 3.14159...$ We zien dat in die functieregel twee parameters voorkomen, namelijk μ en σ^2, het gemiddelde en de variantie van θ. De marginale kans van antwoordpatroon $\boldsymbol{x}$ in het geval we een normale verdeling veronderstellen van θ, is gegeven door

$$\begin{aligned} P(\boldsymbol{x}) &= \int_{-\infty}^{+\infty} P(\boldsymbol{x}|\theta)\, g(\theta)\, d\theta \\ &= \int_{-\infty}^{+\infty} P(\boldsymbol{x}|\theta)\, \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right] d\theta. \end{aligned} \tag{4.56}$$

Formule (4.56) is niet meer afhankelijk van θ, want die is er uitgeïntegreerd, wel van de itemparameters en van de twee verdelingsparameters μ en σ^2. Indien we deze marginale kans nu beschouwen als functie van die parameters, dan krijgen we de marginale aannemelijkheidsfunctie voor het antwoordpatroon $\boldsymbol{x}$. De aannemelijkheidsfunctie voor alle geobserveerde antwoordpatronen samen is dan gegeven door

$$L(\beta,\mu,\sigma^2;\boldsymbol{X}) = \prod_{v=1}^{n} \int_{-\infty}^{+\infty} P(\boldsymbol{x}_v|\theta) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right] d\theta . \tag{4.57}$$

Nemen we hiervan de logaritme,

$$\ln L(\beta,\mu,\sigma^2;\boldsymbol{X}) = \sum_{v=1}^{n} \ln \int_{-\infty}^{+\infty} P(\boldsymbol{x}_v|\theta) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right] d\theta , \tag{4.58}$$

dan stuiten we op de moeilijkheid dat we de logaritme van een integraal moeten nemen. Zulke uitdrukkingen laten zich in de regel niet vereenvoudigen, tenzij men een expliciete uitdrukking kan vinden voor de integraal, dat wil zeggen een uitdrukking zonder integraal- teken. Niemand echter heeft zo'n expliciete uitdrukking gevonden, en waarschijnlijk bestaat die zelfs niet. De uitdrukking in het rechterlid van (4.58) kan dan ook niet teruggebracht worden tot de standaarduitdrukking voor de exponentiële familie, en er kan dus geen beroep gedaan worden op de eigenschappen van de exponentiële familie. Het vinden van het maximum van (4.58) is dan ook geen eenvoudige aangelegenheid. Op de verdere details van dit probleem gaan we niet in. Er zijn verschillende computerprogramma's in de handel die MML-schattingen berekenen, en ook de bijbehorende standaardfouten. Bijvoorbeeld BILOG (Mislevy & Bock, 1986), MULTILOG (Thissen, 1988) en het reeds eerder vermelde OPLM. In de statistiek is bewezen (Kiefer & Wolfowitz, 1956) dat door deze methode consistente schattingen van alle parameters worden verkregen.

We sluiten deze paragraaf af met een korte vergelijking van de CML- en de MML-methode. Het belangrijkste verschil tussen beide methodes bestaat erin dat bij CML geen enkele veronderstelling wordt gemaakt over de verdeling van θ in de populatie, terwijl dat bij MML wel wordt gedaan. Het is bij MML helemaal niet noodzakelijk een normale verdeling te veronderstellen. Men zou ook een andere verdeling kunnen aannemen, zie bijvoorbeeld Andersen en Madsen (1977). Belangrijk is echter in te zien dat de veronderstelling over de verdeling nu deel gaat uitmaken van het model. Dus

als we MML toepassen, dan vermengen we als het ware twee modellen: het Raschmodel dat iets vertelt over de antwoorden gegeven θ, en de normale verdeling die vertelt hoe de θ's in de populatie zijn verdeeld. De verstrengeling van beide modellen gebeurt op een heel diep niveau (zie formule (4.56)), zodanig dat beide onderdelen niet eenvoudig uit elkaar zijn te halen. Maken we een fout in de veronderstelling over de normale verdeling, hetzij omdat θ niet normaal verdeeld is, hetzij omdat de steekproef niet aselect uit de normale verdeling is getrokken, dan heeft dat als gevolg dat er ook systematische fouten geïntroduceerd worden in de schatting van de itemparameters. Een gebruiker die MML gebruikt stelt zich dus iets kwetsbaarder op.

Het voordeel van MML is wel dat de verdelingsparameters gelijktijdig met de itemparameters geschat kunnen worden. Indien men in beide geïnteresseerd is, is MML de meest efficiënte methode. In paragraaf 4.4 en uitvoeriger in hoofdstuk 6, waar onvolledige designs worden besproken, zullen we zien dat in sommige omstandigheden CML helemaal niet kan toegepast worden, maar MML wel.

4.2.5 Een voorbeeld

Een goede manier om een indruk te krijgen van de eigenschappen van schattingen is het analyseren van artificiële of gesimuleerde data. Immers, indien we reële data analyseren, weten we nooit of aan de veronderstellingen van het model is voldaan, en bovendien kennen we de echte waarden van de parameters niet. Artificiële data zijn afkomstig van een computerprogramma dat geïnstrueerd kan worden zich volgens het model te gedragen. Essentieel daarbij is dat er een programma voorhanden is dat een aselecte trekking uit een bepaalde verdeling kan uitvoeren. Zulke programma's bestaan en zijn uitvoerig in de statistische literatuur beschreven.

Stel dat we een antwoordpatroon willen genereren van een artificieel persoon die aselect uit de standaardnormale verdeling is getrokken. De toets bestaat uit k=3 items die aan het Raschmodel voldoen en parameterwaarden hebben van respectievelijk -1, 0 en 1. Het pro- gramma start met het trekken van een θ-waarde uit de standaardnormale verdeling. Neem aan dat θ = 0.2. Dan kan berekend worden met behulp van (4.5) dat

$$f_1(0.2) = 0.769, \quad f_2(0.2) = 0.550, \quad f_3(0.2) = 0.310.$$

Vervolgens wordt uit de uniforme verdeling op het interval (0,1) een toevalsgetal p_1 getrokken. Voor de toevalsvariabele p_1 geldt dus dat

$$P(p_1 \leq x) = x, \quad (0 < x \leq 1)$$

en dus $P(p_1 \leq 0.769) = 0.769$. Indien $p_1 \leq 0.769$, krijgt de toevalsvariabele X_1, het antwoord op item 1, de waarde 1, anders 0. Deze procedure wordt herhaald voor elk item, waarbij voor elk item i dus een nieuw en onafhankelijk toevalsgetal p_i uit de uniforme verdeling wordt getrokken. Voor de getrokken waarde van θ is de antwoordregel dus gegeven door

$$X_i = \begin{cases} 1 \text{ indien } p_i \leq f_i(\theta), \\ 0 \text{ indien } p_i > f_i(\theta). \end{cases}$$

De hele hierboven beschreven procedure wordt herhaald voor elk van de n artificiële personen.

In tabel 4.2 staan de resultaten van een analyse op artificiële data, met $n = 500$ personen aselect getrokken uit de standaardnormale verdeling. Het aantal items is acht en de itemparameters zijn -2, -1.5, -1, -0.5, 0.5, 1, 1.5 en 2.

Tabel 4.2
Parameterschattingen uit artificiële data

	CML met $\Sigma_i \hat{\delta}_i = 0$		CML met $\hat{\beta}_1 = 0$		CML met $\hat{\tau}_2 = 0$		MML met $\Sigma_i \hat{\delta}_i = 0$	
β_i	$\hat{\delta}_i$	SE($\hat{\delta}_i$)	$\hat{\beta}_i$	SE($\hat{\beta}_i$)	$\hat{\tau}_2$	SE($\hat{\tau}_2$)	$\hat{\delta}_i$	SE($\hat{\delta}_i$)
-2.	-2.239	0.133	0	---	-0.724	0.181	-2.264	0.135
-1.5	-1.515	0.111	0.724	0.181	0	---	-1.511	0.113
-1.	-1.073	0.103	1.166	0.177	0.441	0.158	-1.063	0.104
-0.5	-0.283	0.096	1.956	0.175	1.231	0.154	-0.273	0.095
0.5	0.609	0.098	2.848	0.180	2.123	0.159	0.615	0.097
1.	0.924	0.101	3.163	0.183	2.439	0.162	0.930	0.101
1.5	1.560	0.113	3.799	0.193	3.075	0.174	1.561	0.113
2.	2.018	0.128	4.257	0.205	3.533	0.187	2.004	0.125

Voor de werkelijke parameters wordt het symbool β gebruikt, voor de CML-schattingen en de MML-schattingen waarvoor de som van de schattingen gelijk is aan 0, wordt het symbool $\hat{\delta}_i$ gebruikt. Voor de CML-schattingen waarbij de parameterschatting van het eerste item gelijk gesteld is aan 0 gebruiken we het symbool $\hat{\beta}_i$ en voor de schattingen waar de parameter van het tweede item gelijkgesteld is aan 0 wordt $\hat{\tau}_i$ gebruikt. Dit

is in overeenstemming met de notatie die gebruikt is in paragraaf 4.2.3 bij de discussie over de standaardfouten. Uit tabel 4.2 zijn enkele interessante bevindingen af te leiden.

Voor de CML-schattingen en de MML-schattingen met dezelfde normering krijgen we ongeveer dezelfde uitkomsten. In alle gevallen ligt de ware parameter binnen het 95%- betrouwbaarheidsinterval rond de geschatte waarde. Ook de geschatte standaardfouten zijn ongeveer aan elkaar gelijk. Indien men de nauwkeurigheid van de schattingen onvoldoende vindt, dan kan de nauwkeurigheid opgevoerd worden door de steekproef groter te maken. Uit (4.48) volgt dat elke persoon een eigen onafhankelijke bijdrage heeft aan de informatiematrix. Nemen we de steekproef dubbel zo groot, dan verdubbelt ook de informatie, en de variantie van de schatters wordt gehalveerd. De standaardfout neemt dus af met een factor $\sqrt{2}$. Wil men de standaardfouten halveren, dan dient men dus een steekproef te nemen die vier maal zo groot is. Dit geldt zowel voor MML als voor CML.

De drie gerapporteerde CML-schattingen verschillen slechts een constante van elkaar, zoals kan afgeleid worden uit tabel 4.2. Normeren door één parameter gelijk aan 0 of een andere constante te stellen, resulteert in veel grotere standaardfouten voor de andere parameters dan in het geval dat de som van de schattingen gelijk wordt gesteld aan 0. Als voorbeeld berekenen we de correlatie tussen $\hat{\beta}_2$ en $\hat{\beta}_3$. Passen we de formule rechtsonder in tabel 4.1 toe voor *i*=3, dan vinden we

$$0.158^2 = 0.177^2 + 0.181^2 - 2\mathrm{cov}(\hat{\beta}_2, \hat{\beta}_3)$$

waaruit volgt dat $\mathrm{cov}(1\hat{\beta}_2, \hat{\beta}_{31}) = 0.01956$. De correlatie tussen de schatters bedraagt dus

$$\mathrm{corr}(\hat{\beta}_2, \hat{\beta}_3) = \frac{0.01956}{0.181 \times 0.177} = 0.611.$$

De hoogte van de correlatie is niet afhankelijk van de steekproefgrootte, noch van het aantal items, maar wel van de informatie die verkregen wordt over het item waarop genormeerd wordt, dit is het desbetreffende element op de hoofddiagonaal van de informatiematrix; zie formule (4.48). Is deze informatie relatief laag, dan zal de correlatie hoger uitvallen dan wanneer die informatie relatief groot is. Dit wordt geïllustreerd in tabel 4.3. De resultaten in de tweede en derde kolom van deze tabel hebben betrekking op dezelfde gegevens als tabel 4.2; in de vierde en vijfde kolom staan de resultaten op een onafhankelijke steekproef ter grootte van 5000, maar met dezelfde itemparameters als in het eerste voorbeeld.
De twee steekproefgroottes leveren nagenoeg dezelfde schatting van de correlaties op. Merk op dat de getallen in de vierde kolom van tabel 4.3 ongeveer tien maal zo groot

zijn als de getallen in de tweede kolom: de informatie neemt evenredig toe met het aantal observaties. De verhouding is niet exact 10, omdat de kolommen alleen schattingen van de informatie bevatten.

Tabel 4.3

Correlatie tussen CML-schatters als functie van het item waarop genormeerd wordt

	n=500		n=5000	
item	informatie	$\text{corr}(\hat{\beta}_2,\hat{\beta}_3)$	informatie	$\text{corr}(\hat{\beta}_2,\hat{\beta}_3)$
1	47.3	.611	490	.588
4	84.4	.452	836	.450
5	86.6	.461	828	.472
6	77.8	.499	749	.507
7	61.0	.569	633	.557
8	52.7	.607	505	.615

De standaardfouten in tabel 4.2 zijn niet voor alle items even groot. Dit hangt eveneens samen met de informatie die de gegevens over het item opleveren. De iteminformatiefunctie wordt gegeven door de elementen op de diagonaal van de informatiematrix, zie (4.48). In veel toepassingen wordt alleen van deze elementen gebruik gemaakt om een schatting van de standaardfout te maken:

$$SE^{*}(\hat{\beta}_i) = [I_{ii}(\hat{\beta})]^{-\frac{1}{2}} = \left[\sum_v \hat{\pi}_{i|s_v}(1-\hat{\pi}_{i|s_v})\right]^{-\frac{1}{2}}. \qquad (4.59)$$

De uitkomst van (4.59) kan dus om drie redenen van de echte standaardfout verschillen. Ten eerste, in het rechterlid worden niet de echte conditionele kansen $\pi_{i|s}$, maar schattingen ingevuld. Ten tweede wordt een asymptotisch resultaat toegepast op een eindige steekproef en ten derde worden de buitendiagonale elementen van de informatiematrix verwaarloosd. Toch wordt (4.59) in de praktijk vaak gebruikt, en soms niet terecht zoals we verderop zullen zien. De eenvoudige structuur van (4.59) laat ons echter twee zaken duidelijk zien. In de eerste plaats het effect van de steekproefgrootte op de nauwkeurigheid van de schattingen. Elke antwoordpatroon in de steekproef levert precies één term aan de som in het rechterlid van (4.59). Als we uit een bepaalde populatie twee aselecte steekproeven trekken, de eerste van *n* personen, en de tweede van 2*n* personen, dan zal de som in het rechterlid van (4.59) voor de tweede steekproef ongeveer twee keer zo groot zijn als voor de eerste steekproef, en dus zullen de standaardfouten van de itemparemeterschattingen in de eerste steekproef ongeveer $\sqrt{2}$

zo groot zijn als in de tweede steekproef. In de tweede plaats kunnen we (4.59) gebruiken om te laten zien dat we voorzichtig moeten omspringen met het theoretische voordeel van de steekproefonafhankelijkheid. De maximale waarde van het produkt $\pi_{i|s}(1-\pi_{i|s})$ bedraagt 0.25 en wordt bereikt indien $\pi_{i|s}$=0.5. Indien de score 0 is of *k*, is de bijdrage aan de informatie precies 0. Stel nu dat we aan een steekproef van n personen een toets voorleggen die veel te moeilijk is, zodat relatief veel personen een score 0 behalen. De antwoordpatronen van deze personen dragen dus niets bij aan de iteminformatie, en de standaardfouten van de parameterschattingen zullen groter zijn dan in het geval van een even grote steekproef waarbij de moeilijkheidsgraad van de items goed overeenkomt met de vaardigheid van de personen. Het voordeel van steekproefonafhankelijkheid moet dus niet gebruikt worden om een toets voor te leggen aan een willekeurige verzameling personen. In hoofdstuk 7 zullen we twee voorbeelden zien waarbij op een verstandige manier voordeel is gehaald uit de steekproefonafhankelijkheid van de CML-schatters.

De reden waarom (4.59) in de praktijk vaak gebruikt wordt, is dat het uitrekenen en inverteren van de hele informatiematrix erg tijdrovend wordt indien het aantal items groot is. Formule (4.59) kan ook gebruikt worden voor het item waarop genormeerd is. Gaat men naderhand de oplossing centreren, dan wordt ook dezelfde formule gebruikt om de standaardfout van de *k* parameters te berekenen. Standaardfouten kunnen dus op veel verschillende manieren geschat worden, en de resultaten kunnen nogal uiteenlopen. Dit kunnen we zien door een speciaal geval te bestuderen. Veronderstel dat de parameters van de *k* items in een toets allemaal aan elkaar gelijk zijn. De informatie die we over elk item inwinnen zal dus ook dezelfde zijn voor alle items. De elementen op de hoofddiagonaal van de informatiematrix zullen dus ook aan elkaar gelijk zijn. Veronderstel dat die informatie gelijk is aan c^2. De waarde van c^2 is afhankelijk van de grootte van de steekproef en van de moeilijkheid van de items in vergelijking met de gemiddelde vaardigheid. In tabel 4.4 worden de asymptotische standaardfouten, berekend uit de inverse van de informatiematrix, en hun schattingen, gebaseerd op formule (4.59), gegeven.

Tabel 4.4
Standaardfout bij *k* items van dezelfde moeilijkheid

normering	*SE*	formule (4.59)
op één item	$\frac{1}{c}\sqrt{\frac{2(k-1)}{k}}$	$\frac{1}{c}$

gecentreerd	$\frac{1}{c}\frac{k-1}{k}$	$\frac{1}{c}$

Voor de theoretische afleiding van dit resultaat, verwijzen we naar Verhelst (1993). Voor de gecentreerde oplossing is (4.59) dus een goede benadering indien het aantal items niet te klein is. Merk ook op dat (4.59) systematisch de standaardfout overschat. Kiezen we echter een oplossing waarbij op één item genormeerd is, dan geeft (4.59) een grove onderschatting van de standaardfout: het effect van de verwaarlozing van de buitendiagonale elementen van de informatiematrix komt dus ongeveer overeen met het overwaarderen van de steekproefgrootte met een factor 2. De gecentreerde oplossing verdient dus de voorkeur. Tenslotte is het gemakkelijk te controleren dat de correcte standaardfout bij een gecentreerde oplossing kleiner is dan bij normering op één item.

4.3 Het toetsen van het Raschmodel

In paragraaf 4.2.1 hebben we de grootste-aannemelijkheidsschatter voor de parameter van een muntstuk opgesteld. Daar bleek dat we enkel het relatief aantal successen hoefden te kennen om die parameter te schatten. De observaties kunnen ons niet méér informatie opleveren. Indien we zeker zouden zijn dat aan de veronderstellingen van het model was voldaan, dan hoefden we ook niets meer te weten. Bij het opgooien van een muntstuk zijn die veronderstellingen eenvoudig: de kans op succes moet onveranderd blijven en de uitkomst bij elke worp moet onafhankelijk zijn van de uitkomsten van de andere worpen. Veronderstel nu dat het opgooien zo klungelig gebeurt dat de uitkomst bij een bepaalde worp bijna zeker gelijk is aan de uitkomst van de vorige worp, bijvoorbeeld omdat het muntstuk maar een heel klein beetje wordt opgetild en dan weer losgelaten. Als die afhankelijkheid heel sterk is, is het mogelijk dat bij 100 keer opgooien het muntstuk 99 keer op munt valt, ook al is het niet vervalst. We kunnen dan nog wel de techniek van het schatten gaan toepassen, doch de conclusie dat het muntstuk onzuiver is, is niet terecht, omdat niet voldaan is aan de veronderstellingen van het model. Om na te gaan of aan de veronderstellingen van het model is voldaan, kunnen we natuurlijk de experimentele procedure aan een nader onderzoek onderwerpen. Indien het muntstukexperiment is uitgevoerd zoals hierboven beschreven, zullen we niet geneigd zijn de resultaten serieus te nemen. Indien bij de dataverzameling van toetsgegevens de afname niet serieus gebeurt, bijvoorbeeld omdat de leerlingen alle gelegenheid krijgen elkaar te consulteren bij het beantwoorden van de items,

kunnen we beter de statistische verwerking achterwege laten, want de belangrijke eis van experimentele onafhankelijkheid is geschonden, en alle conclusies die uit een statistische analyse volgen, berusten op drijfzand. Echter, een zorgvuldige dataverzameling is wel een noodzakelijke, doch geen voldoende voorwaarde opdat alle veronderstellingen van het model vervuld zijn. De reden hiervoor is dat schendingen van het model erg subtiel kunnen zijn. De algemene strategie om modelschendingen te ontdekken is het nauwkeurig onderzoeken van de fijnere structuur van de data, met name die aspecten van de data die niet zijn gebruikt om de parameters te schatten.

We beginnen met een voorbeeld uit het muntexperiment. Indien we 100 keer een zuivere munt opgooien, en we stellen vast dat het muntstuk de eerste 50 keer munt valt, en vervolgens 50 keer kruis, dan zullen we het muntstuk of het experiment niet vertrouwen. We verwachten dat de afwisseling 'k-m' of 'm-k' meer dan één keer optreedt. Observeren we echter een volmaakte regelmaat waarbij k en m elkaar iedere keer weer afwisselen, dan is dit ook verdacht. Het aantal afwisselingen mag dus niet te groot zijn maar ook niet te klein. De statistische theorie wordt gebruikt om precies aan te geven wat bedoeld wordt met te groot of te klein. De toetsingsprocedure voor dit probleem staat beschreven in Siegel en Castellan (1988, p. 58-64).

In het Raschmodel is de ruwe score, het aantal items juist, een voldoende steekproefgrootheid voor de latente variabele θ. In paragraaf 4.5 zullen we zien dat iedereen met dezelfde score ook dezelfde schatting van θ krijgt, ongeacht welke items juist beantwoord zijn. Dit betekent echter niet dat bij de subgroep van personen die dezelfde score hebben alle mogelijke antwoordpatronen even waarschijnlijk zijn. Beschouwen we een simpel voorbeeld. Laat de toets bestaan uit $2k$ items, waarvan k gemakkelijke met itemparameter $\varepsilon_i = 2$, $i = 1,\ldots,k$ en k moeilijke met itemparameter $\varepsilon_i = 0.5$, $i = k+1,\ldots,2k$ (zie formule (4.37)). In de subpopulatie van personen die precies k items juist hebben, is de kans dat de k gemakkelijkste items juist zijn beantwoord, gegeven door

$$P(X_1 = \ldots = X_k = 1\,,\, X_{k+1} = \ldots = X_{2k} = 0 \,|\, s = k) \;=\; \frac{2^k}{\gamma_k(\boldsymbol{\varepsilon})},$$

en de kans dat de k moeilijkste juist zijn is

$$P(X_1 = \ldots = X_k = 0\,,\, X_{k+1} = \ldots = X_{2k} = 1 \,|\, s = k) \;=\; \frac{2^{-k}}{\gamma_k(\boldsymbol{\varepsilon})}\,.$$

De verhouding tussen die twee kansen is 2^{2k}. Bij 10 items en een score van 5 verwachten we dus 1024 keer zoveel respondenten die de vijf makkelijkste items juist hebben als respon- denten met de vijf moeilijkste items juist. Indien we in een

steekproef ongeveer gelijke aantallen zouden vinden, is dat een voldoende reden om de geldigheid van het model in twijfel te trekken. Dit voorbeeld maakt ook duidelijk dat een theorie die geen absolute uitspraken doet over het gedrag wel degelijk gefalsificeerd kan worden. De kans op een juist antwoord in het Raschmodel is altijd strikt groter dan 0 en strikt kleiner dan 1 ongeacht de waarde van θ. Hoewel dus met elke θ-waarde alle antwoordpatronen mogelijk zijn, zijn ze niet allemaal even waarschijnlijk, en deze ongelijke waarschijnlijkheden dienen weerspiegeld te worden in ongelijke relatieve frequenties in de steekproef. De statistische theorie wordt gebruikt om aan te geven hoe nauwkeurig die weerspiegeling dient te zijn.

4.3.1 De veronderstellingen van het Raschmodel

In paragraaf 4.1 is gezegd dat een belangrijke reden om het Raschmodel als meetmodel te kiezen wiskundige elegantie is. Dit is ongetwijfeld waar, maar men kan zich de vraag stellen of er geen andere modellen bestaan die wiskundig even elegant zijn, en toch drastisch van het Raschmodel verschillen. In de literatuur zijn verschillende pogingen ondernomen om het Raschmodel af te leiden uit een aantal eenvoudige aannames. Deze aannames worden ook axioma's genoemd. Het is mogelijk het Raschmodel af te leiden uit verschillende ver- zamelingen van axioma's. Voor een overzicht, zie Fischer (in voorbereiding). Wij zullen één stel aannames bespreken, zonder echter de afleiding aan te tonen, omdat deze wiskundig nogal moeilijk is. Deze aannames zijn:

(1) de itemresponscurve $f_i(\theta)$ is continue en strikt stijgend voor alle waarden van θ en voor alle items *i* in de beschouwde itemverzameling. θ is een unidimensionale grootheid en kan een willekeurige reële waarde aannemen;

(2) voor alle items *i* zijn de limieten (4.7) geldig:

$$\lim_{\theta\to\infty} f_i(\theta) = 1, \qquad \lim_{\theta\to-\infty} f_i(\theta) = 0;$$

(3) het axioma van de lokale stochastische onafhankelijkheid is geldig;

(4) de ruwe score $s = \sum_i x_i$ is een voldoende steekproefgrootheid voor θ.

Er kan mathematisch worden aangetoond dat de vier bovenstaande axioma's equivalent zijn met het Raschmodel. Voor de praktijk betekent dit dat schending van één of meer van die axioma's automatisch een schending is van het Raschmodel.

Het eenvoudigste voorbeeld van een schending wordt wellicht gegeven door het gebruik van meerkeuzevragen: uit axioma (2) volgt dat de kans op een juist antwoord, gegeven dat de vaardigheid zeer klein is ($\theta\to-\infty$), praktisch gelijk moet zijn aan 0. Indien er in zo'n geval geraden wordt tussen bijvoorbeeld vier alternatieven, is de kans op een juist antwoord 0.25. Voor dit soort items geeft het Raschmodel dus geen juiste

beschrijving. In de praktijk betekent dit dus dat raadgedrag een oorzaak kan zijn van de ongeldigheid van het Raschmodel. In paragraaf 7.2 wordt uitvoerig op dit probleem ingegaan.

Een tweede soort inbreuk die voor de praktijk relevant is, wordt gewoonlijk aangeduid met het begrip multidimensionaliteit. In axioma (1) is sprake van een unidimensionale grootheid θ. Neem aan dat θ staat voor numerieke vaardigheid. Veronderstel verder dat de items bestaan uit een aantal redaktiesommen, die een beroep doen zowel op deze numerieke vaardigheid als op verbale vaardigheid. Het is zeer wel mogelijk dat aan axioma (1) voldaan is, doch beschouwen we nu tegelijkertijd axioma (3). Dit axioma impliceert dat, indien θ constant wordt gehouden, de covariantie tussen alle itemantwoorden 0 is. Als numerieke vaardigheid en verbale vaardigheid niet precies hetzelfde betekenen, is het natuurlijk zo dat in een subpopulatie waar θ constant is, er nog variabiliteit in de verbale vaardigheid zal overblijven, en omdat we aangenomen hebben dat het antwoord gedeeltelijk door de verbale vaardigheid wordt bepaald, zal de covariantie tussen de itemantwoorden niet 0 zijn. Samenvattend kunnen we dus stellen dat, indien de items een beroep doen op meerdere vaardigheden die niet perfect correleren, en θ verwijst naar één van die vaardigheden, dan is automatisch het axioma van de lokale stochastische onafhankelijkheid geschonden. Door het hierboven gegeven voorbeeld iets aan te scherpen is ook duidelijk te zien dat het vierde axioma geschonden is. Veronderstel dat de helft van de items uitsluitend een beroep doen op verbale vaardigheid, en de andere helft uitsluitend op numerieke vaardigheid. Veronderstel bovendien dat verbale en numeriek vaardigheid in de populatie zeer laag correleren. Beschouw nu twee personen, A en B, die beide de helft van de items juist beantwoorden: persoon A heeft uitsluitend de verbale items juist en persoon B uitsluitend de numerieke items. Hoewel beide personen dezelfde ruwe score hebben behaald, ligt het voor de hand de numerieke vaardigheid, θ, van persoon B hoger in te schatten dan die van persoon A, doch dit is hetzelfde als het verwerpen van axioma (4).

Uit het voorgaande mag niet worden afgeleid dat rekenitems alleen aan het Raschmodel voldoen, indien ze uitsluitend een beroep doen op numerieke vaardigheid en niet op verbale vaardigheid. IRT-modellen zijn wiskundige modellen die voorspellingen doen over het gedrag van personen die de items beantwoorden. Indien deze voorspellingen juist zijn kan men daaraan het argument ontlenen dat de items in de toets een unidimensionale vaardigheid meten. Of deze vaardigheid een numerieke dan wel een mengsel van numerieke en verbale vaardigheden is, is een kwestie van interpretatie.

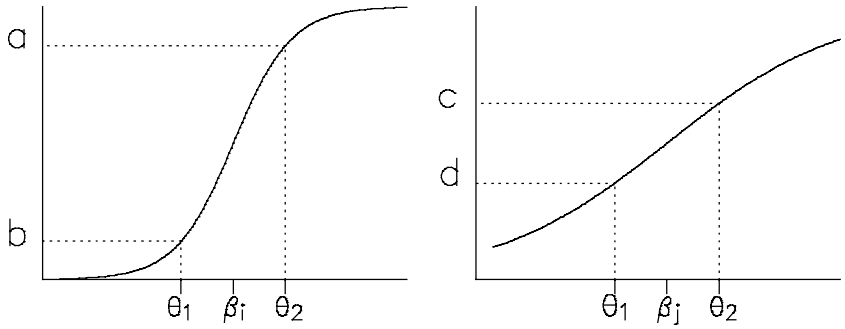
Voor een goed begrip van axioma (4) introduceren we een IRT-model dat algemener is dan het Raschmodel, namelijk het twee parameter logistisch model, dat ook wel

aangeduid wordt als het Birnbaummodel (Birnbaum, 1968). Het onderscheid tussen het Raschmodel en het Birnbaummodel hangt nauw samen met een eigenschap van het Raschmodel die uitgebeeld is in figuur 4.3: de curven van twee itemresponsfuncties snijden elkaar nooit.

Beschouw nu figuur 4.6. Daarin zijn twee itemresponscurves afgebeeld voor de items *i* en *j*. We nemen aan dat $\beta_i = \beta_j$. Beschouw nu twee personen met latente vaardigheden θ_1 en θ_2. Uit de figuur is duidelijk dat

$$f_i(\theta_2) - f_i(\theta_1) = a - b > c - d = f_j(\theta_2) - f_j(\theta_1).$$

Dit betekent dat we op grond van item *i* een beter onderscheid kunnen maken tussen die twee personen dan op grond van item *j*, hoewel beide items even moeilijk zijn. Anders gezegd: item *i* discrimineert beter dan item *j*. Dit betere discriminerend vermogen komt tot uiting in het steilere verloop van de itemresponscurve van item *i*. Merk overigens op dat dit discriminerend vermogen een plaatselijke eigenschap is: twee personen met een verschillende vaardigheid die voor beiden veel groter is dan de moeilijkheidsgraad van het item, zullen beiden bijna zeker het item oplossen en dus kan het item geen onderscheid maken tussen beider vaardigheid. Het discriminerend vermogen van een item wordt dus afgemeten aan de snelheid waarmee de itemresponsfunctie verandert in de buurt van de moeilijkheidsparameter.



Figuur 4.6
Twee items die verschillend discrimineren

Als we binnen de familie van logistische functies blijven, kunnen we dit verschil in discriminerend vermogen uitdrukken door een iets gecompliceerder functievorm te kiezen dan in het Raschmodel. De formule voor functies zoals weergegeven in figuur 4.6 is:

$$P(X_i=1|\theta) = f_i(\theta) = \frac{\exp[a_i(\theta - \beta_i)]}{1 + \exp[a_i(\theta - \beta_i)]}, \quad (a_i>0). \tag{4.60}$$

De grootheid a_i wordt de discriminatieparameter van het item genoemd. Een item wordt dus gekenmerkt door twee parameters: een moeilijkheidsparameter β_i en een (positieve) discriminatieparameter a_i. Merk op dat formule (4.60) onveranderd blijft indien we zowel θ als β_i met een willekeurige positieve constante c vermenigvuldigen en a_i door c delen. Vergelijken we nu (4.60) met (4.5), dan zien we dat in (4.5) het verschil θ-β_i met 1 is vermenigvuldigd. Formule (4.5) is dus een speciaal geval van (4.60), waarbij $a_i = 1$ voor alle items i. Maar omdat we de discriminatieparameters met een willekeurige positieve constante mogen vermenigvuldigen, kunnen we zeggen dat het Raschmodel een speciaal geval is van het Birnbaummodel waarbij alle discriminatieparameters aan elkaar gelijk zijn. In figuur 4.6 heeft item i een grotere discriminatieparameter dan item j.

Voor een antwoordpatroon $\boldsymbol{x}$ is met gebruik van (4.60) gemakkelijk aan te tonen dat de log-aannemelijkheidsfunctie gegeven is door

$$\ln L(\beta,\boldsymbol{a},\theta;\boldsymbol{x}) = \theta\sum_i a_i x_i - \sum_i x_i a_i \beta_i - \sum_i \ln\{1+\exp[a_i(\theta-\beta_i]\}, \qquad (4.61)$$

waaruit duidelijk blijkt dat de gewone somscore geen voldoende steekproefgrootheid is voor θ. In het Birnbaummodel is dus niet voldaan aan axioma (4). De gewogen somscore $\sum_i a_i x_i$ is wel voldoende, doch deze grootheid is een functie van de onbekende discriminatieparameters. Het Birnbaummodel behoort dus ook niet tot de exponentiële familie. Nadere beschouwingen over dit model worden in hoofdstuk 5 gegeven.

Naast de axioma's (1) tot (4) zijn er nog een paar veronderstellingen die strikt genomen geen axioma's zijn, doch die men zou kunnen omschrijven als algemene voorwaarden die vervuld moeten worden om het model te kunnen toepassen en toetsen. De eerste voorwaarde, die reeds ter sprake kwam, is experimentele onafhankelijkheid bij de dataverzameling. Indien niet aan die voorwaarde voldaan is, snijden we onszelf de pas af om iets zinnigs over het model te kunnen zeggen. De tweede voorwaarde heeft te maken met de herhaalbaarheid van de metingen. De axioma's (1) tot (4) worden van toepassing geacht op een persoon die de items van een toets beantwoordt. Om de geldigheid van probabilistische uitspraken te onderzoeken zijn veel waarnemingen nodig, maar die kunnen wegens geheugeneffecten niet allemaal bij dezelfde persoon gedaan worden. We zullen dus de antwoorden van meer personen tegelijkertijd moeten analyseren, doch dit impliceert dat we de geldigheid van de axioma's voor alle personen tegelijkertijd veronderstellen. Deze aanname wordt wel eens aangeduid als de aanname van homogeniteit van de populatie. Het is dus belangrijk bij het toetsen van het model niet alleen met een ja of nee als antwoord te komen, doch eveneens aanwijzingen te

vinden dat het model eventueel geldig is in bepaalde subpopulaties en in andere niet. Op dit aspect wordt nader ingegaan in paragraaf 4.5 over persoonsparameterschattingen en in hoofdstuk 9 over itemonzuiverheid.

4.3.2 Relaties tussen het Raschmodel en het multinomiale model

Om een goed begrip te hebben van de statistische toetsen is het nuttig een zeer algemeen statistisch model te beschouwen, waarvan het Raschmodel een speciaal geval is. Indien bij n personen een toets van k items wordt afgenomen levert elke persoon een bepaald antwoord- patroon op. Bij k items zijn er 2^k mogelijke antwoordpatronen en zonder verlies aan informatie kunnen we de observaties samenvatten in een frequentievector met 2^k elementen, waarbij elk element aangeeft hoe vaak het overeenkomstig antwoordpatroon is geobserveerd. Symbolisch duiden we deze frequentie aan met $np_{\boldsymbol{x}}$, waarbij $p_{\boldsymbol{x}}$ de geobserveerde proportie weergeeft van het antwoordpatroon $\boldsymbol{x}$. $p_{\boldsymbol{x}}$ is dus een realisatie van de toevalsvariabele $P_{\boldsymbol{x}}$. Een statistisch model specificeert voor elk antwoordpatroon de kans dat dit antwoordpatroon optreedt. Zie bijvoorbeeld formule (4.56). Kortheidshalve duiden we deze kans aan met $\pi_{\boldsymbol{x}}$. In formule (4.56) is duidelijk dat deze kans een functie is van de modelparameters ε, μ en σ^2. Duiden we nu op een algemene manier het rijtje parameters aan met de vector φ, dan kunnen we expliciet aangeven dat de theoretische kansen een functie zijn van de modelparameters door te schrijven $\pi_{\boldsymbol{x}}(\varphi)$. Het model wordt dus symbolisch geschreven als een vector van 2^k functies van de modelparameters φ, en de observaties zijn vectoren met 2^k overeen- komstige proporties. Deze vectoren worden geschreven als respectievelijk π en $\boldsymbol{p}$. De aannemelijkheidsfunctie kan dus geschreven worden als

$$L(\varphi;\boldsymbol{p},n) = \frac{n!}{(np_1)!\ldots(np_{2^k})!}\prod_{\boldsymbol{x}}\left[\pi_{\boldsymbol{x}}(\varphi)\right]^{np_{\boldsymbol{x}}}, \tag{4.62}$$

waarbij de breuk in het rechterlid aangeeft op hoeveel verschillende manieren de geobserveerde frequentievector uit n observaties gerealiseerd kan worden. Merk op dat deze grootheid niet afhangt van de parameters, en in de aannemelijkheidsfunctie dus als een constante C behandeld kan worden. Het rechterlid van (4.62) is de kansverdeling van de multinomiale verdeling, waarbij de theoretische kansen een functie zijn van de model- parameters. Deze klasse van verdelingen wordt aangeduid als de geparametriseerde multinomiale verdeling.

De eenvoudigste verdeling uit deze familie ontstaat wanneer de theoretische kansen π zelf de parameters zijn. In dat geval spreekt men kortweg van de multinomiale

verdeling. Merk wel op dat niet alle 2^k parameters vrij kunnen variëren, want hun som moet gelijk zijn aan 1; er zijn dus $2^k - 1$ vrije parameters. Evenzo geldt dat er slechts $2^k - 1$ vrije frequenties zijn, want hun som is gelijk aan n. De logaritme van de aannemelijkheidsfunctie in het multinomiaal model is gegeven door

$$\ln L(\pi; \boldsymbol{p}, n) = \ln C + n \sum_{\boldsymbol{x}} p_{\boldsymbol{x}} \ln \pi_{\boldsymbol{x}}, \tag{4.63}$$

waarin men direct de gedaante van de exponentiële familie herkent, waarbij de proporties $p_{\boldsymbol{x}}$ voldoende steekproefgrootheden zijn. De schattingsvergelijkingen zijn dus gegeven door

$p_{\boldsymbol{x}} = \mathscr{E}(P_{\boldsymbol{x}}) = \pi_{\boldsymbol{x}}$ met als oplossing $\hat{\pi}_{\boldsymbol{x}} = p_{\boldsymbol{x}}$.

In dit multinomiale model worden de observaties dus foutloos voorspeld door het model, een betere voorspelling is niet mogelijk. Daarom wordt dit multinomiale model het verzadigde model genoemd.

Keren we nu terug naar het geparametriseerde multinomiale model waar de theoretische kansen $\pi_{\boldsymbol{x}}$ een functie zijn van de parameters φ. In ons voorbeeld bevat φ $k + 1$ vrije para- meters, en voor $k > 2$ geldt dat $k + 1 < 2^k - 1$. Indien we φ vastleggen, liggen alle 2^k theoretische kansen $\pi_{\boldsymbol{x}}$ vast. In de statistiek drukt men dat als volgt uit. In het multinomiale model is de vector π een rijtje van 2^k getallen dat aan zekere voorwaarden moet voldoen. De verzameling van vectoren die aan deze voorwaarden voldoen wordt de parameterruimte genoemd, en deze verzameling duiden we aan met het symbool Ω. In het multinomiale model geldt dus

$$\Omega = \{(\pi_1,...,\pi_{2^k}) \mid \pi_j \geq 0, (j=1,...,2^k); \textstyle\sum_j \pi_j = 1\}. \tag{4.64}$$

In het geparametriseerde multinomiale model brengen we restricties aan op Ω, door te eisen dat de theoretische kansen welbepaalde functies zijn van de parameters φ, in het voorbeeld gegeven door de functieregel (4.56). Deze beperkte parameterruimte duiden we aan met Ω_φ en de definitie is

$$\Omega_\varphi = \{(\pi_1,...,\pi_{2^k}) \mid \pi_j = \pi_j(\varphi), (j=1,...,2^k); \varepsilon_i > 0, (i=1,...,k); \sigma^2 \geq 0\}. \tag{4.65}$$

Aan de hand van formule (4.56) is gemakkelijk na te gaan dat $\pi_{\boldsymbol{x}} \geq 0$ en dat $\sum_{\boldsymbol{x}} \pi_{\boldsymbol{x}} = 1$. Dus elke vector π die behoort tot Ω_φ behoort eveneens tot Ω, of

$$\Omega_\varphi \subset \Omega. \tag{4.66}$$

Als een tweede voorbeeld beschouwen we de CML-schatting van de itemparameters in het Raschmodel. Voor een willekeurig antwoordpatroon $\boldsymbol{x}$ met score s kunnen we steeds schrijven (zie (4.52)) $P(\boldsymbol{x}) = P(\boldsymbol{x}\,|s)P(s)$, of in een wat compactere notatie

$$\pi_{\boldsymbol{x}} = \pi_{\boldsymbol{x}|s}\,\omega_s, \tag{4.67}$$

waarin $\omega_s = P(s)$. Beschouwen we nu een model waarin de frequentievector van de scores de multinomiale verdeling volgt met parameters ω_s, $(s = 0,\ldots,k)$, en de conditionele kansen gegeven zijn door het rechterlid van (4.40), de conditionele kansen in het Raschmodel, dan zien we dat (4.67) een geparametriseerd multinomiaal model definieert met parametervector $\varphi = (\omega_0,\ldots,\omega_s,\ldots,\omega_k,\varepsilon_1,\ldots,\varepsilon_k)$, waarbij echter niet alle parameters vrij zijn, want één itemparameter kunnen we vrij kiezen, en er moet gelden dat $\Sigma_s\omega_s = 1$. Er zijn dus $2k - 1$ vrije parameters in φ. Glas (1989) heeft aangetoond dat de ML-schatters van de ε-parameters de CML-schatters zijn en dat de schatters van de marginale kansen ω_s gegeven zijn door

$$\hat{\omega}_s = p_s, \quad (s = 0,\ldots,k). \tag{4.68}$$

Door de conditionele aannemelijkheid aan te vullen met een verzadigd model voor de scoreverdeling, construeren we een geparametriseerd multinomiaal model. In de volgende paragrafen wordt de statistische toetsingstheorie behandeld waarbij we vaak een beroep zullen doen op deze multinomiale modellen.

4.3.3 Likelihood-ratio-toetsen

Indien een bepaald niet-verzadigd model juist is, kan men niet verwachten dat bij een eindige dataverzameling het maximum van de aannemelijkheidsfunctie even groot zal zijn als het maximum onder het verzadigde model. Immers, het verzadigde model levert altijd het absolute maximum van de aannemelijkheidsfunctie op, terwijl het beperkte model restricties oplegt aan de multinomiale kansen die in een eindige steekproef niet perfect weerspiegeld hoeven te zijn in de geobserveerde proporties. Er geldt dus altijd

$$\frac{L^*(\varphi;\boldsymbol{p},n)}{L^*(\pi;\boldsymbol{p},n)} \leq 1, \tag{4.69}$$

waarin L^* het maximum van de aannemelijkheidsfunctie aanduidt. Anderzijds verwachten we natuurlijk dat, indien het beperkte model juist is, het maximum van de aannemelijkheidsfunctie niet al te zeer zal afwijken van het absolute maximum. De verhouding aangegeven in het linkerlid van (4.69) moet niet al te zeer afwijken van 1, of haar logaritme moet niet al te ver van 0 afwijken. Meer formeel kunnen we de statistische nulhypothese $H_0 : \pi_{\boldsymbol{x}} \in \Omega_\varphi$ toetsen door de overschrijdingskans van (4.69) te bepalen onder de nulhypothese. Deze toets wordt de likelihood-ratio-toets (LR-toets) genoemd. In de theoretische statistiek wordt aan- getoond dat minus twee maal de logaritme van (4.69), vaak aangeduid als G^2, asymptotisch chi-kwadraat verdeeld is indien de nulhypothese waar is. G^2 is dus gegeven door

$$\begin{aligned} G^2 &= 2[\ln L^*(\pi;\boldsymbol{p},n) - \ln L^*(\varphi;\boldsymbol{p},n)] \\ &= 2n\sum_{\boldsymbol{x}} p_{\boldsymbol{x}} \ln \frac{p_{\boldsymbol{x}}}{\hat{\pi}_{\boldsymbol{x}}}, \end{aligned} \tag{4.70}$$

waarin $\hat{\pi}_{\boldsymbol{x}} = \pi_{\boldsymbol{x}}(\hat{\varphi})$, de functie $\pi_{\boldsymbol{x}}$ geëvalueerd op de ML-schatter van φ. Het aantal vrijheidsgraden is het aantal geschatte parameters in het verzadigde model minus het aantal vrije parameters in het beperkte model. In het geval van MML-schattingen is dit dus $[2^k - 1] - [k + 1] = 2^k - k - 2$; in het geval van CML-schattingen is dit verschil $[2^k - 1] - [2k - 1] = 2^k - 2k$. De uitdrukking dat G^2 asymptotisch chi-kwadraat verdeeld is betekent dat de steekproevenverdeling van G^2 goed door de chi-kwadraatverdeling benaderd wordt als n groot wordt; als n niet zeer groot is kan deze benadering slecht zijn, en het gebruik van de chi-kwadraatverdeling dus onterecht. Het probleem is echter wat er precies bedoeld wordt met groot. Het aantal mogelijke antwoordpatronen stijgt zeer snel met het aantal items. Indien $k = 10$ zijn er meer dan 1000 verschillende antwoordpatronen, doch in het sociaalweten- schappelijk onderzoek in Nederland wordt een steekproef van 1000 personen doorgaans als groot beschouwd. In zo'n situatie zal er meestal een vrij groot aantal antwoordpatronen helemaal niet voorkomen in de steekproef, terwijl voor veel andere antwoordpatronen de geobserveerde frequentie klein zal zijn. Of in zo'n geval de chi-kwadraatverdeling een goede benadering is van de verdeling van G^2 is een vraagstuk waar nog veel discussie over is (zie bijv. Read & Cressie, 1988). De schijnbaar voor de hand liggende oplossing om de steekproef dan maar veel groter te maken, heeft echter naast het kostenaspect nog een ander nadeel. Door de steekproefomvang te laten toenemen vergroot ook het onderscheidend vermogen van de statistische toets, dit is de kans om modelafwijkingen te ontdekken. Nu is het natuurlijk wel zo dat men met het construeren van formele modellen, zoals het Raschmodel, hoopt een acceptabele beschrijving te krijgen van de werkelijkheid

met een beperkt aantal concepten, doch het zou heel naïef zijn te denken dat een eenvoudig model de werkelijkheid tot in de kleinste details correct kan weergeven. Als we nu de steekproef heel groot laten worden, wordt de statistische toets ook gevoelig voor onbelangrijke modelafwijkingen, zodat het model steeds verworpen zal worden. De toetsingsgrootheid G^2 zoals gedefinieerd in (4.70) is dus niet goed bruikbaar in de praktijk.

We kunnen echter de LR-toets uitbreiden tot gevallen waarbij het verzadigd model vervangen wordt door een model dat reeds zekere beperkingen oplegt aan Ω, doch waarin we voldoende vertrouwen hebben. We zullen een toets bespreken die door Andersen (1973a) is ontwikkeld, en die geschikt is voor het geval met CML-schatters gewerkt wordt. In paragraaf 4.2.3 werd er op gewezen dat het grote voordeel van de CML-schattingsmethode erin gelegen is dat geen representatieve steekproef hoeft te worden getrokken. Dit impliceert dat, indien het Raschmodel geldig is in een bepaalde populatie, de parameters geschat kunnen worden uit de antwoorden van een willekeurige steekproef, en dat de schattingen binnen de grenzen van de steekproeffout aan elkaar gelijk moeten zijn. Als nu een gegeven steekproef opgedeeld wordt in $k-1$ substeekproeven, waarin voor elke substeekproef geldt dat iedereen dezelfde score heeft, dan kunnen de itemparameters geschat worden uit de antwoorden van elke substeekproef afzonderlijk. Die schattingen moeten ongeveer gelijk zijn aan elkaar, en aan de schattingen die we verkrijgen door de hele steekproef in één keer te analyseren. Dat 'ongeveer gelijk' kunnen we preciseren door een LR-toets te construeren. Even terzijde dient opgemerkt te worden dat de antwoordpatronen met alle items juist of alle items fout geen informatie over de items bevatten. Deze antwoordpatronen kunnen uit de steekproef ver- wijderd worden.

Als algemeen model nemen we aan dat het Raschmodel geldig is in elke subpopulatie afzonderlijk. Binnen elk van de $k-1$ scoregroepen, voor de scores 1 tot $k-1$, moeten dus $k-1$ vrije itemparameters geschat worden. De parametervector duiden we aan met φ_u - de u staat voor 'unrestricted' - en is gegeven door

$$\begin{aligned}\varphi_u &= (\varepsilon_1^{(1)},\varepsilon_2^{(1)},\ldots,\varepsilon_k^{(1)},\varepsilon_1^{(2)},\ldots,\varepsilon_i^{(s)},\ldots,\varepsilon_k^{(k-1)}) \\ &= (\boldsymbol{\varepsilon}^{(1)},\ldots,\boldsymbol{\varepsilon}^{(k-1)}),\end{aligned} \tag{4.71}$$

waarin $\varepsilon_i^{(s)}$ de parameter is van item i in de scoregroep met score s. In de vector φ_u zijn $k(k-1)$ elementen opgenomen omwille van de symmetrie in de notatie, doch er zijn slechts $(k-1)^2$ vrije parameters. Omdat de $k-1$ scoregroepen onafhankelijk zijn

van elkaar kan de aanneme- lijkheidsfunctie voor alle observaties samen geschreven worden als

$$L(\varphi_u; \boldsymbol{X} \mid \boldsymbol{s}) = \prod_{s=1}^{k-1} L(\boldsymbol{\varepsilon}^{(s)}; \boldsymbol{X}^{(s)} \mid s). \tag{4.72}$$

Indien één enkel lid van de familie van Raschmodellen voor alle scoregroepen geldig is, betekent dit dat de itemparameters voor item *i* in alle scoregroepen aan elkaar gelijk moeten zijn. We voeren dus de restrictie in

$$\boldsymbol{\varepsilon}^{(1)} = \ldots = \boldsymbol{\varepsilon}^{(s)} = \ldots = \boldsymbol{\varepsilon}^{(k-1)} = \boldsymbol{\varepsilon} \tag{4.73}$$

en de parametervector φ_r in het beperkte model, waarbij de *r* staat voor 'restricted', is gegeven door

$$\varphi_r = (\varepsilon_1, \ldots, \varepsilon_k). \tag{4.74}$$

Het is duidelijk dat de parameterruimte in het beperkte model een deelverzameling is van de parameterruimte in het algemene model. De restrictie (4.73) is de statistische nulhypothese. Bovendien is het beperkte model niets anders dan het Raschmodel zoals we het tot nog toe behandeld hebben. De toetsingsgrootheid

$$\begin{aligned} Z &= -2 \ln \frac{L^*(\varphi_r; \boldsymbol{X} \mid \boldsymbol{s})}{L^*(\varphi_u; \boldsymbol{X} \mid \boldsymbol{s})} \\ &= 2 \left[\sum_{i=1}^{k-1} \ln L^*(\boldsymbol{\varepsilon}^{(i)}; \boldsymbol{X}^{(i)} \mid s = i) - \ln L^*(\boldsymbol{\varepsilon}; \boldsymbol{X} \mid \boldsymbol{s}) \right] \end{aligned} \tag{4.75}$$

is asymptotisch chi-kwadraat verdeeld met als aantal vrijheidsgraden het verschil in aantal vrije parameters in φ_u min het aantal vrije parameters in φ_r, dus $(k-1)^2 - (k-1) = (k-1)(k-2)$. Indien de waarde van *Z* klein is, betekent dit dat het maximum van de aannemelijk- heidsfunctie niet belangrijk afneemt indien de restrictie (4.73) wordt ingevoerd; men zou kunnen zeggen dat de gegevens zich niet tegen deze restrictie verzetten, en dat we ze dus redelijkerwijze kunnen aannemen.

Om de toetsingsgrootheid *Z* uit te rekenen, moeten de parameters dus *k* keer geschat worden: één keer in elke scoregroep afzonderlijk en één keer voor alle scoregroepen samen. Indien in één van de scoregroepen de parameters niet schatbaar zijn, bijvoorbeeld omdat een item door niemand of door iedereen juist beantwoord is, kan de toetsingsgrootheid niet berekend worden. Om dit probleem op te lossen kan

men ook een LR-toets construeren door verschillende scoregroepen samen te nemen. Stel dat er G scoregroepen gevormd worden, dan veronderstelt het algemene model dat het Raschmodel geldig is in elke der G score- groepen afzonderlijk. De vector φ_u bevat dus $G(k-1)$ vrije parameters. De toetsingsgrootheid wordt uitgerekend op dezelfde manier als in (4.75) is aangegeven, met dien verstande dat de som in het rechterlid G termen bevat. Het aantal vrijheidsgraden is $(G-1)(k-1)$. Andersen (1973a) toont aan dat de toets gevoelig is voor schendingen van axioma (4), dit wil zeggen dat de toets ernaar zal tenderen een significant resultaat op te leveren als de items niet gelijkelijk discrimineren. Indien men scoregroepen samenneemt is het aan te bevelen aan- liggende scoregroepen in dezelfde groep op te nemen. Van den Wollenberg (1982) heeft laten zien dat de toets niet erg gevoelig is voor schendingen van de unidimensionaliteit.

In principe kan men ook een LR-toets construeren indien men met MML-schatters werkt, in plaats van met CML. Het uitrekenen van de toetsingsgrootheid is echter niet eenvoudig met de bestaande programmatuur. Immers het algemene model heeft als parametervector

$$\varphi_u = (\boldsymbol{\varepsilon}^{(0)},...,\boldsymbol{\varepsilon}^{(k)},\mu,\sigma^2),$$

we veronderstellen wel verschillende itemparameters in de verschillende scoregroepen, doch we nemen tevens aan dat de θ-waarden van alle personen in de steekproef een aselecte trekking zijn uit één enkele normale verdeling. De veronderstelling dat er met elke scoregroep een normale verdeling geassocieerd is, doet erg geforceerd aan. Dit betekent dat φ_u uit alle data samen geschat moet worden en daar is de bestaande programmatuur niet op gebouwd. Praktisch gezien is de LR-toets dus beperkt tot het geval dat er CML-schatters voorhanden zijn.

Uit statistisch oogpunt is er geen dwingende reden om de totale steekproef op te delen in homogene scoregroepen. De opdeling kan ook gebeuren volgens een extern criterium, bijvoorbeeld het geslacht of de leeftijd van de respondenten. Voor het gebruik van de LR-toets in zo'n geval verwijzen we naar Andersen (1980).

Een tweede toets, die door Martin-Löf (1973) is ontwikkeld, is wel gevoelig voor schending van het axioma van unidimensionaliteit. Om de toets onderscheidingsvermogen te geven moet men echter een goede hypothese hebben over welke items de verschillende dimensies vertegenwoordigen. Stel dat een toets bestaande uit k items, k_1 kale sommen bevat en k_2 redactiesommen, en dat men vermoedt dat de vaardigheid om de kale sommen op te lossen toch iets anders voorstelt dan de vaardigheid om de redactiesommen op te lossen. Een willekeurig antwoordpatroon $\boldsymbol{x}$ kunnen we schrijven

als $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})$, waarbij $\boldsymbol{x}^{(1)}$ het partiële antwoordpatroon is op de k_1 kale sommen en $\boldsymbol{x}^{(2)}$ het partiële antwoordpatroon op de k_2 redactiesommen. Het algemene model, geformuleerd als een geparametriseerd multinomiaal model geeft als kans voor een antwoordpatroon $\boldsymbol{x}$ met s_1 juiste antwoorden in $\boldsymbol{x}^{(1)}$ en s_2 juiste antwoorden in $\boldsymbol{x}^{(2)}$

$$\pi_{\boldsymbol{x}} = P(\boldsymbol{x}^{(1)} \mid s_1)\, P(\boldsymbol{x}^{(2)} \mid s_2)\, \omega_{s_1 s_2},$$

waarin $\omega_{s_1 s_2}$ de kans is op een antwoordpatroon met subscores s_1 respectievelijk s_2. In totaal moeten dus $(k_1 - 1) + (k_2 - 1) = k - 2$ vrije itemparameters geschat worden en $(k_1 + 1)(k_2 + 1) - 1 = k_1 k_2 + k$ vrije multinomiale parameters. De schattingen voor de itemparameters zijn de CML-schattingen die men verkrijgt door de twee subtoetsen met k_1 respectievelijk k_2 items afzonderlijk te analyseren. De schatters van de multinomiale parameters zijn gegeven door

$$\hat{\omega}_{s_1 s_2} = \frac{n_{s_1 s_2}}{n}.$$

Het beperkte model is niets anders dan het Raschmodel, aangevuld met een verzadigd multinomiaal model voor de scoreverdeling, berekend op beide toetsen samen. Dit model heeft $k - 1$ vrije itemparameters en k vrije multinomiale parameters, samen dus $2k - 1$. Het verschil in aantal vrije parameters tussen algemeen en beperkt model is dus $k_1 k_2 - 1$, en dat is ook het aantal vrijheidsgraden voor de toetsingsgrootheid

$$A = 2\left[\sum_{s_1=0}^{k_1}\sum_{s_2=0}^{k_2} n_{s_1 s_2} \ln(n_{s_1 s_2}/n) + \ln L^*(\boldsymbol{\varepsilon}^{(1)}; \boldsymbol{X}^{(1)} \mid \boldsymbol{s}_1) + \ln L^*(\boldsymbol{\varepsilon}^{(2)}; \boldsymbol{X}^{(2)} \mid \boldsymbol{s}_2) - \sum_{s=0}^{k} n_s \ln(n_s/n) - \ln L^*(\boldsymbol{\varepsilon}; \boldsymbol{X} \mid \boldsymbol{s})\right]. \tag{4.76}$$

Merk op dat in formule (4.76) de superscripten wijzen op een opdeling van de items in twee deeltoetsen, terwijl in (4.75) de superscripten wijzen op een opdeling van de steekproef van personen in deelgroepen.

4.3.4 Wald-toetsen

Bij de likelihood-ratio-toetsen hebben we gezien dat het maximum van de aannemelijkheids- functie onder het beperkte model niet al te veel kleiner mag zijn dan het

maximum onder het algemene model om het beperkte model aanvaardbaar te maken. Bij de Wald-toetsen gaat men uit van de volgende rationale: stel dat het beperkte model zegt dat twee parameters β_i en β_j aan elkaar gelijk moeten zijn, doch men schat de parameters zonder die gelijkheid op te leggen, dan mag men verwachten dat de schattingen van die twee parameters niet veel van elkaar zullen verschillen, indien het beperkte model waar is. Men verwacht eigenlijk dat het verschil tussen die twee schattingen uitsluitend veroorzaakt is door de steekproeffout. De nulhypothese luidt dus

$$H_0: \beta_i - \beta_j = 0.$$

Het linkerlid van deze gelijkheid is een functie van de parameters, en de nulhypothese stelt dat deze functie gelijk is aan 0. Nu kunnen we deze nulhypothese complexer maken door niet één functie te beschouwen, maar q functies tegelijkertijd waarbij q niet groter mag zijn dan het aantal vrije parameters. We beschouwen een concreet voorbeeld, dat verder in hoofdstuk 11 wordt besproken. Stel dat een onderzoeker twee Raschtoetsen van k items wil construeren die sterk parallel zijn. Daartoe trekt hij uit een grote itembank k paren van items, zodat binnen elk paar de itemparameters gelijk zijn. Om nog eens te controleren of er werkelijk aan de eis van sterke parallelliteit is voldaan, voegt hij alle items samen in één toets van $2k$ items. Neem aan dat de paren gevormd worden door de items i en $k+i$ $(i = 1, \ldots, k)$. De nulhypothese van de onderzoeker luidt dus

$$H_0: \begin{cases} h_1(\beta) = \beta_1 - \beta_{k+1} = 0 \\ \quad . \\ \quad . \\ h_i(\beta) = \beta_i - \beta_{k+i} = 0 \\ \quad . \\ \quad . \\ h_k(\beta) = \beta_k - \beta_{2k} = 0\,. \end{cases} \tag{4.77}$$

Er geldt dus $q = k$, en het aantal vrije parameters is $2k - 1$. Deze q functies kunnen we verzamelen in een q-vector $\boldsymbol{h}(\beta)$ en de nulhypothese luidt dus in deze compacte notatie:

$$H_0: \boldsymbol{h}(\beta) = \mathbf{0}. \tag{4.78}$$

Beschouw nu de toetsingsgrootheid

$$W = \boldsymbol{h}'(\hat{\beta})[T'(\hat{\beta})\hat{\Sigma}T(\hat{\beta})]^{-1}\boldsymbol{h}(\hat{\beta}), \tag{4.79}$$

waarin T een $2k \times q$ matrix is met elementen t_{ij} gedefinieerd door

$$t_{ij} = \frac{\partial h_j(\beta)}{\partial \beta_i}. \tag{4.80}$$

Σ is de variantie-covariantiematrix van de parameterschatters, en het dakje duidt aan dat alle functies geëvalueerd moeten worden op het punt van de ML-schatters. Wald (1943) heeft aangetoond dat W asymptotisch chi-kwadraat verdeeld is met q vrijheidsgraden, als de nul- hypothese waar is. In het algemeen is het aantal vrijheidsgraden gelijk aan het aantal lineair onafhankelijke restricties die samen de nulhypothese vormen. Het uitrekenen van deze toetsingsgrootheid is niet erg moeilijk omdat de geschatte covariantiematrix meestal voor- handen is als resultaat van de schattingsprocedure. Uit (4.77) volgt direct dat

$$\frac{\partial h_j(\beta)}{\partial \beta_i} = \begin{cases} 1 \text{ indien } i = j, \\ -1 \text{ indien } i = j+k, \\ 0 \text{ in andere gevallen}. \end{cases} \tag{4.81}$$

De matrix T' kan dus geschreven worden als de supermatrix $[I_k | -I_k]$, en de matrix $T'\Sigma T$ is gegeven door

$$T'\Sigma T = \begin{bmatrix} I_k & -I_k \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} I_k \\ -I_k \end{bmatrix}$$

$$= \Sigma_{11} + \Sigma_{22} - \Sigma_{12} - \Sigma_{21}.$$

Bij een significant resultaat is het heel natuurlijk om te gaan onderzoeken of het gebrek aan parallelliteit niet te wijten is aan één of meer specifieke itemparen. Dit kan men doen door de k gelijkheden in (4.77) achtereenvolgens als nulhypothese te hanteren en te toetsen. Voor elke afzonderlijke toets geldt dus dat $q = 1$, en de matrix T is een $2k \times 1$ matrix. De matrix $T'\Sigma T$ is dus een 1×1 matrix, en de toetsingsgrootheid W_j krijgt de eenvoudige vorm

$$W_j = \frac{(\hat{\beta}_j - \hat{\beta}_{j+k})^2}{\mathrm{var}(\hat{\beta}_j) + \mathrm{var}(\hat{\beta}_{j+k}) - 2\,\mathrm{cov}(\hat{\beta}_j, \hat{\beta}_{j+k})}, \quad (j = 1, \ldots, q), \tag{4.82}$$

waarin var(.) en cov(.,.) respectievelijk de variantie en covariantie aanduiden. W_j is asymptotisch chi-kwadraat verdeeld met 1 vrijheidsgraad, en $\pm\sqrt{W_j}$ is dus asymptotisch standaardnormaal verdeeld. Het teken ± beduidt dat de vierkantswortel hetzelfde algebraïsch teken krijgt als het verschil $\hat{\beta}_j - \hat{\beta}_{j+k}$ in de teller van (4.82).

Men zou natuurlijk ook kunnen starten met het uitvoeren van de k één-vrijheidsgraad toetsen, en de berekening van de meer ingewikkelde toetsingsgrootheid W achterwege laten. Dit kan men doen als men de volgende overwegingen in acht neemt: de toetsingsgrootheden W_j zijn niet onafhankelijk van elkaar. Hun som is niet gelijk aan W, en de som is ook niet chi-kwadraat verdeeld. Maar de toetsingsgrootheden W_j zijn ook niet volledig afhankelijk van elkaar. Dit betekent dat, indien alle q nulhypothesen waar zijn, de kans dat minstens één toets significant zal uitvallen groter is dan het nominaal significantieniveau α. Men kan dan bijvoorbeeld de Bonferroni toetstechniek gaan gebruiken waar bij de q afzonderlijke toetsen een significantieniveau van α/q wordt gehanteerd, doch deze techniek leidt meestal tot een zeer conservatieve globale toets: de kans dat een fout van de eerste soort gemaakt wordt is weliswaar niet groter dan α, maar kan heel veel kleiner zijn, met als gevolg dat het onderscheidingsvermogen van de toets onnodig klein is. Een toetsingsprocedure die uitgewerkt is door Hommel (1983), neemt dit onnodig strenge criterium weg, terwijl de kans op een fout van de eerste soort toch niet groter is dan α. Voor elk van de q toetsingsgrootheden W_j kan de overschrijdingskans p_j worden uitgerekend. Deze overschrijdingskansen worden geordend van klein naar groot. Deze geordende overschrijdingskansen worden aangeduid als $p_{(j)}$. Dus $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(k)}$. De algemene nulhypothese (4.77) wordt verworpen indien

$$p_{(j)} \leq \frac{j\alpha}{q C_q}, \tag{4.83}$$

waarin $C_q = \sum_{j=1}^{q} \frac{1}{j}$.

Tabel 4.5 bevat een voorbeeld, waarbij $q = 5$. $C_5 = 2.283$ en α wordt op 0.05 gesteld.

Tabel 4.5
Voorbeeld van Hommels toetsingsprocedure

j	W_j	p_j	$p_{(j)}$	$(j\alpha)/(qC_q)$
1	0.748	.387	.008	.0044
2	4.019	.045	.017	.0088
3	7.033	.008	.045	.0131
4	1.840	.175	.175	.0175
5	5.696	.017	.387	.0219

Hoewel van drie toetsingsgrootheden W_j de overschrijdingskans kleiner is dan α, leidt de procedure niet tot verwerping van de nulhypothese (4.77) op niveau α. Natuurlijk is het ook mogelijk dat men a priori verdenking koestert tegen de hypothese van parallelliteit van één of meer specifieke paren van items. In zo'n geval is het wel zinvol deze specifieke hypothesen te toetsen op het nominale α-niveau van 5%.

Het is wellicht interessant even na te gaan dat de hypothese (4.77) ook nog op een andere manier getoetst kan worden. Men had bijvoorbeeld de twee deeltoetsen aan twee onafhankelijke steekproeven kunnen aanbieden. In de schattingsprocedure worden de parameters van beide steekproeven dan afzonderlijk geschat. Noemen we de covariantiematrices van de schatters in beide steekproeven Σ_{11} respectievelijk Σ_{22}, dan volgt uit het feit dat de twee steekproeven onafhankelijk zijn van elkaar dat de matrix Σ in (4.79) gegeven is door

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \mathbf{0} \\ \mathbf{0} & \Sigma_{22} \end{bmatrix},$$

de submatrices Σ_{12} en Σ_{21} zijn nul-matrices. Voor de toetsingsgrootheden W_j is de covariantieterm in de noemer dus ook gelijk aan 0, waardoor we bij onafhankelijke steekproeven krijgen dat

$$W_j = \frac{(\hat{\beta}_j - \hat{\beta}_{j+k})^2}{\text{var}(\hat{\beta}_j) + \text{var}(\hat{\beta}_{j+k})}. \qquad (4.84)$$

Let wel: de items in de tweede steekproef zijn genummerd $k+1,\ldots,2k$. Hoewel beide toetsingsgrootheden (4.82) en (4.84) allebei asymptotisch chi-kwadraat verdeeld zijn met

1 vrijheidsgraad, zijn beide toetsingsprocedures niet equivalent. Indien de nulhypothese niet waar is, heeft de toetsingsprocedure met afhankelijke steekproeven een veel groter onder-scheidend vermogen dan de procedure met onafhankelijke steekproeven. De toetsings- procedure met onafhankelijke steekproeven heeft echter interessante toepassingen bij het onderzoek naar itemonzuiverheid. Deze toepassingen worden besproken in hoofdstuk 9.

Een toetsingsgrootheid die erg lijkt op W_j zoals gedefinieerd in (4.84) is voorgesteld door Fischer en Scheiblechner (1970), en wordt soms aangeduid als de Fischer-Scheiblechner z_i-toetsingsgrootheid. Hoewel deze toetsingsgrootheid dezelfde formele gedaante heeft als de vierkantswortel-met-teken van (4.84) is er toch een belangrijk verschil. De varianties in de noemer van (4.84) dienen berekend te worden uit de inverse van de informatiematrix. Fischer en Scheiblechner gebruiken echter alleen de hoofddiagonaal van de informatiematrix, dit is, ze gebruiken het kwadraat van (4.59) om de variantie uit te rekenen. Als de schattingen in beide steekproeven gecentreerd worden, dan wordt hierdoor de variantie waarschijnlijk overschat, en is hun toetsings-grootheid dus te klein. Zie voor een exact resultaat bij items van gelijke moeilijkheid paragraaf 4.2.5 en vooral tabel 4.4.

De nulhypothese (4.77) kan ook getoetst worden met een likelihood-ratio-toets. Immers (4.77) is een restrictie op de parameterruimte en de parameters kunnen geschat worden zon-
der en met deze restrictie. Zonder in te gaan op de technische details van het schatten onder restricties, zie daarvoor hoofdstuk 5, is het duidelijk dat voor het construeren van de LR-toets twee maal geschat moet worden, terwijl voor de Wald-toetsen alleen onder het algemene model geschat hoeft te worden. Indien we bovendien de afzonderlijke hypothesen $h_j = 0$ (j=1,...,k) zouden willen toetsen met een LR-toets, dan moeten voor elke hypothese de parameters met die specifieke restrictie opnieuw worden geschat. Voor de toetsing van de k afzonderlijke hypothesen moeten dus $k+1$ schattingsproce-dures uitgevoerd worden, terwijl de Wald-toetsen slechts één enkele schatting vereisen, wat een belangrijke werkbesparing betekent. Bovendien is er een zeer interessant resultaat uit de theoretische statistiek, dat zegt dat beide toetsen asymptotisch equivalent zijn. Dit betekent dat als n toeneemt, de toetsings- grootheden voor beide toetsen ongeveer dezelfde waarde zullen aannemen. De vrijheidsgraden voor beide toetsen zijn gelijk: het aantal restricties q in de Wald-toetsen is precies gelijk aan het verschil in het aantal vrije parameters tussen het algemene model en het beperkte model. Hoewel de keuze tussen de twee procedures voor de hand lijkt te liggen, is het opmerkelijk dat in de bestaande programmatuur bijna geen mogelijkheden zijn voorzien om de Wald toetsen routinematig uit te voeren.

4.3.5 Veralgemeende Pearson X^2-toetsen

De uitkomst van likelihood-ratio-toetsen en van Wald-toetsen is van de data afhankelijk. Bij de likelihood-ratio-toetsen worden de maxima van de aannemelijkheidsfunctie gebruikt onder verschillende restricties op de parameters, maar deze maxima zelf zijn afhankelijk van de data. Bij de Wald-toetsen wordt een functie $\boldsymbol{h}$ berekend op de schattingen van de parameters, en deze schattingen zijn eveneens van de data afhankelijk. Het verband tussen de toetsingsgrootheid en de data is in beide toetsen echter niet zeer doorzichtig. Bij de toetsen die in deze paragraaf worden besproken is het verband tussen de toetsingsgrootheid en de data veel duidelijker: de predicties die uit het model volgen worden op een directe manier met de data vergeleken. De toetsen zijn een veralgemening van de welbekende chi-kwadraat-toetsen die gebruikt worden bij de analyse van contingentietabellen. Allereerst wordt ingegaan op de algemene theorie van deze toetsen. Daarna wordt de theorie op verschillende wijzen toegepast op het Raschmodel, en dit levert toetsen op die gevoelig zijn voor bepaalde schendingen van het Raschmodel.

Algemene theorie

Hoewel de chi-kwadraat-toetsen in de sociale wetenschappen routinematig worden toegepast, kan het nuttig zijn even in te gaan op de theorie achter die toetsen. Daarom beginnen we met een voorbeeld. Stel dat we willen nagaan of de antwoorden op twee vragen in een enquête statistisch afhankelijk zijn van elkaar. De observaties waarover we beschikken zijn weergegeven in tabel 4.6. De eerste variabele kan drie waarden aannemen, *a*, *b* en *c*; de tweede variabele kan de waarden *A* en *B* aannemen. De eerste variabele duiden we aan met *X*, en de uitspraak $X=a$ betekent dus dat de eerste variabele de waarde *a* aanneemt. De tweede variabele zullen we aanduiden met *Y*. In het corpus van de tabel staan bivariate frequenties: voor 25 personen uit de steekproef geldt de uitspraak "$X=a$ en $Y=B$".

Tabel 4.6
Tweedimensionale contingentietabel

a	*b*	*c*	totaal

A	25	17	2	44
B	67	42	9	118
totaal	92	59	11	162

We kunnen van de tweedimensionale tabel 4.6 gemakkelijk een ééndimensionale tabel maken door de frequenties achter elkaar te schrijven. Dit is gebeurd in tabel 4.7.

Tabel 4.7

Tweedimensionale tabel omgevormd tot een ééndimensionale tabel

aA	*bA*	*cA*	*aB*	*bB*	*cB*
25	17	2	67	42	9

Door dit te doen, definiëren we impliciet een nieuwe variabele Z die zes verschillende waarden kan aannemen, zoals aangeduid in de bovenste regel van tabel 4.7. Het spreekt vanzelf dat beide tabellen precies dezelfde informatie bevatten. De uitspraak "$Z=aB$" is dus equivalent met de gecombineerde uitspraak over de twee oorspronkelijke variabelen "$X=a$ en $Y=B$", de waarden van Z zijn dus antwoordpatronen, en tabel 4.7 bevat de geobserveerde frequenties van alle zes mogelijke antwoordpatronen.

Om te onderzoeken of de variabelen X en Y afhankelijk zijn van elkaar, moeten we zorgvuldig een aantal stappen zetten. We moeten een model formuleren, de parameters van het model schatten, een toetsingsgrootheid definiëren en nagaan wat de overschrijdingskans is van de uit de gegevens berekende toetsingsgrootheid. Het eenvoudigste, verzadigde model is dat de zes frequenties uit tabel 4.6 een multinomiale verdeling volgen: bij een aselecte trekking uit de populatie is er de kans $\pi_{ij} = P(X=i, Y=j)$, $(i = a,b,c; j = A,B)$ dat de observatie in cel (i,j) van tabel 4.6. terechtkomt. Omdat de som van de kansen gelijk moet zijn aan 1, betekent dit dat in het verzadigde model vijf parameters geschat moeten worden. De ML-schatters in het multinomiale model zijn gelijk aan de celproporties: $\hat{\pi}_{ij} = n_{ij}/n$, zodat onmiddellijk duidelijk is dat het model de geobserveerde frequenties perfect voorspelt. Om de afhankelijkheid te onderzoeken, stellen we een nulhypothese op die afhankelijkheid ontkent. De variabelen X en Y zijn stochastisch onafhankelijk indien:

$$\pi_{ij} = \pi_i \pi_j, \quad (i = a,b,c; j = A,B) \tag{4.85}$$

waarin $\pi_i = P(X=i)$ en $\pi_j = P(Y=j)$. Omdat $\Sigma_i \pi_i = \Sigma_j \pi_j = 1$, zijn er in het beperkte model slechts drie parameters. Hun ML-schatters zijn gegeven door de relatieve frequenties

van de marginale totalen: $\hat{\pi}_i = n_i/n$ en $\hat{\pi}_j = n_j/n$. In het beperkte model is de ML-schatter van π_{ij} dan gegeven door:

$$\hat{\pi}_{ij} = \hat{\pi}_i \hat{\pi}_j = \frac{n_i n_j}{n^2} \tag{4.86}$$

en de verwachte frequentie in de *(i,j)*-de cel van tabel 4.6 is gegeven door de welbekende formule:

$$E_{ij} = n\hat{\pi}_{ij} = \frac{n_i n_j}{n}. \tag{4.87}$$

Indien de restrictie (4.85) geldig is, mogen de verwachte frequenties E_{ij} niet al te veel afwijken van de geobserveerde frequenties O_{ij}, niet meer dan door de steekproeffout kan worden verklaard. Pearson heeft aangetoond dat de toetsingsgrootheid

$$X^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{4.88}$$

asymptotisch chi-kwadraat verdeeld is. Het aantal vrijheidsgraden is gelijk aan het aantal vrije cellen in de tabel verminderd met het aantal geschatte parameters. In het voorbeeld dus 5-3=2. De grootheid X^2, berekend op de gegevens van tabel 4.6, bedraagt 0.53, terwijl de kritieke waarde voor α=0.05 in de chi-kwadraatverdeling met twee vrijheidsgraden 5.99 is. Er is dus geen reden om het model van onafhankelijkheid (4.85) te verwerpen. Het is belangrijk het aantal termen in de som van het rechterlid van (4.88) niet te verwarren met het aantal vrije cellen. Er moet gesommeerd worden over alle cellen van de tabel en niet alleen over de vrije cellen.

Er is vrij uitvoerig op dit voorbeeld ingegaan, opdat duidelijk zou worden dat er een aantal stappen is gezet die in de routinematige uitvoering van de toets vaak niet meer worden opgemerkt. We becommentariëren deze stappen een voor een.

(1) Er is steeds sprake van een model, en van restricties op de parameterruimte. Pearson heeft zijn toets ontwikkeld voor het geval het model een multinomiaal model is. Daarom is het belangrijk bij toepassingen van Pearsons toets steeds precies na te gaan of het model waarmee men werkt beschouwd kan worden als een multinomiaal model. De nulhypothese komt steeds overeen met een restrictie op de parameterruimte. In het voorbeeld is deze restrictie gegeven door (4.85). Het is belangrijk op te merken dat Pearsons toets niet beperkt is tot deze restrictie alleen. De methode die Pearson heeft ontworpen is geldig voor een zeer grote klasse van restricties. Voor alle gevallen die in dit boek worden beschouwd,

kan de methode worden toegepast. Een uiteenzetting van de statistische theorie kan men vinden in hoofdstuk 14 van Bishop, Fienberg en Holland (1975). Men zou bijvoorbeeld het beperkte model (4.85) nog verder kunnen beperken met de extra eis:

$$\pi_a = \pi_b = \pi_c = 1/3. \tag{4.89}$$

(2) Er moeten parameters geschat worden, en deze parameters worden geschat onder de nulhypothese. Gebruiken we bijvoorbeeld (4.85) en (4.89) samen als nulhypothese, dan hoeft alleen de parameter π_A te worden geschat, want de andere parameters zijn precies vastgelegd. Merk bovendien op dat de parameters geschat worden uit dezelfde data als waarop de grootheid X^2 wordt berekend.

(3) De verwachte frequenties moeten worden uitgerekend met de schattingen van de para- meters onder de nulhypothese. De eerste gelijkheid in (4.87) is dus algemeen geldig, de tweede gelijkheid niet: deze geldt alleen onder de nulhypothese van onafhankelijkheid. Nemen we (4.85) en (4.89) samen als nulhypothese, dan krijgen we als verwachte frequenties

$$E_{ij} = n\hat{\pi}_{ij} = n\pi_i\hat{\pi}_j = \frac{n_j}{3}. \tag{4.90}$$

(4) De steekproevenverdeling van X^2 in (4.88) is niet bekend. Pearson heeft aangetoond dat, indien n toeneemt deze steekproevenverdeling steeds beter gaat lijken op de theoretische chi-kwadraatverdeling. De chi-kwadraatverdeling wordt dus gebruikt als een benadering voor de echte steekproevenverdeling van X^2. Hoe goed die benadering in concrete gevallen is, weten we niet exact. Wel is door veel onderzoek bekend dat voor praktische doeleinden het gebruik van de chi-kwadraatverdeling gerechtvaardigd is indien n niet al te klein is en indien er niet al te veel cellen zijn met kleine verwachte frequenties. Soms wordt de vuistregel gehanteerd dat het aantal cellen met verwachte frequentie kleiner dan 5 niet meer mag bedragen dan 20% van het aantal cellen. Wat men in zulke gevallen meestal doet is overgaan tot het samennemen van cellen. In tabel 4.6 zou men bijvoorbeeld alle cellen 'b' en 'c' kunnen samennemen, zodat er een 2×2 tabel ontstaat. Deze procedure is zeker gerechtvaardigd, mits men goed in het oog houdt dat hierdoor een nieuwe variabele X' gecreëerd wordt, die niet drie maar slechts twee antwoord- categorieën heeft. Het toepassen van Pearsons toets gebeurt dan op de twee variabelen X' en Y, die samen maar vier waarden kunnen aannemen. Kortom, er wordt een nieuw model geformuleerd, de

parameters worden opnieuw geschat en het besluit dat men trekt is alleen van toepassing op de variabelen X' en Y, en niet op X en Y.

(5) Het besluit dat men neemt, aanvaarden of verwerpen van de nulhypothese, betreft de nulhypothese als geheel. Is de nulhypothese bijvoorbeeld de combinatie van (4.85) en (4.89), die in het voorbeeld zeker verworpen moet worden, dan volgt uit de toetsing niet of de significantie te wijten is aan (4.85) of aan (4.89). Werkt men met heel complexe nulhypothesen, zoals het Raschmodel, dan geeft de toetsingsgrootheid dus niet de mogelijkheid een modelschending precies te lokaliseren. Pearsons toets is dus een globale toets van het model.

Passen we nu het voorgaande toe op het Raschmodel, dan is het vrij eenvoudig om de toetsingsgrootheid X^2 te construeren. Naar analogie met de tabellen 4.6 en 4.7 kunnen we de observaties onderbrengen in een k-dimensionale frequentietabel, of in een unidimensionale tabel. De tweede voorstelling is voor onze doeleinden het handigst. Bij een toets met k items zijn er 2^k antwoordpatronen mogelijk, en elke persoon die de toets beantwoordt, levert precies één antwoordpatroon op. Bij n personen kunnen we dus de frequentie $n_{\boldsymbol{x}}$ bepalen waarmee antwoordpatroon $\boldsymbol{x}$ is opgetreden. Alle frequenties samen volgen dus de multinomiale verdeling; het model is zeker niet verzadigd want er zijn $2^k - 1$ vrije cellen en er zijn maar $k-1$, in het geval van CML, of $k+1$, in het geval van MML, parameters geschat. De grootheid X^2 is dus gegeven door:

$$\begin{aligned} X^2 &= \sum_{\boldsymbol{x}} \frac{(n_{\boldsymbol{x}} - n\hat{\pi}_{\boldsymbol{x}})^2}{n\hat{\pi}_{\boldsymbol{x}}} \\ &= n \sum_{\boldsymbol{x}} \frac{(p_{\boldsymbol{x}} - \hat{\pi}_{\boldsymbol{x}})^2}{\hat{\pi}_{\boldsymbol{x}}}, \end{aligned} \tag{4.91}$$

waarin $p_{\boldsymbol{x}} = n_{\boldsymbol{x}}/n$. X^2 is asymptotisch chi-kwadraat verdeeld met $2^k-1-(k-1) = 2^k - k$ vrij- heidsgraden (CML) of 2^k-k-2 vrijheidsgraden (MML). Het bezwaar tegen het gebruik van deze toetsingsgrootheid is natuurlijk dat reeds bij middelgrote k, zeg 20, het aantal cellen van de tabel vele malen groter zal zijn dan de steekproef, zodat automatisch zeer veel, zo niet alle cellen een heel kleine verwachte waarde zullen hebben. Bij k=20 en n=1000 is de gemiddelde verwachte frequentie kleiner dan .001. Het is wel zeker dat het gebruiken van de chi-kwadraatverdeling als benadering van de verdeling van X^2 niet terecht is. Er zit dus niet veel anders op dan onze toevlucht te nemen tot het samenvoegen van cellen. Doch dan zouden strikt genomen de parameters opnieuw geschat moeten worden, waarbij in de schattings-procedure geen gebruik

gemaakt mag worden van de afzonderlijke frequenties van de samengevoegde cellen. Zo'n schattingsprocedure opzetten is echter vrij moeilijk en omslachtig.

Glas en Verhelst (1989) hebben een methode ontwikkeld om een soort correctie op de gewone grootheid X^2 aan te brengen, zonder dat de parameters opnieuw geschat moeten worden. Bovendien is hun methode algemener toepasbaar dan in de situatie waar cellen worden samengenomen. Bij het samennemen van cellen worden de cellen van de oorspronkelijke contingentietabel ingedeeld in een aantal groepen, en elke van de oorspronkelijke cellen wordt aan precies één groep toegewezen. Bij de methode van Glas en Verhelst is het ook mogelijk bepaalde cellen aan meer groepen groep toe te wijzen of cellen buiten beschouwing te laten. Later zullen we zien dat deze mogelijkheid ons in staat stelt om gerichte toetsen te construeren in plaats van alleen maar een globale toets.

De methode is vrij complex en zal in een aantal stappen worden uiteengezet. Eerst wordt aangetoond hoe Pearsons grootheid X^2 als een matrix-expressie kan worden geschreven. Deze matrix-expressie wordt een kwadratische vorm genoemd. Vervolgens wordt getoond hoe het samennemen of groeperen van cellen kan gebeuren door gebruik te maken van een speciaal daartoe geconstrueerde matrix Y. De toetsingsgrootheid Q, waarmee we gaan werken, is ook een kwadratische vorm. De waarde die deze kwadratische vorm aanneemt is afhankelijk van de observaties, maar ook van de matrix Y die we geconstrueerd hebben. Om deze afhankelijkheid expliciet aan te geven zullen we de toetsingsgrootheid aanduiden als $Q(Y)$. De centrale vraag is natuurlijk of $Q(Y)$ asymptotisch chi-kwadraat verdeeld is, en wat het geassocieerde aantal vrijheidsgraden is. Met een voorbeeld zullen we aantonen dat $Q(Y)$ niet chi-kwadraat verdeeld is voor elke matrix Y. Glas en Verhelst hebben een klasse van Y-matrices gekarakteriseerd waarvoor $Q(Y)$ wel asymptotisch chi-kwadraat verdeeld is. We zullen dit resultaat niet in zijn algemeenheid bespreken, maar ons beperken tot het geval waar het geparametriseerd multinomiaal model tot de exponentiële familie behoort.

Pearsons X^2 als een kwadratische vorm

Om elegant te kunnen werken is het nuttig (4.91) als een matrix-expressie te schrijven. Definieer $m = 2^k$, m is dus het aantal mogelijke antwoordpatronen. De geobserveerde proporties $p_{\boldsymbol{x}}$ worden verzameld in de vector $\boldsymbol{p}$ en de geschatte kansen $\hat{\pi}_{\boldsymbol{x}}$ in de vector $\hat{\pi}$. Bovendien definiëren we een diagonaalmatrix $D_{\hat{\pi}}$, met de elementen van $\hat{\pi}$ op de diagonaal. Het is gemakkelijk na te gaan dat (4.91) geschreven kan worden als:

$$X^2 = n(\boldsymbol{p} - \hat{\pi})' D_{\hat{\pi}}^{-1} (\boldsymbol{p} - \hat{\pi})$$
$$= n(\boldsymbol{p} - \hat{\pi})' I_m (I_m D_{\hat{\pi}} I_m)^{-1} I_m (\boldsymbol{p} - \hat{\pi}), \quad (4.92)$$

waarbij I_m de $m \times m$ identiteitsmatrix is. De algemene gedaante van (4.92) is het produkt van een rijvector met een symmetrische matrix met een kolomvector, waarbij de twee vectoren in het produkt gelijk zijn aan elkaar. Een dergelijk produkt wordt in de lineaire algebra een kwadratische vorm genoemd. Door het toevoegen van de identiteitsmatrix wordt expliciet aangegeven dat de som in (4.91) uit m termen bestaat: elke afwijking tussen geobserveerde (p) en verwachte (π) proportie wordt gekwadrateerd, en draagt dus bij tot de som X^2.

Het samennemen van cellen

De manier waarop cellen moeten worden samengenomen kan worden aangegeven in een speciaal daartoe geconstrueerde matrix Y. De matrix Y in tabel 4.8 is een voorbeeld voor een geval met $k = 3$ items. De matrix bevat alleen enen en nullen, en voorlopig kunnen we er vanuit gaan dat de enen op willekeurige plaatsen zijn neergezet. De acht mogelijke antwoordpatronen zijn afgebeeld onder het kopje T_1; de matrix T_2 komt later aan de orde.
Beschouw nu het produkt $(\boldsymbol{p} - \hat{\pi})' \boldsymbol{y}_2$, waarin $\boldsymbol{y}_2$ de tweede kolom van Y is. Dit produkt geeft de som van de afwijkingen $p_x - \pi_x$ voor het vijfde en het zevende antwoordpatroon, dit is voor de twee antwoordpatronen waarvoor een 1 staat in de overeenkomstige rij van de tweede kolom van Y. Op analoge manier is het produkt $(\boldsymbol{p} - \hat{\pi})' \boldsymbol{y}_1$ de som (met één term) van alle antwoordpatronen waarbij een 1 staat in de eerste kolom van Y. Men kan ook zeggen dat in elke kolom alle afwijkingen meedoen: ze worden eerst vermenigvuldigd met een constante die in hun rij staat. In het voorbeeld zijn die constanten 1 of 0, maar we hadden ook andere constanten kunnen invullen. Het vermenigvuldigen van een aantal elementen, de afwijkingen, met een constante en die produkten bij elkaar optellen geeft een som die men een lineaire combinatie van die elementen noemt. De constanten waarmee vermenigvuldigd is, worden de coëfficiënten genoemd. Het produkt $(p - \hat{\pi})' \boldsymbol{Y}$ definieert dus in het algemeen evenveel
lineaire combinaties als er kolommen zijn in Y. Merk op dat de antwoordpatronen 1, 2, 4,

Tabel 4.8

Constructie van de matrix voor de veralgemeende Pearson toetsen

T_1	T_2	Y

0	0	0	1	0	0	0	0	0
1	0	0	0	1	0	0	0	0
0	1	0	0	1	0	0	1	0
0	0	1	0	1	0	0	0	0
1	1	0	0	0	1	0	0	1
1	0	1	0	0	1	0	0	0
0	1	1	0	0	1	0	0	1
1	1	1	0	0	0	1	0	0

6 en 8 in geen van beide groepen zijn opgenomen. Het zal duidelijk zijn dat een matrix Y die de antwoordpatronen groepeert in de gebruikelijke zin van het woord, aan de volgende eis moet voldoen: in elke rij van de matrix moet precies één 1 voorkomen, de andere elementen van de rij zijn gelijk aan nul. Het groeperen is dus ook het nemen van een aantal lineaire combinaties.

Beschouw nu de kwadratische vorm

$$Q(Y) = n(\boldsymbol{p} - \hat{\pi})' Y (Y' D_{\hat{\pi}} Y)^{-} Y' (\boldsymbol{p} - \hat{\pi}), \tag{4.93}$$

waarin de aanduiding '-' in superscript een veralgemeende inverse aanduidt. Indien de matrix Y niet van volle rang is, dat wil zeggen, indien één of meer van zijn kolommen kunnen worden geschreven als een lineaire combinatie van de andere kolommen, dan is de matrix $Y' D_{\hat{\pi}} Y$ singulier en heeft geen reguliere inverse. Singuliere matrices hebben echter wel oneindig veel zogenaamde veralgemeende inversen. De kwadratische vorm $Q(Y)$ heeft echter altijd dezelfde waarde, ongeacht welke veralgemeende inverse men kiest. Indien de matrix van de kwadratische vorm niet singulier is, is de inverse matrix uniek. Een vergelijking van (4.93) met (4.92) leert ons onmiddellijk dat $X^2 = Q(I_m)$, dus X^2 is een speciaal geval van (4.93) met $Y = I_m$. Daaruit volgt echter niet dat $Q(Y)$ asymptotisch chi-kwadraat verdeeld is voor elke Y.

$Q(Y)$ is niet voor elke Y chi-kwadraat verdeeld

De antwoordpatronen waarbij een 1 staat in de tweede kolom van de matrix Y in tabel 4.8 kunnen als volgt worden omschreven: het zijn alle antwoordpatronen die een juist antwoord hebben op item 2 en een score 2. Indien de parameters met CML geschat zijn geldt: $\hat{\pi}' \boldsymbol{y}_2 = n^{-1}(n_2 \hat{\pi}_{2|2})$. Voor de geobserveerde proporties geldt analoog dat $\boldsymbol{p}' \boldsymbol{y}_2 = n^{-1}(n_2 p_{2|2})$. De ene 1 in de eerste kolom heeft betrekking op het antwoordpatroon met score 1 en een juist antwoord op item 2, zodat ook hier soortgelijke

uitdrukkingen gelden voor de produkten $\hat{\pi}'\boldsymbol{y_1}$ en $\boldsymbol{p}'\boldsymbol{y}_1$. Omdat in de rijen van de matrix Y nooit meer dan één element verschilt van 0 is de matrix $Y'D_{\hat{\pi}}Y$ een diagonaalmatrix. De kwadratische vorm (4.93) kan dan ook expliciet geschreven worden als

$$Q(Y) = \sum_{s=1}^{2} \frac{n_s(p_{2|s} - \hat{\pi}_{2|s})^2}{\hat{\pi}_{2|s}}. \tag{4.94}$$

Hoewel deze uitdrukking erg lijkt op het laatste lid van (4.91), zijn er enkele belangrijke verschillen. Deze kunnen we het beste toelichten door de score × itemantwoord-contingentietabel te construeren (zie tabel 4.9).

Tabel 4.9

Verwachte frequenties in de score × itemantwoord-tabel voor item 2

	x_2=0	x_2=1
s=0	---	---
s=1	$n_1(1-\hat{\pi}_{2\|1})$	$n_1\hat{\pi}_{2\|1}$
s=2	$n_2(1-\hat{\pi}_{2\|2})$	$n_2\hat{\pi}_{2\|2}$
s=3	---	---

Er zijn twee opmerkelijke verschillen met de situatie die leidde tot formule (4.91). Het eerste is dat in de som (4.94) maar twee termen zijn opgenomen en niet vier, zoals door tabel 4.9 wordt gesuggereerd. Bovendien zijn vier van de mogelijke cellen helemaal uit de kwadratische vorm weggelaten. Nu is het wel zo dat in die vier cellen de score 0 of 3 bedraagt, waardoor de geobserveerde en verwachte frequenties precies aan elkaar gelijk zijn, maar in het algemeen kan natuurlijk een matrix Y geconstrueerd worden waarbij cellen worden weggelaten, waarvoor de overeenkomst tussen geobserveerde en verwachte proporties niet perfect is. De wel ingevulde cellen waarvoor $x_2 = 0$ zijn ten onrechte niet meegeteld.

Het tweede verschil heeft te maken met de parameterschattingen en het aantal vrijheidsgraden. In totaal zijn er vijf vrije parameters geschat: twee itemparameters en drie parameters ω_s voor het verzadigde multinomiale model van de scorefrequenties. In tabel 4.9 zijn vier vrije cellen, en het mechanisch toepassen van de regel voor het bepalen van de vrijheidsgraden zou 4-5=-1 vrijheidsgraden opleveren, hetgeen natuurlijk

onzin is. De vijf parameters kunnen natuurlijk niet geschat worden als alleen de frequenties gegeven zijn die overeenkomen met de ingevulde cellen van tabel 4.9. Dit toont duidelijk aan dat $Q(Y)$ niet asymptotisch chi-kwadraat verdeeld is voor elke willekeurige matrix Y.

Een klasse van Y-matrices waarvoor $Q(Y)$ asymptotisch chi-kwadraat verdeeld is

Glas en Verhelst (1989) hebben een klasse van Y-matrices gekarakteriseerd waarvoor geldt dat $Q(Y)$ asymptotisch chi-kwadraat verdeeld is. Hier geven we alleen het resultaat voor exponentiële-familiemodellen. Om de uiteenzetting niet nodeloos abstract te maken, zullen we de principes eerst uiteenzetten aan de hand van een concreet voorbeeld, het Raschmodel, waarbij de parameters met CML geschat worden. Zoals reeds is opgemerkt zijn de CML-schatters in het Raschmodel equivalent met de gewone ML-schatters van de itemparameters, als we het Raschmodel aanvullen met een verzadigd multinomiaal model voor de scoreverdeling.

Het resultaat van Glas en Verhelst is het gemakkelijkst te begrijpen door gebruik te maken van voldoende steekproefgrootheden. Om te laten zien dat het Raschmodel, aangevuld met een verzadigd multinomiaal model voor de verdeling van de scores een lid van de exponentiële familie is, definiëren we $k+1$ zogenaamde indicatorvariabelen $t_j, j=0,...,k$, die de waarde 1 of 0 kunnen aannemen. De variabele $t_j=1$ indien de score op de k items gelijk is aan j, anders is t_j gelijk aan 0. Merk op dat de waarde van t_j eenduidig uit de antwoord- vector $\boldsymbol{x}$ kan worden berekend. Voorbeeld: als $k=3$ en $\boldsymbol{x}=(1\ 0\ 1)$, dan is de score 2, en de indicatorvector heeft de waarde $t=(0\ 0\ 1\ 0)$. We kunnen dus evengoed zeggen dat de observatie bestaat uit het antwoordpatroon $\boldsymbol{x}$, als uit de combinatie van antwoordpatroon en indicatorvector $(\boldsymbol{x},\boldsymbol{t})$. De uitdrukking (4.67) kunnen we dus ook schrijven als $\pi_{\boldsymbol{x},\boldsymbol{t}} = \pi_{\boldsymbol{x}|\boldsymbol{t}}\,\pi_{\boldsymbol{t}}$, waarin de eerste factor in het rechterlid de conditionele kans op het antwoordpatroon is, gegeven de indicator van de score. De log-aannemelijkheidsfunctie is gegeven door

$$\ln\ L(\boldsymbol{\varepsilon},\pi;\boldsymbol{x},\boldsymbol{t}) = \sum_i x_i \ln \varepsilon_i + \sum_j t_j \ln \omega_j - \ln \gamma_s(\boldsymbol{\varepsilon}) \qquad (4.95)$$

waaruit duidelijk blijkt dat de vector $(\boldsymbol{x},\boldsymbol{t})$ een voldoende steekproefgrootheid is voor de parameters: de vector $\boldsymbol{t}$ is voldoende voor de multinomiale parameters ω_s en de vector $\boldsymbol{x}$ is voldoende voor de itemparameters. Het feit dat de vector $(\boldsymbol{x},\boldsymbol{t})$ $2k+1$ elementen bevat, terwijl er maar $2k$-1 vrije parameters zijn is voorlopig niet belangrijk; we komen er later op terug.

Om er voor te zorgen dat de kwadratische vorm $Q(Y)$ asymptotisch chi-kwadraat verdeeld is, kan aangetoond worden dat de voldoende steekproefgrootheden ($\boldsymbol{x},\boldsymbol{t}$) op een of andere manier te vinden moeten zijn in elke rij van de matrix Y. Dit is, kort samengevat, het belangrijkste resultaat van Glas en Verhelst. Voor de matrix Y in tabel 4.8 is dit zeker niet het geval. Een eenvoudige manier om de voldoende steekproefgrootheden in de matrix te brengen, bestaat erin een gegeven matrix Y uit te breiden met die steekproefgrootheden. Dit is gebeurd in tabel 4.8. De rijen van de matrix T_1 zijn de antwoordpatronen $\boldsymbol{x}$ en de rijen van T_2 zijn de erbij behorende indicatorvectoren $\boldsymbol{t}$. Definieer nu $T=[T_1|T_2]$, en $Z=[T_1|T_2|Y]= [T|Y]$. In plaats van $Q(Y)$ wordt $Q(Z)$ uitgerekend, en omdat in de rijen van Z de afdoende steekproefgrootheden aanwezig zijn, geldt het volgende resultaat:

(1) $Q(Z)=Q([T|Y])$ is asymptotisch chi-kwadraat verdeeld waarbij het aantal vrijheidsgraden gelijk is aan de rang van de matrix Z min 1, min het aantal geschatte parameters. Dit geldt voor elke matrix Y.

Men zou natuurlijk kunnen opperen dat dit allemaal goed en wel is, doch dat daarmee het oorspronkelijke probleem is veranderd. Bij de behandeling van het voorbeeld zijn we immers begonnen met het beschouwen van slechts twee lineaire combinaties van afwijkingen, namelijk $(\boldsymbol{p}-\hat{\boldsymbol{\pi}})'\boldsymbol{y}_1$ en $(\boldsymbol{p}-\hat{\boldsymbol{\pi}})'\boldsymbol{y}_2$, terwijl de matrix Z negen kolommen heeft, en het produkt $(\boldsymbol{p}-\hat{\boldsymbol{\pi}})'Z$ dus negen lineaire combinaties definieert. Er kan echter bewezen worden (Glas, 1989) dat, indien de parameters zijn geschat met de ML methode, geldt:

(2) $(\boldsymbol{p}-\hat{\boldsymbol{\pi}})'T = \boldsymbol{0}$. Daaruit volgt onmiddellijk dat $Q(T) = 0$.

De lineaire combinaties die we toegevoegd hebben zijn dus gelijk aan 0. Dit betekent echter niet dat $Q(Y)=Q(Z)$. Het belangrijkste verschil is dat de matrix $Z'D_{\hat{\pi}}Z$ gebruikt moet worden in de kwadratische vorm en niet de diagonale matrix $Y'D_{\hat{\pi}}Y$. De reden hiervoor is dat de parameters uit de oorspronkelijke data geschat zijn en niet uit de lineaire combinaties $\boldsymbol{p}'Y$ die minder informatie bevatten.

Hiervoor werd gezegd dat de voldoende steekproefgrootheden 'aanwezig' moesten zijn in de matrix Z van lineaire combinaties. We hebben ons van die aanwezigheid verzekerd door een gegeven matrix uit te breiden. Dit is een handige methode, maar ze is niet noodzakelijk. De precieze definitie van aanwezig zijn is als volgt. Stel dat een geparametriseerd multinomiaal model met s vrije parameters tot de exponentiële familie behoort. Het aantal verschillende antwoordpatronen is m. Beschouw de $m \times s$

matrix U, waarvan elke rij de minimaal voldoende steekproefgrootheden voor het desbetreffende antwoordpatroon bevat. Voor een gegeven $m \times r$ matrix Z, waarbij $r>s+1$, is de kwadratische vorm $Q(Z)$, gedefi- nieerd door (4.93) asymptotisch chi-kwadraat verdeeld als aan de volgende twee voorwaarden is voldaan:

(3) elke kolom van de matrix U kan geschreven worden als een lineaire combinatie van de kolommen van Z;
(4) de m-vector **1**, dit is de vector waarvan alle elementen gelijk zijn aan 1, kan geschreven worden als een lineaire combinatie van de kolommen van Z.

Voor de matrix $Z=[T_1|T_2|Y]$ uit tabel 4.8 is dit het geval. Er zijn slechts twee vrije itemparameters en drie vrije marginale kansen ω_s. De matrix U kunnen we dus vormen door in de matrix $T=[T_1|T_2]$ bijvoorbeeld de eerste kolom van T_1 en de eerste kolom van T_2 te schrappen. Aan voorwaarde (3) is dan op een triviale manier voldaan. Door de kolommen van de matrix T_2 bij elkaar op te tellen zien we ook dat aan voorwaarde (4) is voldaan.

We beschikken dus over twee manieren om aan te tonen dat, binnen de exponentiële familie, de kwadratische vorm $Q(Z)$ asymptotisch chi-kwadraat verdeeld is: ofwel we breiden een gegeven matrix Y uit met een matrix die de voldoende steekproefgrootheden en de vector **1** bevat, ofwel we tonen aan dat aan de voorwaarden (3) en (4) is voldaan.

Voor een gedetailleerde uiteenzetting van bovenstaande resultaten, zie Glas (1989), Glas en Verhelst (1989) en Verhelst en Eggen (1989).

Praktische problemen

Het resultaat dat hierboven is gegeven, heeft zeer veel toepassingsmogelijkheden omdat de matrix Y die in resultaat (1) staat volkomen willekeurig is. Alle toetsen van het Raschmodel die hierna nog besproken zullen worden, zijn speciale gevallen van (4.93). De algemeenheid van het resultaat dient echter niet overschat te worden, want er duiken een viertal praktische problemen op waarmee men in de toepassing terdege rekening moet houden.

Het eerste probleem heeft te maken met het uitrekenen van de kwadratische vorm $Q(Y)$. De matrix Y heeft $m=2^k$ rijen. Indien we de kwadratische vorm $Q(Y)$ uitrekenen met de matrixvermenigvuldigingen als aangegeven in (4.93), moet gigantisch veel rekenwerk worden uitgevoerd, zelfs voor niet al te grote k. We zullen dus moeten

zoeken naar een aangepaste definitie van de matrix Y waardoor het rekenwerk snel en efficiënt kan verlopen.

Het tweede probleem heeft te maken met het aantal vrijheidsgraden. Dat aantal is gegeven door rang(Y)-s-1, waarin s het aantal vrije parameters van het model is. Het bepalen van de rang van Y moet met de nodige zorgvuldigheid gebeuren. Ook als we de methode van de toegevoegde matrix T gebruiken, en de kwadratische vorm $Q([T|Y])$ beschouwen, is het niet automatisch zo dat het aantal vrijheidsgraden gelijk is aan het aantal kolommen van Y. In het voorbeeld van tabel 4.8 is het aantal vrije parameters s gelijk aan 5, de rang van de matrix $T=[T_1|T_2]$ is s+1=6, maar de rang van $Z=[T|Y]$ is niet 6+2=8, maar 7, omdat de kolom- men van Y lineair afhankelijk zijn van de kolommen van T. Dit kan men in tabel 4.8 gemakkelijk controleren: de som van de twee kolommen van Y is gelijk aan de tweede kolom van T_1 min de laatste kolom van T_2. Het aantal vrijheidsgraden geassocieerd met $Q(Z)$ is dus niet 2 maar 1.

Het derde probleem heeft te maken met het feit dat van $Q(Y)$ alleen de asymptotische verdeling bekend is, maar niet de exacte verdeling. De chi-kwadraatverdeling wordt dus gebruikt als een benadering van de exacte verdeling. Het is echter niet bekend hoe goed die benadering is in concrete gevallen. Het enige wat we eigenlijk kunnen doen, is waarschuwen tegen het gebruik van (4.93) en de chi-kwadraatverdeling bij zeer kleine steekproeven, en het vermijden van lineaire combinaties in de matrix Y die zeer kleine proporties van het totale aantal observaties vertegenwoordigen. Zo is de eerste kolom van de matrix Y in tabel 4.8 een lineaire combinatie waarin alleen het antwoordpatroon (0 1 0) is betrokken. Als het aantal personen in de steekproef met dit antwoordpatroon zeer klein is, kan betwijfeld worden of de chi-kwadraatverdeling wel een goede benadering is van de exacte verdeling van de kwadratische vorm.

Het vierde probleem is het belangrijkste en luidt: "hoe moet men de matrix Y kiezen ?" Het feit dat $Q(Y)$ voor een grote klasse van Y-matrices asymptotisch chi-kwadraat verdeeld is, betekent niet dat het er niet toe doet welke matrix we uit die klasse kiezen. De kwadratische vorm is alleen chi-kwadraat verdeeld onder de nulhypothese, dat wil zeggen indien het model waar is. Indien één of meer veronderstellingen van het model geschonden zijn, is het onderscheidend vermogen van de statistische toets afhankelijk van de matrix Y die we gekozen hebben. Dit kunnen we reeds zien door een inspectie van formule (4.94). De afwijkingen die in de teller staan hebben betrekking op item 2. Het is dus te verwachten dat het gebruik van de matrix Y uit tabel 4.8 een toets zal opleveren die vooral gevoelig is indien er, in termen van het model, iets mis is met item 2, eerder dan met item 1 of item 3.

Bij de specifieke toetsen voor het Raschmodel die hierna worden besproken, zal aan deze vier problemen aandacht worden geschonken.

De S_i-toetsen

De S_i-toetsen zijn bedoeld om modelschendingen op itemniveau te kunnen ontdekken. Voor elk item wordt een toets geconstrueerd, en de matrix Y heeft betrekking op een bepaald item. In deze paragraaf wordt dit specifieke item aangeduid met de index i. Om dit expliciet aan te geven krijgt de matrix Y een index i mee. Deze toetsen zijn alleen van toepassing indien de parameters met de CML-methode zijn geschat.

Het totale scorebereik wordt opgedeeld in r intervallen, dat wil zeggen de scores worden opgedeeld in r scoregroepen van aaneengesloten scores. Daarbij mogen de score 0 en de perfecte score buiten beschouwing gelaten worden. Deze scoregroepen duiden we aan als de verzamelingen G_q, q=1,...,r. Bijvoorbeeld, stel k=10 en r=3, dan is een mogelijke opdeling G_1={1,2,3,4}, G_2={5,6} en G_3={7,8,9}. De matrix Y heeft r kolommen waarbij elke kolom overeenkomt met een scoregroep. De waarden in de Y_i-matrix zijn 0 of 1; een 1 in de q-de kolom wordt ingevuld voor elke rij (antwoordpatroon) indien de score van dit antwoordpatroon behoort tot de q-de scoregroep, en indien het een antwoordpatroon betreft met een juist antwoord op item i. De matrix Y in tabel 4.8 is volgens deze regel geconstrueerd, waarbij r=2, G_1={1}, G_2={2} en i=2. Merk op dat uit deze regel volgt dat in elke rij van de Y-matrix niet meer dan één element kan verschillen van 0. Dit heeft het prettige voordeel dat de matrix $Y_i' D_{\hat{\pi}} Y_i$ een diagonale matrix is. De kolommen van Y_i zijn echter lineair afhankelijk van de kolommen van T, zoals hierboven reeds is aangetoond. Definiëren we nu twee vectoren met lineaire combinaties van afwijkingen tussen $\boldsymbol{p}$ en π:

$$\boldsymbol{d}_1 = (\boldsymbol{p} - \hat{\pi})' T, \quad \boldsymbol{d}_2 = (\boldsymbol{p} - \hat{\pi})' Y_i,$$

dan weten we uit de vorige paragraaf dat $\boldsymbol{d}_1 = \boldsymbol{0}$. Door een vrij lange afleiding, die we hier niet bespreken, zie Verhelst en Eggen (1989) voor details, kan aangetoond worden dat de kwadratische vorm $Q([T|Y_i])$ gegeven is door:

$$Q([T|Y_i]) = n\boldsymbol{d}'_2 [Y_i' D_{\hat{\pi}} Y_i - \Delta_i - A_i]^{-} \boldsymbol{d}_2. \tag{4.96}$$

De matrix Δ_i in (4.96) is een $r \times r$ diagonale matrix waarvan de elementen op de diagonaal gegeven zijn door

$$(\Delta_i)_{qq} = \sum_{s \in G_q} \frac{n_s}{n} \hat{\pi}^2_{i|s}. \tag{4.97}$$

De matrix A_i is een symmetrische $r \times r$ matrix waarvan de elementen afhankelijk zijn van de informatiematrix, zie (4.48). De precieze definitie van de elementen van A_i is nogal omslachtig en wordt hier achterwege gelaten. Theoretisch gezien echter is deze matrix uiterst belangrijk, omdat hij precies de correctie bevat die noodzakelijk is, omdat de parameters niet zijn geschat uit de gegevens die bevat zijn in een score × itemantwoord-contingentietabel, maar uit de oorspronkelijke data, die meer informatie bevatten. Bovendien is het uitrekenen van de matrix A_i in de praktijk een tijdrovend karwei, dat bij grote k zelfs niet goed meer uit te voeren is. Daarom stellen we ons vaak tevreden met een benaderende kwadratische vorm door de matrix A_i in (4.96) gewoon weg te laten. Deze benaderende kwadratische vorm kan geschreven worden als:

$$Q^*([T|Y_i]) = n\boldsymbol{d}_2'[Y_i' D_{\hat{\pi}} Y_i - \Delta_i]^{-1}\boldsymbol{d}_2$$

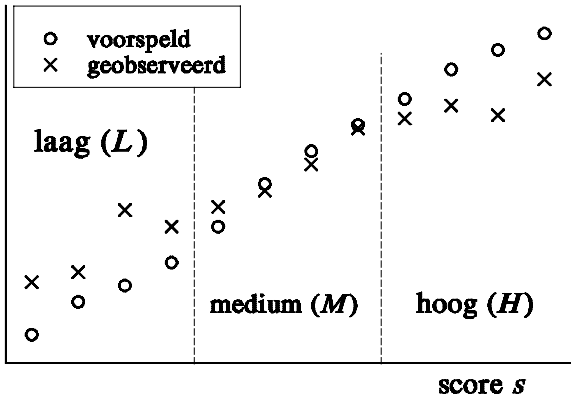
$$= \sum_{q=1}^{r} \frac{\left[\sum_{s \in G_q} n_s(p_{i|s} - \hat{\pi}_{i|s})\right]^2}{\sum_{s \in G_q} n_s \hat{\pi}_{i|s}(1 - \hat{\pi}_{i|s})} . \qquad (4.98)$$

De kwadratische vorm $Q([T|Y_i])$ is asymptotisch chi-kwadraat verdeeld met r-1 vrijheids- graden; van de benaderende vorm Q^* gegeven in (4.98) is de asymptotische verdeling niet bekend. Ervaring heeft echter geleerd dat beide grootheden heel vaak niet veel van elkaar afwijken, maar dat de vorm Q^* meestal een iets grotere uitkomst oplevert. Door Q^* te inter- preteren als een chi-kwadraat verdeelde variabele met r-1 vrijheidsgraden zal men dus de nulhypothese iets vaker verwerpen dan aangegeven door het nominale significantieniveau α.

In het vervolg zullen we de kwadratische vorm $Q([T|Y_i])$ aanduiden als S_i en de benaderende grootheid $Q^*([T|Y_i])$ als S_i^*.

Een nadere beschouwing van de teller in het rechterlid van (4.98) kan ons iets leren over het onderscheidend vermogen van de S_i-toetsen. De uitdrukking tussen vierkante haken is een som van afwijkingen tussen geobserveerde en verwachte frequenties. Deze afwijkingen kun- nen positief of negatief zijn. Indien nu binnen een scoregroep G_q zowel positieve als nega- tieve afwijkingen voorkomen, dan heffen die elkaar (ten dele) op. Doordat alleen hun som wordt gekwadrateerd is het dus mogelijk dat grote afwijkingen door dit compensatiemechanisme slechts een geringe bijdrage leveren aan de toetsingsgrootheid. Of er compensatie optreedt, is afhankelijk van de manier van groeperen in scoregroepen. In figuur 4.7 is een voorbeeld gegeven van een item dat slechter discrimineert dan door het Raschmodel is voorspeld.

De geobserveerde proporties, gezien als functie van de score, vertonen een vlakker verloop dan de voorspelde proporties. De verticale stippellijnen in de figuur geven aan dat er drie scoregroepen zijn, die zijn aangeduid als laag, medium en hoog. Omdat de modelafwijkingen systematisch zijn, zien we dat in de twee extreme groepen geen compensatie optreedt, de afwijkingen hebben allemaal hetzelfde teken; in de medium-groep echter zal de som van de afwijkingen nagenoeg nul zijn. Deze groep draagt dus weinig of niets bij aan de toetsingsgrootheid S_i. Hadden we de twee extreme groepen, laag en hoog, als één enkele groep behandeld, door de twee overeenkomstige kolommen in de matrix Y_i bij elkaar op te tellen, dan zou in deze gecombineerde groep ook cancellatie optreden, en de resulterende kwadratische vorm zou nauwelijks van nul verschillen.



Figuur 4.7
Een item dat slechter discrimineert dan voorspeld door het Raschmodel

Aan dit voorbeeld zien we dat het onderscheidend vermogen van de toets afhankelijk is van de manier waarop de scoregroepen gevormd worden en de bijbehorende Y matrix wordt geconstrueerd. Men zou nu kunnen denken dat maximaal onderscheidend vermogen bereikt kan worden door eerst een plaatje te construeren analoog aan figuur 4.7, en dan de groepsindeling te maken zodanig dat er geen cancellatie van positieve en negatieve afwijkingen optreedt binnen de scoregroepen. Of andersom, als men liever geen significantie heeft, de groepen zo maken dat er zoveel mogelijk cancellatie optreedt. Op zo'n manier echter wordt de toetsingsprocedure afhankelijk gemaakt van de data, of preciezer gezegd, van de afwijkingen tussen geobserveerde en voorspelde frequenties. Dus is de Y-matrix geen matrix van constanten maar een matrix van toevalsvariabelen waarvan de waarde van steekproef tot steekproef zal gaan verschillen. Maar in dat geval is de toetsingsgrootheid S_i niet meer chi-kwadraat verdeeld. In de

praktijk echter zal men er niet helemaal onderuit kunnen om de groepsindeling toch enigszins van de data te laten afhangen. De noemer van het rechterlid van (4.97) zal klein zijn indien voor alle scores in G_q de geobserveerde frequenties zeer klein zijn of de verwachte proporties $\hat{\pi}_{i/s}$ zeer dicht bij 0 of 1 liggen. Het is twijfelachtig of in zo'n geval de benadering door de chi-kwadraatverdeling nog wel gerechtvaardigd is. Door een andere groepsindeling te kiezen kan men die kleine noemers vermijden. Maar een groepsindeling 'op maat' vereist dat de data geconsulteerd worden. Hoewel een dergelijke handelwijze niet helemaal orthodox is, maakt ze de S_i-toetsen niet waardeloos. Immers om de groepsindeling zo te maken dat de noemer van (4.97) niet al te klein wordt, hoeven de afwijkingen tussen geobserveerde en verwachte proporties niet geconsulteerd te worden. In het programma OPLM (Verhelst, Glas & Verstralen, 1993) wordt de minimale waarde van de noemers in (4.97) op 5 gesteld.

In de literatuur zijn verschillende toetsingsgrootheden voorgesteld waarvan de formule erg veel lijkt op het rechterlid van (4.98). We noemen als voorbeelden Wright en Panchapakesan (1969), Bock (1972), Wright en Mead (1977), Elliott, Murray en Saunders (1977) en Yen (1981). Er zijn echter twee belangrijke punten waarop de toetsingsgrootheden van al deze auteurs verschillen van (4.98).

Het eerste is de wijze waarop de verwachte proporties worden uitgerekend. Wij gebruiken de conditionele kans gegeven de score, en deze kans is onafhankelijk van θ; bovengenoemde auteurs gebruiken echter allemaal een schatting die gebaseerd is op een schatter van θ, die bovendien gebaseerd is op een JML-procedure. Deze benadering heeft het schijnbare voordeel dat de toetsen dan ook gebruikt kunnen worden voor andere modellen dan het Raschmodel, zoals het twee- en drieparameter-logistische model, doch het bewijs dat de toetsingsgrootheden asymptotisch chi-kwadraat verdeeld zijn ontbreekt, en de bewering is waarschijnlijk ook onjuist. In ieder geval kan men voor het bewijs geen beroep doen op standaardresultaten uit de statistiek, want die vereisen allemaal schatters met bepaalde eigenschappen. Een van de eisen is consistentie van de parameterschatters. In het Raschmodel zijn JML-schatters niet consistent en voor het tweeparameter-logistische model is geen bewijs van consistentie gegeven. Afgezien hiervan hebben alle formules die door bovengenoemde auteurs worden gepresenteerd in de teller dezelfde gedaante als het rechterlid van (4.98).

Het tweede punt is dat de noemers nogal verschillen. Wright en Panchapakesan (1969) presenteren dezelfde noemer als in (4.98), doch hun toets is alleen ontworpen voor het Raschmodel waarbij scores niet worden gegroepeerd. De noemer van (4.98) is een som van varianties, waarbij elke term de variantie is van het aantal juiste antwoorden in de scoregroep met s juiste antwoorden. In de toets die Yen (1981) voorstelt, wordt deze som vervangen door de variantie van het aantal juiste items in de

groep, waarbij gedaan wordt alsof alle personen in de groep dezelfde kans op een juist antwoord hebben. Het effect hiervan is dat de noemer te groot wordt. Wright en Mead (1971) houden hier rekening mee, en voeren een correctiefactor in. Hun formule heeft in de noemer dezelfde gedaante als de noemer van (4.98). De meest afwijkende vorm komt voor in de formule die Elliott e.a. (1977) gebruiken: daar bevat de noemer geen varianties maar verwachte aantallen juiste antwoorden. Hun toetsingsgrootheid is te vergelijken met (4.94), en komt erop neer dat in termen van contingentietabellen de helft van de cellen ten onrechte niet meegeteld wordt. Hun toetsingsgrootheid is dan ook systematisch veel te klein. Een overzicht van al deze formules wordt gegeven door Yen (1981).

De M_i-toetsen

Stel dat we een item onderzoeken dat beter discrimineert dan het merendeel van de andere, en we construeren voor dit item een figuur analoog aan figuur 4.7, dan zullen we zien dat de geobserveerde proporties een steiler verloop vertonen dan de verwachte, maar de S_i-toets kan geen onderscheid maken tussen te grote en te kleine discriminatie, want in beide gevallen is de toetsingsgrootheid positief. Er kunnen natuurlijk nog andere afwijkingen optreden die niet zo'n systematisch patroon te zien geven, maar die, als ze voldoende groot zijn, ook een significant (positief) resultaat opleveren. Door een slimme constructie van de matrix Y_i kan onderscheid gemaakt worden tussen items die te weinig en die te veel discriminerend vermogen hebben. De scores worden opgedeeld in drie groepen, een laag-, een medium- en een hoog-groep, precies zoals in figuur 4.7 is aangegeven. De Y_i-matrix bestaat echter uit één enkele kolom, waar een 1 staat indien de score van het antwoordpatroon een juist antwoord bevat op item *i*, en de bijbehorende score tot de laag-groep behoort. In geval de score tot de hoog-groep behoort, vult men -1 in en voor de medium-groep komt overal 0 te staan. De kwadratische vorm $Q([T|Y_i])$ is asymptotisch chi-kwadraat verdeeld met één vrijheidsgraad. De 'vierkantswortel-met-teken', dat wil zeggen, de positieve vierkantswortel vermenigvuldigd met -1 indien de één-elements vector $\boldsymbol{d}_2$ negatief is, volgt dus de standaardnormale verdeling. De benaderende waarde van deze toetsingsgrootheid, gebaseerd op (4.98), is gegeven door

$$M_i^* = \frac{\sum_{s\in L} n_s(p_{i|s} - \hat{\pi}_{i|s}) - \sum_{s\in H} n_s(p_{i|s} - \hat{\pi}_{i|s})}{\left[\sum_{s\in L,H} n_s\hat{\pi}_{i|s}(1-\hat{\pi}_{i|s})\right]^{½}}, \tag{4.99}$$

waarin L en H verwijzen naar respectievelijk de laag- en de hoog-groep. Uit figuur 4.7 volgt duidelijk dat in die situatie de eerste som in de teller van (4.99) een positieve waarde zal aannemen, en de tweede som een negatieve waarde. Het verschil zal dus een positieve waarde krijgen, en omdat de noemer van (4.99) steeds positief is, krijgen we dus bij een te weinig discriminerend item een positieve uitkomst. Bij een te sterk discriminerend item zal de uitkomst negatief zijn.

Door de bovenstaande omschrijving liggen de M-toetsen echter niet eenduidig vast, omdat de begrippen laag-groep en hoog-groep niet nauwkeurig gedefinieerd zijn. In het programma OPLM worden drie varianten van de M-toetsen uitgerekend, waarbij drie verschillende definities van laag-groep en hoog-groep worden gehanteerd. De drie toetsingsgrootheden worden aangeduid als respectievelijk M_i, $M2_i$ en $M3_i$. De definities van de verschillende score groepen is als volgt:

M_i : $s \in$ L indien $\hat{\pi}_{i|s} \leq 0.4$ en $s \in$ H indien $\hat{\pi}_{i|s} \geq 0.6$;

$M2_i$: de scores worden in een laag-groep en een hoog-groep verdeeld zodanig dat $\sum_{s \in L} n_s \approx \sum_{s \in H} n_s \approx n/2$. De medium-groep is leeg. Het is niet steeds mogelijk dat precies de helft van de observaties in beide groepen valt, omdat alle antwoordpatronen met dezelfde score tot dezelfde groep moeten behoren;

$M3_i$: analoog aan de situatie bij $M2_i$, doch nu is de opdeling in drie groepen die elk ongeveer een derde van de observaties vertegenwoordigen.

Door Molenaar (1983) is een toets ontwikkeld die als een speciale variant van de hier besproken M-toetsen kan worden opgevat. In de inleiding van deze paragraaf hebben we gezien dat de matrix Y een willekeurige matrix is. Indien we in een bepaalde rij een 1 invullen, en in een andere rij 2, blijven de theoretische resultaten geldig. Alleen kennen we verschillende gewichten toe aan verschillende antwoordpatronen. Molenaar stelt voor de afwijkingen $n_s(p_{i|s} - \hat{\pi}_{i|s})$ te wegen met het omgekeerde van hun standaardafwijking. Op de plaatsen waar in de Y-vector voor de M_i-toetsen een 1 of -1 komt, plaatst Molenaar de grootheid $\pm[n_s \hat{\pi}_{i|s}(1 - \hat{\pi}_{i|s})]^{\frac{1}{2}}$, waarbij de positieve wortel genomen wordt voor de laag-groep en de negatieve voor de hoog-groep. De toetsingsgrootheid, door Molenaar U_i, genoemd is gegeven door

$$U_i = \frac{\sum_{s \in L} \frac{n_s(p_{i|s} - \hat{\pi}_{i|s})}{[n_s \hat{\pi}_{i|s}(1 - \hat{\pi}_{i|s})]^{\frac{1}{2}}} - \sum_{s \in H} \frac{n_s(p_{i|s} - \hat{\pi}_{i|s})}{[n_s \hat{\pi}_{i|s}(1 - \hat{\pi}_{i|s})]^{\frac{1}{2}}}}{(|L| + |H|)^{\frac{1}{2}}} \qquad (4.100)$$

waarin $|L|$ en $|H|$ het aantal verschillende scores is in respectievelijk de laag- en de hoog-groep. Het is niet moeilijk om aan te tonen dat U_i hetzelfde is als $Q^*(T|Y_i])$, met verschillende gewichten in de een-koloms matrix Y_i. De U_i-toetsen zijn geïmplemen-

teerd in het programma PML (Gustafsson, 1979, aanpassing door Molenaar, 1981). Voor deze U_i-toetsen wordt ook een andere definitie van de laag-groep en de hoog-groep gebruikt dan in de *M*-toetsen. De laag-groep bevat de 25% laagst scorende en de hoog-groep de 25% hoogst scorende observaties.

De R_{1c}-toets

Hoewel de S_i-toetsen allemaal asymptotisch chi-kwadraat verdeeld zijn, zijn ze niet onafhankelijk van elkaar. Dit betekent dat hun som niet chi-kwadraat verdeeld is. Bovendien moet men voorzichtig zijn bij de interpretatie van de S_i-toetsen. Indien het model geldig is, dan kan men verwachten dat ongeveer 100α% van de toetsen een significant resultaat zal opleveren bij toetsen op niveau α. Dit resultaat is niet exact, omdat de toetsen niet onafhankelijk zijn van elkaar. De kans dat een of meer toetsen significant zijn is echter behoorlijk groter dan het nominale significantieniveau α. Om een globale toets te construeren kan men de toetsingsprocedure van Hommel gebruiken die reeds werd besproken in paragraaf 4.3.4, of men kan gebruik maken van een globale toets die beschouwd kan worden als een combinatie van alle S_i-toetsen. Deze toets is de R_{1c}-toets die door Glas (1989) werd ontwikkeld.

De rationale van deze toets is uiterst eenvoudig: hij is niets anders dan de kwadratische vorm $Q(Y)$, gegeven door (4.93), waarbij $Y = [Y_1|Y_2|...|Y_k]$.

Het uitrekenen van deze kwadratische vorm is in het algemeen echter zeer ingewikkeld omdat de matrix $Y' D_{\hat{\pi}} Y$ niet langer diagonaal is. Dit is precies de reden waarom de S_i-toetsen niet onafhankelijk zijn van elkaar. Glas (1989) heeft aangetoond dat een belangrijke vereenvoudiging aangebracht kan worden indien de opdeling in scoregroepen G_q voor alle items dezelfde is. In tabel 4.10 zijn de drie Y_i-matrices afgebeeld voor een toets met drie items, waarbij echter de kolommen gepermuteerd zijn. Elke kolom draagt een dubbele index *iq*, waarbij de eerste index verwijst naar het item en de tweede naar de scoregroep. Er zijn ook maar zes rijen afgebeeld, omdat de antwoordpatronen (0 0 0) en (1 1 1) niets aan de toetsingsgrootheid bijdragen. Indien men de parameters schat met CML komt het weglaten van die antwoordpatronen overeen met het aannemen van een verzadigde multinomiale verdeling van de scorefrequenties voor de scores 1, 2,...,*k*-1. Blokken van de totale *Y*-matrix die volledig uit nullen bestaan zijn wit gelaten.

Het is gemakkelijk na te gaan dat de matrix $Y' D_{\hat{\pi}} Y$ in dit geval een blokdiagonale structuur heeft, waarbij elk blok betrekking heeft op één scoregroep. Bovendien is gemakkelijk in te zien dat de kolommen van de matrices T_1 en T_2 geschreven kunnen

worden als lineaire combinaties van de kolommen van Y. De i-de kolom van de matrix T_1 in tabel 4.8
is gegegeven door $Y_i1 + Y_{i2}$, de tweede kolom van T_2 is gegeven als $\Sigma_i Y_{i1}$ en de derde ko-

Tabel 4.10
De Y-matrix voor de R_{1c}-toets (k=3)

Y_{11}	Y_{21}	Y_{31}	Y_{12}	Y_{22}	Y_{32}
1	0	0			
0	1	0			
0	0	1			
			1	1	0
			1	0	1
			0	1	1

lom als $\Sigma_i Y_{i2}/2$. De eerste en de laatste kolom van T_2 kunnen buiten beschouwing worden gelaten omdat de patronen met score 0 en 3 verwijderd zijn. De matrix Y bevat dus de matrix T, als lineaire combinaties van zijn kolommen, en daarom is $Q(Y)$ asymptotisch chi-kwadraat verdeeld. Het aantal vrijheidsgraden is hier 3, en in het algemeen $k(r-1)$. De benaderende vorm $Q^*(Y)$, in dit geval aangeduid als R_{1c}^*, is een eenvoudige veralgemening van (4.98):

$$R_{1c}^* = Q^*(Y) = \sum_{q=1}^{r} \sum_{i=1}^{k} \frac{\left[\sum_{s \in G_q} n_s (p_{i|s} - \hat{\pi}_{i|s})\right]^2}{\sum_{s \in G_q} n_s \hat{\pi}_{i|s}(1 - \hat{\pi}_{i|s})}. \qquad (4.101)$$

Meestal is de benaderende vorm $Q^*(Y)$ groter dan de exacte vorm $Q(Y)$; de asymptotische verdeling is echter niet bekend. Uit een vergelijking van (4.98) en (4.101) is direct duidelijk dat, indien voor alle items dezelfde groepering is gebruikt, geldt dat

$$R_{1c}^* = \sum_i S_i^* .$$

In de literatuur is op verschillende plaatsen aan deze globale toets aandacht gegeven. Martin-Löf (1973) heeft een zogenaamde T-toets ontwikkeld, vanuit een iets andere rationale dan hier werd gebruikt (zie bijvoorbeeld Van den Wollenberg, 1979). Er kan echter aangetoond worden (Glas, 1981) dat Martin-Löfs T-toets equivalent is met de R_{1c}-toets. De R_{1c}-toets is geïmplementeerd in het programma OPLM, de T-toets wordt uitgerekend in het programma PML. Merk echter op dat beide toetsingsgrootheden, uitgerekend met dezelfde data niet noodzakelijkerwijze dezelfde uitkomst geven: de

uitkomst is natuurlijk afhankelijk van de wijze waarop de scores zijn gegroepeerd, en dit gebeurt in de twee programma's niet op identieke wijze.

Van den Wollenberg (1979, 1982) heeft de Q_1-toets voorgesteld. De toetsingsgrootheid Q_1 is een kleine modificatie van (4.101):

$$Q_1 = \frac{k-1}{k} R_{1c}^{*} \quad .$$

Uit simulatiestudies blijkt dat de verdeling van Q_1 goed te benaderen is door de chi-kwadraat verdeling.

Bij het gebruik van de R_{1c}-toets dient men aan twee zaken aandacht te geven. In de eerste plaats is dat de grootte van de noemer in (4.101). Door het feit dat voor de R_{1c}-toets dezelfde scoregroepering gebruikt wordt voor alle items, is het soms onvermijdelijk dat één of meer noemers in (4.101) zeer klein worden, waardoor sommige termen erg groot worden. In zo'n geval is het twijfelachtig of nog wel een beroep gedaan kan worden op de chi-kwadraat verdeling. Het tweede probleem betreft het gecombineerde gebruik van itemgerichte toetsen, bijvoorbeeld de S_i-toetsen, en een globale toets als R_{1c}. Het is mogelijk dat de R_{1c}-toets niet significant is, terwijl één of meer S_i-toetsen een zeer significant resultaat opleveren. De reden hiervoor is dat de R_{1c}-toets minder onderscheidend vermogen heeft dan de S_i-toetsen voor zeer specifieke modelschendingen. Men zou kunnen stellen dat de R_{1c}-toets een 'slecht' item niet opmerkt als het ingebed is in een toets waarvan de meeste items aan het Raschmodel voldoen. Omgekeerd is het ook mogelijk dat de modelschendingen niet zonder meer aan specifieke items kunnen worden toegeschreven, zodat de itemgerichte toetsen niet significant zijn, maar bijvoorbeeld in meerderheid een kleine overschrijdingskans hebben, bijvoorbeeld kleiner dan 0.5. In zo'n geval kan de 'niet zo schitterende prestatie' van de afzonderlijke S_i- toetsen gecombineerd worden in de R_{1c}-toets die wel tot significantie kan leiden. Daarom is het in de praktijk aan te raden itemgerichte toetsen en globale toetsen gecombineerd te gebruiken.

Van den Wollenberg (1979, 1982) heeft laten zien dat de R_{1c}- (of de Q_1-) toets niet erg geschikt is om schendingen van het unidimensionaliteitsaxioma te ontdekken. Een theoretisch eenvoudige generalisatie van de R_{1c}-toets, namelijk de R_{2c}-toets is wel gevoelig voor deze schendingen. De teller van (4.98) en (4.101) bevat zogenaamde-eerste-orde-afwijkingen $n_s(p_{i|s}-\hat{\pi}_{i|s})$. Nu kan ook een toetsingsgrootheid worden opgesteld die tweede-orde-afwijkingen onderzoekt: de proportie personen die zowel item *i* als item *j* juist beantwoordt, wordt vergeleken met de voorspelde proportie. Er wordt dus een vector $\boldsymbol{d}$ van afwijkingen opgesteld die als elementen de afwijkingen $n_s(p_{ij|s}-\hat{\pi}_{ij|s})$ heeft, voor alle scores s=2,...,k-2 en voor alle paren (i,j), i>j=1,...,k. De bijbehorende Y-matrix heeft dan $rk(k-1)/2$ kolommen, en voor grote k is de R_{2c}-

toetsingsgrootheid niet goed uit te rekenen. Details over de berekeningswijze kan men vinden in Glas (1989). Van den Wollenberg (1979, 1982) geeft een benaderende toetsingsgrootheid Q_2.

De R_0- en de R_{1m}-toetsen

De S_i-toetsen, de M_i-toetsen, en de R_{1c}-toets zijn allemaal toepasbaar indien de parameters geschat zijn met de CML-schattingsmethode. Gebruiken we echter MML, dan ligt de zaak heel wat gecompliceerder. Immers, MML is niet zomaar een methode, maar veronderstelt een ander model dan alleen maar het Raschmodel; er dient een hypothese toegevoegd te worden over de verdeling van de latente variabele θ. De combinatie van het Raschmodel en de verdeling van θ zorgt er voor dat het model als geheel niet meer tot de exponentiële familie behoort, en dat we voor de constructie van statistische toetsen niet zonder meer een beroep kunnen doen op de resultaten (1) en (2) die hiervoor werden gegeven.

Voor de normale verdeling geldt wel resultaat (1), namelijk dat $Q([T|Y])$ asymptotisch chi-kwadraat verdeeld is indien T is opgebouwd volgens de beschrijving die hiervoor werd gegeven. Het tweede resultaat, namelijk $(\boldsymbol{p}\text{-}\hat{\boldsymbol{\pi}})'T = \mathbf{0}$, geldt echter niet meer. Glas (1989) heeft in zijn onderzoekingen geconstateerd dat $(\boldsymbol{p}\text{-}\hat{\boldsymbol{\pi}})'T_1 = \mathbf{0}$, zonder dat hij evenwel deze gelijkheid in het algemeen kon bewijzen. Bij gebruik van MML is echter de vector $(\boldsymbol{p}\text{-}\hat{\boldsymbol{\pi}})'T_2 \neq \mathbf{0}$. Met behulp van tabel 4.8 is het gemakkelijk na te gaan dat $n\boldsymbol{p}'T_2$ niets anders is dan de $(k+1)$-vector met geobserveerde scorefrequenties $(n_0, n_1, \cdots, n_k)$, dus de vector $(\boldsymbol{p}\text{-}\hat{\boldsymbol{\pi}})'T_2$ geeft de afwijkingen aan tussen de geobserveerde en voorspelde proportie van elke score. Bij CML was de overeenkomst perfect door het invoeren van een verzadigd multinomiaal model met k parameters. Door de invoering van de veronderstelling van een normale verdeling van θ zal de overeenkomst niet meer perfect zijn. Als de hypothese van een normale verdeling echter juist is, moeten de afwijkingen toe te schrijven zijn aan de steekproeffout. Dus de grootheid

$$R_0 = Q([T_1|T_2]) \tag{4.102}$$

is asymptotisch chi-kwadraat verdeeld. Het aantal vrijheidsgraden is k-2. De R_0-toets is gevoelig voor schendingen van de normaliteitsassumptie.

De R_{1m}-toets wordt op precies dezelfde manier geconstrueerd als de R_{1c}-toets. De afwijkingen tussen voorspelde en geobserveerde proporties kunnen nu echter toegeschreven worden zowel aan schendingen van het Raschmodel, dus de combinatie van S_i-achtige toetsen, als aan schendingen van de assumptie van normaliteit van de

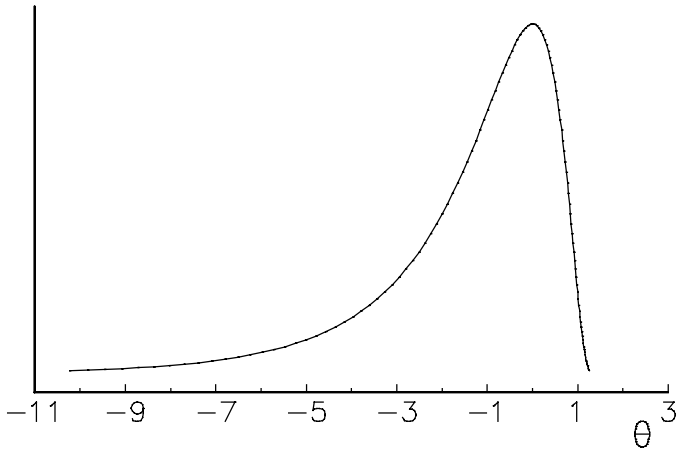
verdeling van theta. Het aantal vrijheidsgraden van R_{1m} bedraagt dan ook k-2 meer dan van de R_{1c}-toets: de k multinomiale parameters ω_s zijn niet meer nodig, doch worden vervangen door de twee parameters van de normale verdeling. De R_{1m}-toets kan echter geen onderscheid maken tussen die twee soorten schendingen. Een goede strategie is daarom, eerst de R_0-toets toe te passen en als er geen duidelijke schending is van de normaliteit gebruik te maken van de R_{1m}-toets. Men hoede zich echter voor een al te absolute interpretatie. Een significante R_{1m}-toets, samen met een niet significante R_0-toets is geen bewijs dat aan de assumptie van normaliteit is voldaan, en dat de modelschendingen dus bij het Raschmodel moeten liggen. Wil men deze twee assumpties duidelijk scheiden, dan verdient het de voorkeur de assumptie van normaliteit helemaal niet te maken, en CML als schattingsmethode te gebruiken.

4.3.6 Een voorbeeld

Als voorbeeld wordt een artificiële dataset geanalyseerd waarbij de itemantwoorden aan het Raschmodel voldoen, maar waarbij de verdeling van θ scheef is. De θ-waarden zijn gedefinieerd als

$$\theta = \frac{[\exp(-0.7\,z) - 1]}{-0.7}$$

waarbij z een aselecte trekking is uit de standaardnormale verdeling. De verdeling van θ is weergegeven in figuur 4.8, en wijkt dus sterk af van de normale verdeling. De toets bestaat uit 7 items met itemparameters (-1.5, -1, -0.5, 0, 0.5, 1, 1.5); n = 1000.



Figuur 4.8
Links scheve verdeling van θ

De schattingen en enkele statistische grootheden staan in tabel 4.11. De standaardfouten van de parameterschattingen zijn ongeveer 0.07. Vergeleken met deze grootte, verschillen CML- en MML-schattingen niet veel van elkaar.

Tabel 4.11

Schattingen en toetsingsgrootheden

item	$\hat{\beta}_i$ (CML)	$\hat{\beta}_i$ (MML)	S_i	vg	p	M_i	$M2_i$	$M3_i$
1	-1.460	-1.420	2.325	3	.508	0.76	1.50	0.84
2	-0.924	-0.933	0.817	3	.845	0.36	0.68	0.07
3	-0.506	-0.535	1.361	3	.715	-0.32	0.14	-0.15
4	0.053	0.021	2.853	3	.415	-0.94	-1.36	-0.97
5	0.394	0.371	1.255	3	.740	0.22	0.72	-0.04
6	0.964	0.972	7.288	3	.063	-2.60	-1.95	-1.65
7	1.480	1.526	6.752	3	.080	-2.52	0.61	2.30

R_{1c} = 19.17	vg = 18	p = .381
R_0 = 68.74	vg = 5	p < .00005
R_{1m} = 87.12	vg = 23	p < .00005

Voor de itemgerichte toetsen die in de tabel 4.11 zijn gerapporteerd is er niet veel reden om het model te verwerpen, hoewel voor de laatste twee items de overeenkomst met het model niet schitterend is. Vergelijken we dit echter met de uitkomsten van de *R*-toetsen, dan zien we dat de R_0- en de R_{1m}-toets zeer verschillende resultaten opleveren: de R_{1c}-toets, die niet beïnvloed wordt door de veronderstelling van de normale verdeling is niet significant. De conclusie is dus dat er geen reden is om het Raschmodel te verwerpen, maar een zeer overtuigende reden om deassumptie van een normale verdeling te verwerpen. In tabel 4.12 zijn de geobserveerde voorspelde scoreverdelingen weergegeven, waarbij het patroon van de afwijkingen niet erg duidelijk is. Het aantal geobserveerde nul-scores, bijvoorbeeld, is duidelijk groter dan verwacht, doch bij de daaropvolgende lage scores, 1 en 2, is de geobserveerde frequentie kleiner dan verwacht. Het patroon van afwijkingen tussen geobserveerde en voorspelde scorefrequenties hangt op een ingewikkelde manier af van de itemparameters en de verdeling van θ. In het algemeen is het niet mogelijk een duidelijke aanwijzing te krijgen over de onderliggende verdeling van θ door deze afwijkingen te bestuderen.

Tabel 4.12

Geobserveerde en verwachte scorefrequenties

score	geobs.	verwacht
0	98	61.3
1	94	131.1
2	147	180.2
3	188	197.4
4	212	180.9
5	176	137.6
6	72	81.3
7	13	29.7

Tenslotte zij er nog op gewezen dat, hoewel de assumptie van normaliteit op grove wijze geschonden is, de parameterschattingen met CML en MML erg goed op elkaar lijken. Het Raschmodel aangevuld met de normale verdeling voor θ is blijkbaar erg robuust tegen schendingen van de normaliteit. Men dient zich echter te hoeden voor klakkeloze generalisatie van dit resultaat. Een meer gedetailleerde studie is te vinden in Zwinderman (1991, hoofdstuk 4). In hoofdstuk 7 wordt een voorbeeld gegeven waarbij een verkeerde specificatie van de verdeling van θ leidt tot serieuze systematische fouten in de schatting van de itemparameters.

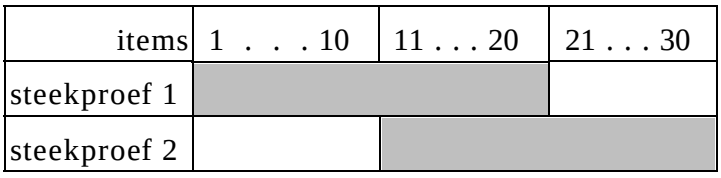
4.4 Het Raschmodel en onvolledige designs

In de vorige paragrafen is het Raschmodel uitvoerig besproken voor een situatie waarin alle personen uit de steekproef alle items beantwoorden. In de praktijk zal dit heel vaak niet het geval zijn, omdat sommigen door gebrek aan tijd de laatste items niet meer kunnen beantwoorden of omdat om een of andere reden bepaalde items worden overgeslagen. Het ontbreken van itemantwoorden in deze gevallen is dan afhankelijk van de persoon zelf die de items beantwoordt. De gaten die aldus in de data ontstaan zijn niet gepland. Analyse van zulke data is niet eenvoudig, en kan leiden tot systematische fouten in de parameter- schattingen, afhankelijk van de reden die tot het niet beantwoorden van bepaalde items heeft geleid. Als bijvoorbeeld items worden overgeslagen omdat ze moeilijk zijn, of er moeilijk uitzien, is het redelijk om aan te nemen dat de kans dat een item wordt overgeslagen groter is naarmate de vaardigheid waarop een beroep wordt gedaan lager is. In zo'n geval dient men uiterst voorzichtig te zijn met schattingsmethoden. Details hierover zijn het onderwerp van hoofdstuk 6.

Soms echter worden de gaten in de data gepland. Bij het construeren van een itembank van 1000 items zal het in de meeste gevallen om praktische redenen ondoenlijk zijn om alle personen alle items te laten beantwoorden. Daarom wordt aan elke persoon slechts een gedeelte van de items ter beantwoording voorgelegd volgens een vooropgezet design. In zo'n geval spreekt men van structureel onvolledige designs. De planning van een design kan echter verschillende vormen aannemen. Uitgaande van enige voorkennis over de moeilijkheidsgraad van de items zou een onderzoeker als volgt te werk kunnen gaan: aan de hand van een kleine voortoets van bijvoorbeeld 10 items die direct na afname nagekeken wordt, neemt men de beslissing voor de vervolgtoets. Personen met een lage score, zeg 5 of minder items juist, krijgen een relatief gemakkelijke natoets, de anderen een moeilijke natoets. Deze regel is eenduidig, maar er kan niet van te voren gezegd worden wie welke natoets zal krijgen. Het design staat dus onder de controle van degenen die de items beantwoorden. Daartegenover staat een design dat volledig van te voren is gepland. Bijvoorbeeld, de kinderen van school 1 krijgen toets 1, die van school 2 krijgen toets 2. Hier hebben de kinderen geen enkele controle op het design.

In deze paragraaf worden schattings- en toetsingsprocedures besproken die toepasbaar zijn in volledig door de onderzoeker gecontroleerde designs. De vraag welke procedures te gebruiken in andere gevallen, wordt in hoofdstuk 6 besproken.

In figuur 4.9 is een schematische weergave gegeven van een onvolledig design. De gearceerde oppervlakken stellen items voor die wel zijn aangeboden, de witte oppervlakken komen overeen met niet aangeboden items.



Figuur 4.9
Een onvolledig design met twee boekjes

Steekproef 1 heeft de items 1 tot 20 beantwoord en steekproef 2 de items 11 tot 30. Deze twee deelverzamelingen items worden doorgaans als een toetsboekje aangeboden, en om die reden zullen deelverzamelingen items die aan een groep personen worden aangeboden kortweg aangeduid worden als een boekje. Let wel dat in figuur 4.9 de boekjes elkaar overlappen.

In het algemeen zijn er B boekjes, en we definiëren de indexverzameling I_b $(b = 1, \cdots, B)$ als

$$I_b = \{i \mid \text{item } i \text{ komt voor in boekje } b\} \tag{4.103}$$

Het aantal items in boekje b wordt aangeduid als k_b. Het aantal personen dat boekje b heeft gekregen duiden we aan als n_b, en het aantal personen dat boekje b heeft gekregen en bovendien een score s (s = 0,...,k_b) heeft behaald, wordt aangeduid als n_{sb}. Een analoge notatie wordt ook gebruikt voor het aangeven van proporties en kansen. Zo betekent $p_{i|sb}$ de proportie juiste antwoorden op item i in de subgroep van personen die boekje b hebben gekregen en een score s hebben behaald.

Het totale aantal items dat in de analyse is betrokken duiden we aan met k. In figuur 4.9 geldt dus dat k=30. De antwoordvariabele X_i die bij volledige designs slechts twee waarden, 0 en 1, kon aannemen, laten we bij onvolledige designs drie waarden aannemen. We kennen X_i de waarde c toe indien het item niet is aangeboden, waarbij c een willekeurige waarde is die verschilt van 0 en 1. Voor een persoon met vaardigheid θ kunnen we nu twee conditionele kansverdelingen van X_i beschouwen, een voor het geval item i is aangeboden, en een voor het geval dat item i niet is aangeboden. Deze twee verdelingen zijn weergegeven in de rijen van tabel 4.12.

Tabel 4.12
Verdeling van X_i, conditioneel op θ en op het design

	$X_i = 0$	$X_i = 1$	$X_i = c$
aangeboden	$1 - f_i(\theta)$	$f_i(\theta)$	0
niet aangeboden	0	0	1

In de verdeling waarbij het item niet is aangeboden, kan X_i maar één waarde aannemen met een kans groter dan 0. In zo'n geval zegt men dat de verdeling van X_i gedegenereerd is. Formeel echter kunnen we de gewone algebra bedrijven met deze variabele en haar kans- verdeling.

Om expliciet aan te geven naar welke van de twee verdelingen we verwijzen voeren we de indicatorvariabelen D_{bi} in, die gedefinieerd zijn als

$$D_{bi} = \begin{cases} 1 \text{ indien } i \in I_b \\ 0 \text{ indien } i \notin I_b. \end{cases}$$

Eerst wordt de CML-schattingsprocedure besproken. Om het model te kunnen schrijven als een multinomiaal model moeten we de designvariabelen D_{bi} als toevalsvariabelen beschouwen. Dit kunnen we doen door voor de verschillende boekjes

een verzadigd multinomiaal model te beschouwen met parameters ω_b, de kans dat boekje b wordt aange-
boden. De ML-schatter van deze parameters is gegeven door

$$\hat{\omega}_b = \frac{n_b}{n}, \quad (b = 1, \cdots, B). \tag{4.104}$$

De multinomiale kans op een antwoordpatroon $\boldsymbol{x}$ is dan gegeven door

$$\begin{aligned} P(\boldsymbol{x}) &= P(\boldsymbol{x}|s,b)\, P(s,b) \\ &= P(\boldsymbol{x}|s,b)\, P(s|b)\, P(b) \\ &= \pi_{\boldsymbol{x}|sb}\, \omega_{s|b}\, \omega_b , \end{aligned} \tag{4.105}$$

waarbij de laatste regel niets anders is dan een verkorte notatie van de regel erboven. Voor de verdeling van de scores binnen een boekje nemen we, net als in het geval van een volledig design, een verzadigd multinomiaal model aan. De ML-schatters van de parameters van dit model zijn dus gegeven door

$$\hat{\omega}_{sb} = \frac{n_{sb}}{n_b}. \tag{4.106}$$

Gebruik makend van (4.104) en (4.106) zien we dus dat in (4.105) alleen de factor $\pi_{\boldsymbol{x}|sb}$ afhangt van de itemparameters, maar ook dat de conditie niet louter en alleen de score s is, maar de combinatie (s,b). Verzamelen we nu de itemparameters van alle items die behoren tot boekje b in de vector $\boldsymbol{\varepsilon}_b$, dan is $\pi_{\boldsymbol{x}|sb}$ gegeven door

$$\pi_{\boldsymbol{x}|sb} = \frac{\prod_{i=1}^{k} \varepsilon_i^{d_{bi}x_i}}{\gamma_s(\boldsymbol{\varepsilon}_b)} = \frac{\prod_{i\in I_b} \varepsilon_i^{x_i}}{\gamma_s(\boldsymbol{\varepsilon}_b)}. \tag{4.107}$$

De middelste uitdrukking in (4.107) geeft duidelijk aan hoe, door gebruik te maken van de waarde d_{bi} alle k antwoordvariabelen in de kansuitdrukking kunnen worden opgenomen, terwijl het rechterlid overeenkomt met het rechterlid van (4.40): het is gewoon de conditio- nele kans op het antwoordpatroon gegeven de score, maar beperkt tot de items die zijn aangeboden. Omdat in de totale steekproef alle antwoordpatronen onafhankelijk zijn van elkaar, is de aannemelijkheidsfunctie het produkt van

uitdrukkingen zoals het rechterlid van (4.107), en de log-aannemelijkheidsfunctie is de som van hun logaritmen.

Als dat duidelijk is, ligt de afleiding van de schattingsvergelijkingen, de uitdrukkingen voor de informatiematrix en de toetsingsgrootheden S_i^*, M_i^* en R_{1c}^* voor de hand. We geven ze hier volledigheidshalve, een gedetailleerde afleiding kan men vinden in Verhelst en Eggen (1989) en in Glas (1989).

De schattingsvergelijkingen zijn gegeven door

$$t_i = \sum_{b:i\in \mathrm{I}_b} \sum_{s=0}^{k_b} n_{sb} \frac{\varepsilon_i \gamma_{s-1}(\boldsymbol{\varepsilon}_b)}{\gamma_s(\boldsymbol{\varepsilon}_b)}, \tag{4.108}$$

waarin t_i het totaal aantal juiste antwoorden is dat op item i is uitgebracht.

De uitdrukkingen voor de informatiematrix zijn een veralgemening van (4.48):

$$I_{ij}(\beta) = \begin{cases} \sum_{b:i\in I_b} \sum_{s}^{k_b} n_{sb}[\pi_{i|s}(1-\pi_{i|s})] & \text{indien } i=j, \\ \sum_{b:i,j\in I_b} \sum_{s}^{k_b} n_{sb}[\pi_{ij|s} - \pi_{i|s}\pi_{j|s_v}] & \text{indien } i\neq j. \end{cases} \tag{4.109}$$

Voor de S_i-toetsen verandert er heel weinig. Het enige dat aangepast moet worden is de groepering van scores in scoregroepen G_q. Bij een volledig design konden we volstaan met het groeperen van scores; hier moeten de combinaties (s,b) gegroepeerd worden. De manier van groeperen is bepalend voor het onderscheidend vermogen van de toets tegen bepaalde schendingen van het model. Een concreet voorbeeld hiervan zal besproken worden in hoofdstuk 9 bij de discussie over itemonzuiverheid. De formule voor de benaderende grootheid S_i^* komt dan neer op een eenvoudige verandering van (4.98):

$$S_i^* = \sum_{q=1}^{r} \frac{\left[\sum_{(s,b)\in G_q} n_{sb}(p_{i|sb} - \hat{\pi}_{i|sb})\right]^2}{\sum_{(s,b)\in G_q} n_{sb}\hat{\pi}_{i|sb}(1-\hat{\pi}_{i|sb})}. \tag{4.110}$$

Voor de M-toetsen geldt precies hetzelfde: alle (s,b) combinaties worden opgedeeld in een laag- een midden- en een hoog-groep. Om die combinaties te ordenen moeten we echter beschikken over een ordeningsprincipe, dit wil zeggen we moeten een rationele methode vinden om alle combinaties (s,b) een rangnummer $w(s,b)$ te geven. In het programma OPLM worden de rangnummers zo toegekend dat

$$w(s,b) < w(s',b') \text{ indien } \hat{\pi}_{i|sb} < \hat{\pi}_{i|s'b'} \ . \tag{4.111}$$

Indien de twee geschatte kansen aan elkaar gelijk zijn beslist het toeval over de nummering. Op deze manier kunnen scores geordend worden, ook als ze afkomstig zijn van verschillende boekjes.

Bij de veralgemening van de R_{1c}-toets tot onvolledige designs treedt er een complicatie op. In paragraaf 4.3.5 werd gezegd dat de opdeling in scoregroepen voor alle items dezelfde moet zijn, omdat anders de Y matrix van de kwadratische vorm niet teruggebracht kan worden tot een blokdiagonale structuur. Bij onvolledige designs kan deze gelijke opdeling natuurlijk niet, want het ordeningsprincipe (4.111) is zinloos indien item i niet voorkomt in boekje b of b'. Daarom wordt een opdeling gemaakt per boekje in r_b scoregroepen G_{bq}, $(q=1,...,r_b)$, en de veralgemening van (4.101) is dan gegeven door

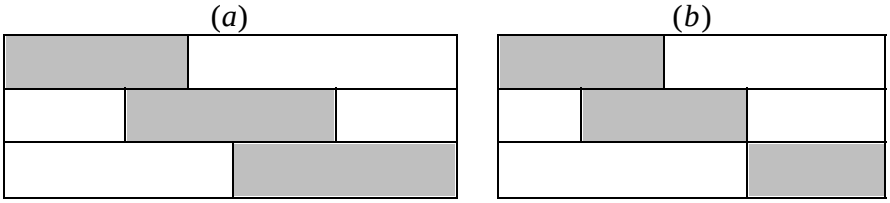
$$R_{1c}^{*} = \sum_{b} \sum_{q=1}^{r_b} \sum_{i \in I_b} \frac{\left[\sum_{s \in G_{bq}} n_s (p_{i|sb} - \hat{\pi}_{i|sb}) \right]^2}{\sum_{s \in G_{bq}} n_s \hat{\pi}_{i|sb} (1 - \hat{\pi}_{i|sb})} \ . \tag{4.112}$$

Het aantal vrijheidsgraden is gegeven door

$$\sum_{b=1}^{B} \left[r_b (k_b - 1) \right] - (k - 1) \ .$$

Hoewel de technische aspecten van het schatten van de parameters eigenlijk alleen neerkomen op iets meer gecompliceerde formules, waar een gebruiker bij zijn eigen toepassingen niet veel last van heeft, als programmatuur gebruikt wordt waar deze formules in zijn geïmplementeerd, is er een ander probleem waarmee bij het plannen van onderzoek terdege rekening moet worden gehouden. In figuur 4.9 zijn twee boekjes afgebeeld die overlappen. In zo'n geval zal men zeggen dat het design verbonden is. Bij ingewikkelder designs is de definitie van verbondenheid iets ingewikkelder. In figuur 4.10 zijn twee designs afgebeeld met elk drie boekjes. Het design (a) is verbonden,

hoewel boekje 1 en boekje 3 geen gemeenschappelijke items hebben, maar boekje 1 vertoont overlap met boekje 2, en boekje 2 heeft overlap met boekje 3, hoewel er geen enkel item is dat in alle drie de boekjes voorkomt. Het design (*b*) is niet verbonden want boekje 3 heeft geen enkele overlap met boekje 1 of boekje 2.



Figuur 4.10
Een verbonden (*a*) en een niet-verbonden design (*b*)

In een niet-verbonden design bestaan geen unieke CML-schatters van de itemparameters. Dit hoeft ook geen verwondering te wekken, omdat het nu eenmaal onmogelijk is om de relatieve moeilijkheid van twee items te schatten als niemand beide items heeft beantwoord. Willen we toch gegevens die verzameld zijn onder design (*b*) in figuur 4.10 met het Raschmodel analyseren, dan kan dat alleen door een MML-procedure te gebruiken.

Bij de MML-schattingsprocedure hebben we iets meer vrijheid om de verdeling van θ te specificeren dan bij volledige designs. In het design gegeven in figuur 4.9 bijvoorbeeld zou het kunnen zijn dat de twee steekproeven aselect uit dezelfde populatie zijn getrokken. In dat geval moeten naast de itemparameters de twee parameters van die gemeenschappelijke verdeling worden geschat. Het zou echter ook kunnen dat die twee steekproeven uit twee verschillende populaties zijn getrokken. Dan moeten, behalve de itemparameters, ook twee gemiddelden en twee varianties worden geschat. Voor het design (*a*) uit figuur 4.10 hebben we nog meer mogelijkheden: we kunnen een enkele verdeling veronderstellen, of twee of drie. Bij twee verdelingen zijn twee van de drie steekproeven afkomstig uit dezelfde populatie. In het algemeen kunnen we dus A populaties of verdelingen beschouwen, en uit elke populatie hebben we een of meer steekproeven die een boekje voorgelegd krijgen. Dus $A \leq B$, en er moeten $2A$ populatieparameters geschat worden: μ_a en σ_a^2, $(a = 1,\ldots,A)$. De log-aannemelijkheidsfunctie is dan een voor de hand liggende veralgemening van (4.58)

$$\ln L(\beta,\boldsymbol{\mu},\boldsymbol{\sigma^2};\boldsymbol{X}) = \sum_{b=1}^{B}\sum_{v=1}^{n_b} \ln \int_{-\infty}^{+\infty} P(\boldsymbol{x}_v|\theta)\,\frac{1}{\sqrt{2\pi\sigma_a^2}}\exp\left[-\frac{(\theta-\mu_a)^2}{2\sigma_a^2}\right]\mathrm{d}\theta\,, \qquad (4.113)$$

waarin $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_A)$ en $\sigma^2 = (\sigma_1^2, \ldots, \sigma_A^2)$. De index *a* in (4.113) dient begrepen te worden als een functie van het boekjesnummer en dient dus gelezen te worden als $a(b)$, de populatie waaruit de steekproef, die boekje *b* heeft gekregen, afkomstig is.

Bij niet-verbonden designs is men niet helemaal vrij om steekproeven aan verschillende populaties toe te wijzen. In design (*b*) van figuur 4.10, bijvoorbeeld, kan men wel een analyse uitvoeren met de hypothese van één of twee verschillende populaties, maar in de tweede geval kan men niet veronderstellen dat steekproef 1 en 2 afkomstig zijn uit dezelfde populatie en steekproef 3 uit een andere. Veronderstelt men echter dat steekproef 1 en steekproef 3 uit dezelfde populatie komen, dan zijn alle parameters in principe wel schatbaar, omdat de items uit die twee boekjes met elkaar verbonden worden door een gemeenschappelijke verdeling.

Tot slot van deze paragraaf, nog een opmerking over schatbaarheid van parameters in het algemeen. Als gezegd wordt dat voor het design in figuur 4.9 CML-schatters bestaan, dan betekent dit niet dat in alle gevallen waar dit design wordt toegepast CML-schattingen kunnen worden gevonden. Het zou bijvoorbeeld kunnen voorkomen dat in een bepaalde steekproef een item door iedereen juist beantwoord is. In zo'n geval bestaat er geen eindige CML-schatting voor de parameter van dit item. Bij onvolledige designs zijn de voorwaarden waar- onder eindige en unieke CML-schattingen van de parameters bestaan echter veel ingewikkelder dan het voorbeeldje hiervoor suggereert. Algemene voorwaarden, die ook redelijk gemakkelijk met de computer kunnen gecontroleerd worden, zijn gegeven in Fischer (1981) en worden in hoofdstuk 6 besproken. Voor het bestaan van MML-schattingen zijn de algemene voorwaarden niet precies bekend. In het algemeen zijn die voorwaarden echter milder dan voor CML-schattingen: als CML-schattingen bestaan, bestaan ook MML-schattingen; maar MML-schattingen kunnen ook bestaan waar CML onmogelijk is. Design (*b*) uit figuur 4.10 is daar een voorbeeld van.

4.5 Het schatten van de persoonsparameters

Het uiteindelijke doel bij het ontwikkelen van een meetinstrument is het meten van eigenschappen van objecten of personen, dat wil zeggen het toekennen van getallen aan die objecten of personen zodanig dat de toegekende getallen ook de mate van aanwezigheid van de bedoelde eigenschap aangeven. In de context van het Raschmodel betekent dit de waarde van θ 'berekenen' voor een willekeurige persoon. De observaties die we nodig hebben, zijn de itemantwoorden van die persoon. De waarde van θ is dus een functie van de itemantwoorden. Als we een toets tweemaal afnemen

aan dezelfde persoon, zullen de item-antwoorden niet tweemaal dezelfde zijn. Itemantwoorden zijn dus toevalsvariabelen, en bijgevolg is de waarde van θ die we uit deze antwoorden berekenen ook een toevalsvariabele. Vergelijk met lichaamslengte: de observatie die we nodig hebben om lichaamslengte te bepalen is iemands verticale uitgestrektheid en die varieert ook: na een dag vol activiteiten is iemands verticale uitgestrektheid minder dan na een nacht slaap. Het is dus niet zonder meer duidelijk wat bedoeld wordt met lichaamslengte: ook als we de observatie-omstandigheden standaardiseren (bijvoorbeeld altijd 's morgens na minstens zes uur rust), zullen de meetuitslagen variabiliteit vertonen, en als we slechts een keer meten, weten we niet of we een 'lage' dan wel een 'hoge' uitkomst hebben. Meestal maken we ons echter niet druk over dit probleem omdat voor de praktische bedoelingen waar we deze uitkomsten voor nodig hebben, de variabiliteit van de uitkomsten te verwaarlozen is. Bij het meten van schoolse of cognitieve vaardigheden met de meetinstrumenten waarover we beschikken, is die variabiliteit meestal niet te verwaarlozen. We zullen er dus enige aandacht aan moeten besteden.

Er zijn bovendien nog twee overwegingen van technische aard waar men rekening mee moet houden bij de interpretatie van de berekende θ-waarde, namelijk de normalisering van de itemparameters en de toegepaste rekenregel. We illustreren beide wederom aan de hand van het voorbeeld over lichaamslengte.

Gewoonlijk bedoelen we met lichaamslengte de afstand tussen iemands voetzolen en kruin bij gestrekte houding. De eenheid waarin we meten wordt gewoonlijk toegevoegd aan de meetuitslag. Zo spreken we van een lichaamslengte van 176 cm of 69 inch. Bij het meten van vaardigheden worden meestal geen eenheden toegevoegd, doch zoals uiteengezet in paragraaf 4.3.1 is er wel degelijk van een eenheid sprake die we kunnen kiezen: de waarde van de gemeenschappelijke discriminatieparameter is willekeurig en bepaalt de eenheid waarin we meten. Als twee meetuitslagen met elkaar worden vergeleken, dienen we er dus zeker van te zijn dat ze in dezelfde eenheid zijn uitgedrukt. Een analoog argument geldt ook voor het nulpunt van de schaal. We zouden iemands lichaamslengte ook kunnen definiëren als de afwijking tot het populatiegemiddelde of het aantal centimeters dat hij in rechtopstaande houding uitsteekt boven een tafel van één meter hoog. Het nulpunt van de schaal wordt bepaald door wat we de normalisatie genoemd hebben. Twee meetuitslagen zijn dus alleen zinvol te vergelijken als ze afkomstig zijn van twee meetinstrumenten met hetzelfde nulpunt en dezelfde eenheid.

Het belang van de rekenregel kan als volgt geïllustreerd worden voor het voorbeeld van de lichaamslengte. Voor het bepalen van iemands lichaamslengte laten we tien beoordelaars een 'schatting-op-zicht' van de lichaamslengte maken. Als eerste

rekenregel nemen we het gemiddelde van de tien schattingen. Bij de tweede rekenregel verwijderen we eerst de hoogste en de laagste schatting en we nemen als uitkomst het gemiddelde van de acht overblijvende schattingen. Het is duidelijk dat we bij het bepalen van iemands lichaamslengte volgens de twee rekenregels, in het algemeen twee verschillende uitkomsten zullen krijgen. Bovendien is het niet meteen duidelijk welke de beste regel is: de eerste regel is iets nauwkeuriger dan de tweede omdat hij gebaseerd is op tien schattingen en de tweede slechts op acht. Daartegenover staat echter dat de tweede regel beter beschermd is tegen grove vergissingen van de beoordelaars. Voor de schattingen van de vaardigheden hebben we ook verschillende rekenregels, die verschillende uitkomsten geven. Welke rekenregel we moeten kiezen is afhankelijk van het gebruik van de meetresultaten. Omdat hieraan soms serieuze ethische implicaties verbonden zijn, zullen we tamelijk uitvoerig op deze regels ingaan.

In paragraaf 4.5.1 worden de verschillende rekenregels besproken. Omdat elke regel een schatting van θ geeft worden die regels gewoonlijk aangeduid als schattingsmethode. Paragraaf 4.5.2 behandelt een voorbeeld.

Bij de bespreking van de veronderstellingen die aan het Raschmodel ten grondslag liggen, is er op gewezen dat homogeniteit met betrekking tot het Raschmodel wordt verondersteld. Dit betekent dat er van uit gegaan wordt dat het Raschmodel voor iedere persoon in de steekproef geldt, of, indien er schendingen zijn van de axioma's, dat die schendingen in gelijke mate voor iedere persoon gelden. Nu is het natuurlijk mogelijk dat het Raschmodel geldt voor de overgrote meerderheid van de personen in de steekproef, maar voor een enkeling niet. In zo'n geval is het goed mogelijk dat dit gebrek aan homogeniteit niet ontdekt wordt door de statistische toetsen die in paragraaf 4.3 werden besproken. Door individuele antwoordpatronen nader te onderzoeken kan men soms overtuigende evidentie vinden dat in individuele gevallen het Raschmodel als nulhypothese verworpen moet worden. Dit is het onderwerp van paragraaf 4.5.3.

4.5.1 Drie methoden om de persoonsparameter θ te schatten

De drie methoden die we hier bespreken, worden aangeduid als ML, Warm of WML en EAP, en staan respectievelijk voor 'Maximum Likelihood', 'Weighted Maximum Likelihood' en 'Expected A Posteriori'. The WML-methode is ontwikkeld door Warm (1989). Vooraleer we de verschillende methoden uiteenzetten, is het belangrijk te wijzen op een overeenkomst in de drie methoden. Om θ te schatten, moeten we de waarde van de itemparameters kennen. In de praktijk kennen we die natuurlijk nooit, en daarom

gebruiken we geschatte waarden. Bij het schatten van θ wordt gedaan alsof die geschatte waarden van de itemparameters de echte waarden zijn. Daarmee wordt dus een extra fout geïntroduceerd in de schatting van θ. Hoe erg die fout is, hangt af van de standaardfout van de itemparameterschattingen, en deze hangt op haar beurt weer in belangrijke mate af van de grootte van de calibratiesteekproef. In het gebruik wordt echter zelden met die fout rekening gehouden, er wordt gedaan alsof die fout er niet is, waardoor de nauwkeurigheid van de θ-schatting doorgaans overschat wordt. Het precieze onderzoek naar de invloed van die schattingsfout op de nauwkeurigheid van de schatting van θ is nogal moeilijk, en wordt hier verder niet besproken.

De ML-schatter van θ

Indien de itemparameters bekend zijn, en we observeren één antwoordpatroon $\boldsymbol{x}$, dan is de logaritme van de aannemelijkheidsfunctie gegeven als een speciaal geval van (4.28):

$$\ln L(\theta;\boldsymbol{x},\beta) = s\theta + \sum_{i=1}^{k} x_i(-\beta_i) - \sum_{i=1}^{k} \ln[1+\exp(\theta-\beta_i)], \qquad (4.114)$$

waarin $s = \Sigma_i x_i$ de score is. Merk op dat in (4.114) de itemparameters β_i als constanten worden behandeld: de tweede term in het rechterlid is dus uitsluitend een functie van de data. De derde term is alleen functie van de parameter θ, zodat duidelijk is dat (4.114) de gedaante heeft van een log-aannemelijkheidsfunctie in de exponentiële familie. De schattings- vergelijking is dus onmiddellijk gegeven door

$$s = \mathscr{E}(S) = \sum_{i=1}^{k} \mathscr{E}(X_i) = \sum_{i=1}^{k} f_i(\theta)\ . \qquad (4.115)$$

Hoewel de formule erg eenvoudig is, is voor het berekenen van de waarde van θ een iteratieve procedure nodig; een expliciete oplossing bestaat niet. De meeste computer-programmatuur geeft de oplossingen echter standaard. Vergelijking (4.115) heeft echter niet altijd een oplossing. Omdat $0 < f_i(\theta) < 1$ is het rechterlid van (4.115) altijd groter is dan 0 en altijd kleiner dan de maximale toetsscore k. Voor de scores 0 en k is er dus geen enkele waarde van θ waarvoor aan (4.115) voldaan is. Voor alle andere scores bestaat de ML-schatting wel. Men dient dus voorzichtig te zijn bij het berekenen van

steekproefgrootheden, zoals de gemiddelde ML-schatting. Het invullen van een willekeurige lage θ-waarde voor personen met een nul-score en een willekeurige hoge waarde in geval van perfecte scores is uit den boze. Wil men toch per se een gemiddelde berekenen, dan zit er niets anders op dan personen met zulke extreme scores uit de steekproef te verwijderen, maar daardoor kunnen groepsvergelijkingen onzuiver gaan worden. Stel dat in een steekproef 5% perfecte scores voorkomen. Hoewel er geen ML-schattingen bestaan voor die 5%, weten we toch dat we de vaardigheid van die personen hoog moeten inschatten. Door ze te verwijderen gaan we de gemiddelde vaardigheid in die steekproef, en bij veralgemening dus ook in de geassocieerde populatie, onderschatten. Komen in een andere steekproef (uit een andere populatie) slechts 2% perfecte scores voor, dan treedt er ook een onderschatting op, maar die is minder erg. De twee berekende gemiddelden kunnen dan niet zinvol met elkaar worden vergeleken.

De nauwkeurigheid waarmee θ gemeten wordt is de nauwkeurigheid waarmee θ geschat wordt en deze kan, zoals in paragraaf 4.2.1 werd uiteengezet, worden afgeleid uit de informatiefunctie, die hier de naam toetsinformatiefunctie draagt:

$$I(\theta) = \sum_{i=1}^{k} f_i(\theta)[1-f_i(\theta)]. \qquad (4.116)$$

Het produkt $f_i(\theta)[1-f_i(\theta)]$ bereikt zijn grootste waarde indien $f_i(\theta) = 0.5$, en dit is het geval indien $\theta = \beta_i$. Dit produkt wordt kleiner naarmate θ verder afwijkt van β_i. Vullen we nu in (4.116) een waarde in die ver afligt van alle β's, dan blijkt dat de toets zeer weinig informatie oplevert over die θ. Indien de waarde van θ middenin tussen de β's is gelegen, levert de toets meer informatie op over θ. Een toets kan dus voor bepaalde personen zeer informatief zijn, en voor andere niet. Deze geschiktheid wordt ook weerspiegeld in de standaardfout van de schatting van θ:

$$SE(\hat{\theta}) \approx \sqrt{1/I(\theta)}\ . \qquad (4.117)$$

Om (4.117) te evalueren moet men θ kennen. In een concrete toepassing waarbij men θ gewoonlijk niet kent, vult men in het rechterlid de ML-schatting van θ in. Het resultaat is natuurlijk een schatting van de standaardfout. Bovendien zijn rechter- en linkerlid van (4.117) slechts asymptotisch aan elkaar gelijk, dus indien $k \to \infty$. In toepassingen met een klein aantal items moet er rekening mee worden gehouden dat gebruik van (4.117) een forse onderschatting van de standaardfout kan opleveren.

De ML-schatter van θ heeft nog een tweede nadeel naast het feit dat hij niet bestaat voor perfecte en nulscores. Hij is namelijk zeer onzuiver. Het begrip zuiverheid dient

als volgt opgevat te worden. Stel dat een persoon met een bepaalde waarde θ een gegeven toets een zeer groot aantal keren maakt, in de veronderstelling van volledige 'brain wash' na elke afname, dan verwachten we niet dat hij telkens dezelfde score haalt. We zullen dus een verdeling van scores vinden. Als we even de gevallen waarin hij 0 of een perfecte score haalt buiten beschouwing laten, kunnen we voor elke score de ML-schatting berekenen. We beschikken dus ook over de verdeling van ML-schattingen. Een schatter heet zuiver als het gemiddelde van die verdeling gelijk is aan de echte θ-waarde. De afwijking tussen het gemiddelde van die verdeling en de echte waarde wordt de onzuiverheid of bias genoemd: bias = $\mathscr{E}(\hat{\theta}|\theta) - \theta$. De ML-schattingen zijn onzuiver in een heel speciale zin. Voor kleine waarden van θ is de onzuiverheid negatief en voor grote waarden positief. Wat precies bedoeld wordt met groot en klein is nogal ingewikkeld, doch in grote lijnen komt het op het volgende neer: meestal is de toetsinformatiefunctie ééntoppig, dat wil zeggen dat de infor- matie heel klein is voor zeer kleine waarden van θ, toeneemt tot een bepaalde θ-waarde, zeg θ_0, en vanaf daar weer afneemt. Met klein wordt nu grofweg bedoeld kleiner dan θ_0, en met groot, groter dan θ_0. Bovendien neemt de onzuiverheid toe naarmate θ meer van θ_0 afwijkt. Het effect van die onzuiverheid is dus als het ware een uitrekken van de schaal van geschatte θ's in vergelijking met de schaal van de echte θ's (zie Lord, 1983a, voor een gedetailleerde uiteenzetting).

Samenvattend: de ML-schatter van θ bestaat niet voor perfecte en nulscores, en is behoorlijk onzuiver. Dit zijn voldoende redenen om die schatter niet te gebruiken. Hij is in de literatuur vrij lang gebruikt omdat er geen goed alternatief was. Warm heeft in 1989 een θ-schatter ontwikkeld die beide euvels verhelpt. Die schatter wordt in de volgende paragraaf besproken.

De WML-schatter van θ (Warm-schatter)

Warm (1989) heeft aangetoond dat de onzuiverheid van de θ-schatter grotendeels kan worden opgeheven door niet de aannemelijkheidsfunctie te maximaliseren, maar een gewogen aannemelijkheidsfunctie. (WML staat voor Weighted Maximum Likelihood.) In het Raschmodel is deze weegfunctie de vierkantswortel uit de informatiefunctie. De WML-schatting van θ is dus die waarde van θ die de functie

$$W(\theta) = L(\theta)\sqrt{I(\theta)} \qquad (4.118)$$

maximaliseert.

De WML-schatter vertoont bijna geen onzuiverheid meer, tenzij voor zeer extreme θ-waarden. De overblijvende onzuiverheid vertoont daarenboven het omgekeerde beeld van de onzuiverheid voor de ML-schatter. Voor zeer kleine waarden van θ is de onzuiverheid positief, en voor zeer grote waarden negatief. De schaal van de geschatte θ's (met WML) vertoont dus een zekere krimping in vergelijking met de echte θ-waarden.

Een gelukkige bijkomstigheid van de WML-schatter is dat hij altijd bestaat, ook voor perfecte en nulscores.

De WML-schatter, samen met een schatting van de standaardfout en een schatting van de bias, wordt berekend in het programmapakket OPLM. De formule voor de standaardfout is ingewikkelder dan in het geval van de ML-schatter en wordt hier niet besproken.

De EAP-schatter van θ

Bij de ML- en de WML-schatter wordt alleen gebruik gemaakt van het geobserveerde antwoordpatroon om θ te schatten. Twee personen met dezelfde score behalen steeds dezelfde schatting van θ. Men zou echter ook andere informatie kunnen gebruiken om θ te schatten, bijvoorbeeld kennis omtrent de populatie waaruit de betrokken persoon afkomstig is. Dit is wat er gebeurt bij de EAP-schatter: daarin wordt informatie die men heeft over de populatie waaruit de betrokken persoon afkomstig is, gecombineerd met informatie die het antwoordpatroon oplevert. Deze combinatie levert in de regel een uitkomst op die ligt tussen de ML-schatting en het populatiegemiddelde. Bijvoorbeeld, stel dat men weet dat een persoon aselect uit een θ-populatie is getrokken en dat de gemiddelde θ-waarde in die populatie 0 is en de standaarddeviatie 1. Stel dat die persoon een hoge toetsscore haalt, met een ML-schatting van 3. Op grond van de toetsuitslag alleen zouden we besluiten tot een vaardigheids- schatting van 3, doch het veel lager gemiddelde van de populatie suggereert dat dit overdreven is. Immers, de kans dat er aselect een persoon met een θ-waarde van 3 of hoger wordt getrokken is zo klein, dat zich als het ware een correctie op de ML-schatter in de richting van het populatiegemiddelde opdringt. De EAP-schatter kan dus beschouwd worden als een soort compromis tussen de informatie die de toetsafname oplevert en de informatie over de populatie waarover we beschikken, net zoals de formule van Kelley die in hoofdstuk 3 is besproken.

Formeel is de EAP-schatter het gemiddelde van de a posteriori verdeling van θ, dit wil zeggen, de verdeling van θ indien de observaties gecombineerd worden met de a

priori verdeling van θ. Deze laatste verdeling is niets anders dan de verdeling van θ die aan het Raschmodel is toegevoegd om MML-schattingen te kunnen maken. De formules voor deze schatter volgen rechtstreeks uit het theorema van Bayes:

$$h(\theta|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\theta)g(\theta)}{P(\boldsymbol{x})}$$

$$= \frac{P(\boldsymbol{x}|\theta)g(\theta)}{\int_{-\infty}^{+\infty} P(\boldsymbol{x}|\theta)g(\theta)\,d\theta}, \qquad (4.119)$$

waarbij de tweede gelijkheid rechtstreeks uit (4.56) volgt. De functie $h(\theta|\boldsymbol{x})$ is de a posteriori dichtheid van θ, waarbij duidelijk te zien is dat deze functie afhankelijk is zowel van de data en de itemparameters, via $P(\boldsymbol{x}|\theta)$, als van de a priori verdeling en de daarmee geassocieerde parameters, via $g(\theta)$. Het gemiddelde van de a posteriori verdeling is dan gegeven door

$$\mathscr{E}(\theta|\boldsymbol{x}) = \int_{-\infty}^{+\infty} \theta\, h(\theta|\boldsymbol{x})\,d\theta. \qquad (4.120)$$

De schatter zegt dus eigenlijk dat de persoon beschouwd dient te worden als een aselecte trekking uit een populatie van θ-waarden met dichtheidsfunctie $h(\theta|\boldsymbol{x})$. De schatter zelf is het gemiddelde van die populatie. Daaruit volgt geenszins dat twee personen met hetzelfde antwoordpatroon ook dezelfde θ-waarde hebben. Immers de a posteriori verdeling heeft ook een variantie ongelijk 0. Deze variantie, of de vierkantswortel eruit, de a posteriori standaarddeviatie, kan dus gehanteerd worden als een maat van onzekerheid. Deze variantie is gegeven door

$$\mathrm{var}(\theta|\boldsymbol{x}) = \int_{-\infty}^{+\infty} \theta^2 h(\theta|\boldsymbol{x})\,d\theta - [\mathscr{E}(\theta|\boldsymbol{x})]^2. \qquad (4.121)$$

De term 'expected a posteriori' is afkomstig uit de bayesiaanse statistiek. 'Echte' Bayesianen voeren de a priori verdeling, zowel de vorm, bijvoorbeeld de normale verdeling, als de waarde van de parameters, op als een soort geformaliseerde overtuiging. Bij toepassingen met MML-schattingen wordt alleen de vorm van de verdeling ingevoerd als hypothese, terwijl de parameters uit de data worden geschat. Deze benadering wordt aangeduid als empirisch bayesiaans. Bij de EAP-schattingsprocedure worden dus de geschatte populatieparameters gebruikt om de a priori verdeling te specificeren.

Stel nu dat men bij de schatting van de item- en populatieparameters twee steekproeven, afkomstig uit twee verschillende populaties, heeft gebruikt, die dezelfde toets hebben gekregen. Eénzelfde antwoordpatroon zal leiden tot verschillende EAP-schatters voor beide populaties, en wel in die zin dat de EAP-schatter voor een persoon uit de populatie met het laagste gemiddelde kleiner zal zijn dan voor een persoon uit de andere populatie. Indien men schattingen van θ gebruikt om beslissingen te nemen die individuen raken, dient men zich terdege bewust te zijn van de ethische implicaties bij het gebruik van EAP-schatters. Immers, de beslissing wordt niet uitsluitend gebaseerd op de itemantwoorden, doch ook op achtergrondinformatie, waarvan het gebruik in bepaalde contexten discriminerend of onrechtvaardig kan zijn. De beslissing om ze dan maar niet te gebruiken is echter een beetje simplistisch. Als men ze niet gebruikt is men aangewezen op ML- of WML-schatters, waarvan de standaardfout in de regel groter is dan de a posteriori standaarddeviatie, en grotere standaardfouten betekenen automatisch meer verkeerde beslissingen. Een goed gefundeerde verhandeling over dit onderwerp ontbreekt echter nog in de psychometrische literatuur.

4.5.2 Een voorbeeld

Als illustratie bij het commentaar dat in de vorige paragraaf gegeven is, beschouwen we het volgende artificiële voorbeeld. Veronderstel dat er twee populaties, A en B zijn waarin de vaardigheid normaal verdeeld is met een standaarddeviatie gelijk aan 1. Het gemiddelde van populatie A is -0.6 en dat van populatie *B* is +0.6. Uit beide populaties wordt aselect een steekproef getrokken van 250 personen. De toets die aan beide steekproeven wordt voorgelegd bestaat uit 21 Raschitems met parameters -2.0, -1.8, -1.6,...,1.6, 1.8, 2.0. De parameters worden geschat met CML, en vervolgens wordt voor ieder antwoordpatroon de ML- en de WML-schatter berekend. Daarnaast zijn ook MML-schatters berekend, waarbij naast de itemparameters ook twee gemiddelden en twee varianties worden geschat. Na de parameterschattingen zijn de schattingen van θ berekend volgens de drie methodes: ML, WML en EAP. Voor WML en EAP geldt, net als voor ML-schatters, dat de schatting alleen afhankelijk is van de score. De resultaten staan in tabel 4.13.

De getallen tussen haakjes in tabel 4.13 zijn de a posteriori standaarddeviaties (voor MML) of de standaardfouten (voor WML en ML). Omdat populatie B gemiddeld vaardiger is krijgen leden uit populatie B ook systematisch een hogere θ-schatting dan leden van populatie A voor dezelfde score. De a posteriori standaarddeviaties zijn ook systematisch kleiner dan de standaardfouten van de WML- en de ML-schatters. De

toets bereikt haar maximale informatie voor θ in de buurt van 0, en we zien ook dat de standaardfouten van WML en ML hun kleinste waarde bereiken rond dit punt. De plaats waar de a posteriori standaarddeviatie haar kleinste waarde bereikt is niet alleen afhankelijk van de informatiefunctie maar ook van de waarde van het gemiddelde en de standaarddeviatie, dus van de a priori verdeling. Merk tenslotte nog op dat de ML-schattingen meer 'uitgerekt' zijn dan de WML-schattingen, terwijl de EAP-schattingen meer samengedrukt zijn.

Tabel 4.13
EAP-, WML- en ML-schattingen van θ

score	EAP (pop. A)		EAP (pop. B)		WML		ML	
0	-3.194	(.574)	-2.748	(.532)	-4.416	(1.844)	---	---
1	-2.883	(.544)	-2.477	(.510)	-3.210	(.966)	-3.570	(1.052)
2	-2.600	(.520)	-2.227	(.492)	-2.590	(.757)	-2.769	(.781)
3	-2.341	(.500)	-1.991	(.479)	-2.141	(.658)	-2.251	(.669)
4	-2.098	(.485)	-1.768	(.467)	-1.773	(.601)	-1.848	(.606)
5	-1.870	(.473)	-1.553	(.459)	-1.453	(.565)	-1.505	(.568)
6	-1.651	(.463)	-1.346	(.453)	-1.161	(.541)	-1.198	(.542)
7	-1.441	(.455)	-1.143	(.448)	-.888	(.524)	-.914	(.525)
8	-1.236	(.450)	-.944	(.444)	-.628	(.513)	-.645	(.514)
9	-1.035	(.446)	-.748	(.443)	-.367	(.507)	-.385	(.507)
10	-.838	(.443)	-.551	(.443)	-.127	(.504)	-.130	(.504)
11	-.641	(.443)	-.355	(.443)	.120	(.504)	.123	(.504)
12	-.444	(.444)	-.157	(.446)	.369	(.507)	.379	(.507)
13	-.247	(.445)	.044	(.451)	.622	(.514)	.640	(.515)
14	-.048	(.449)	.249	(.456)	.884	(.526)	.910	(.526)
15	.156	(.454)	.460	(.463)	1.159	(.542)	1.196	(.544)
16	.364	(.460)	.679	(.473)	1.453	(.567)	1.505	(.569)
17	.579	(.469)	.909	(.485)	1.776	(.603)	1.850	(.608)
18	.804	(.480)	1.152	(.501)	2.146	(.660)	2.255	(.670)

19	1.041	(.494)	1.412	(.520)	2.597	(.758)	2.775	(.782)
20	1.293	(.511)	1.694	(.544)	3.219	(.967)	3.578	(1.053)
21	1.565	(.533)	2.006	(.574)	4.425	(1.845)	---	---

4.5.3 Passingsindices voor individuele antwoordpatronen

In de vorige paragraaf is gesteld dat de schatter van θ alleen afhankelijk is van de score. Dit kan enigszins paradoxaal klinken. Stel dat van twee personen die precies de helft van de items juist hebben beantwoord, de eerste de *k*/2 gemakkelijkste items juist had, en de tweede de k/2 moeilijkste. Is het dan niet redelijk de vaardigheid van de tweede hoger te schatten ? De oplossing van deze paradox is gelegen in het dubbele standpunt dat men tegenover statistische gegevens kan innemen. Statistische gegevens veronderstellen bij analyse steeds een model. Een gedeelte van de informatie die de gegevens bevatten gebruikt men voor het schattingsprobleem. Men kan de schattingen gebruiken en interpreteren, en de juistheid van de interpretatie is alleen gegarandeerd als de modelveronderstellingen juist zijn. Of deze veronderstellingen juist zijn weet men nooit met absolute zekerheid, doch men kan de juistheid statistisch toetsen door het gebruik van andere informatie in de data. In het gegeven voorbeeld is het inderdaad terecht aan beide personen dezelfde schatting van θ toe te kennen indien het model juist is. Beide antwoordpatronen zijn echter in een bepaalde betekenis vrij extreem, zodat men er aan kan twijfelen of de antwoorden wel volgens het Raschmodel tot stand zijn gekomen. Naast de vaardigheid kunnen natuurlijk tal van andere factoren het gedrag bepaald hebben, en de invloed van deze factoren kan zo belangrijk zijn dat het Raschmodel niet meer geldig is.

Redenen voor niet-passing van het model voor individuele respondenten kunnen bijvoor- beeld zijn: vermoeidheid, oneerlijk gedrag, systematisch verkeerd invullen van schrapkaarten waarbij het antwoord voor item *i* wordt ingevuld op de plaats *i*+1, enzovoort. Een discussie van deze en nog andere redenen voor systematische afwijkingen van het model kan men vinden in Hulin, Drasgow en Parsons (1983), die ook een groot aantal indices bespreken waarmee niet-passende antwoordpatronen ontdekt kunnen worden. Een recente en heel interessante bijdrage op dit gebied kan men ook vinden in Klauer (1991). Bij wijze van voorbeeld bespreken we hier een zeer eenvoudige index, die we aanduiden als $z(\theta,\boldsymbol{x})$:

$$z(\theta,\boldsymbol{x}) = \sum_{i=1}^{k} [\, f_i(\theta) - x_i \,] \; . \qquad (4.122)$$

De interpretatie van (4.122) is eenvoudig: hij geeft de som van de afwijkingen tussen het itemantwoord x_i en de verwachte waarde $f_i(\theta)$, elke term is dus verschillend van 0. Grote absolute afwijkingen ontstaan indien een verkeerd antwoord wordt gegeven bij gemakkelijke items of een juist antwoord bij moeilijke items. Indien een volmaakt Guttmanpatroon optreedt waarbij de *s* gemakkelijkste items juist worden beantwoord en de *k*-*s* moeilijkste fout, zijn de absolute afwijkingen relatief klein. Bij een antwoordpatroon waarbij het omgekeerde het geval is, krijgen we wel grote absolute afwijkingen, doch hun teken is verschillend: juiste antwoorden op een moeilijk item resulteren in een negatieve afwijking en verkeerde antwoorden op een gemakkelijk item geven een positieve afwijking, met als gevolg dat die in de som tegen elkaar zullen wegvallen, en kunnen resulteren in een kleine waarde van de index, net zoals bij een Guttmanpatroon. Deze index is dus niet erg geschikt. Een index die wel onderscheid maakt tussen Guttmanpatronen en hun omgekeerde is de ζ_2-index van Tatsuoka (1984):

$$\zeta_2(\theta,\boldsymbol{x}) = \frac{\sum_{i=1}^{k}[f_i(\theta)-x_i][f_i(\theta)-\bar{f}(\theta)]}{\left(\sum_{i=1}^{k} f_i(\theta)[1-f_i(\theta)][f_i(\theta)-\bar{f}(\theta)]^2\right)^{\frac{1}{2}}} \qquad (4.123)$$

waarin $\bar{f}(\theta) = \Sigma_i f_i(\theta)/k$. De interpretatie van ζ_2 is het gemakkelijkst indien we veronderstellen dat de items geordend zijn volgens oplopende moeilijkheid, en de score s ongeveer *k*/2 bedraagt. Voor een Guttmanpatroon waarbij de *s* makkelijkste items juist zijn beantwoord, zullen de eerste *s* termen van (4.123) overwegend negatief zijn, want $f_i(\theta)$-$x_i < 0$ voor $i<s$, en als de verdeling van de moeilijkheidsparameters niet al te scheef is, zal gelden dat $f_i(\theta) > \bar{f}(\theta)$ voor het merendeel van de eerste *s* items. Een omgekeerd Guttman- patroon zal resulteren in een positieve index. Bovendien kan aangetoond worden dat de verwachte waarde van ζ_2 gelijk is aan 0 en de variantie gelijk aan 1. Indien *k* niet al te klein is kan ζ_2 geïnterpreteerd worden als een standaardnormaal verdeelde variabele: waarden van de index groter dan 2 in absolute waarde hebben een kleine kans om geobserveerd te worden indien de nulhypothese, het Raschmodel, waar is. De ζ_2-index is in het programma OPLM geïmplementeerd.

Bij het interpreteren van deze indices dient men de nodige voorzichtigheid aan de dag te leggen. Indien de index gebruikt wordt om beslissingen te nemen die

verstrekkende gevolgen kunnen hebben voor een bepaalde persoon, dient men te bedenken dat het voorkomen van een ongebruikelijk of vreemd antwoordpatroon geen waterdicht bewijs is van bijvoorbeeld oneerlijk gedrag. Immers, indien men toetst met een significantieniveau van 5%, dan kan men verwachten dat ongeveer 5% van de antwoordpatronen in de steekproef een significante index zal opleveren indien het model juist is. Is dit percentage in de steekproef substantieel groter, dan wijst dit er op dat er iets niet in de haak is met het model. Nader onderzoek kan dan gewenst zijn, doch de index op zichzelf is een zwakke basis om individuele beslissingen te rechtvaardigen. Hij kan hoogstens leiden tot een grotere voorzichtigheid. De Amerikaanse naam voor dit soort indices, caution indices, is dan ook heel terecht.

Op deze en vele andere indices die in de literatuur zijn gebruikt hebben Molenaar en Hoijtink (1990) vanuit statistisch standpunt nogal wat kritiek geleverd. Deze kritiek komt erop neer dat we, om deze indices uit te rekenen een schatting van θ in de formule moeten invullen, maar deze schatting is een functie van de score, en met een bepaalde score *s* zijn niet alle mogelijke antwoordpatronen verdraagbaar. Indien bijvoorbeeld $s=1$, dan zijn er maar *k* verschillende antwoorden mogelijk bij deze score, en dus is het redelijk om alleen deze *k* antwoordpatronen te beoordelen op hun 'vreemdheid' onder het Raschmodel. Molenaar en Hoijtink hebben een index ontwikkeld waarbij dit ook gebeurt. De statistische significantie-toetsing van deze index is echter behoorlijk ingewikkeld.

5

Een overzicht van itemresponsmodellen

In hoofdstuk 4 is uitvoerig ingegaan op het Raschmodel, waarbij de nadruk vooral kwam te liggen op de statistische aspecten van schatting en toetsing. Het is niet zo dat dit model een allesoverheersende plaats inneemt in de IRT-literatuur. Er zijn zeer veel IRT-modellen ontwikkeld, en een volledig overzicht geven van de bestaande modellen is in het bestek van een hoofdstuk niet mogelijk. De selectie die zal worden gepresenteerd, weerspiegelt naast een zekere voorkeur van de auteurs, enkele aspecten die voor de praktijk belangrijk zijn, eerder dan diepe theoretische overwegingen. Een van de aspecten is het algemeen beschikbaar zijn van computerprogrammatuur.

Thematisch valt dit hoofdstuk uiteen in twee onderdelen, die men zou kunnen aanduiden als specificatie en generalisatie van het Raschmodel. Nadere specificatie van het Raschmodel is het antwoord op de vraag 'wat kun je verder nog doen als het Raschmodel bij de data past' en generalisatie van het Raschmodel is het antwoord op de vraag 'wat te doen indien het Raschmodel niet bij de gegevens past ?'.

Indien het Raschmodel overtuigend bij de data past, hoeft dit niet noodzakelijkerwijze het einde van de psychometrische bemoeienissen met deze data te betekenen. Naast de praktische toepassingsmogelijkheden van een deugdelijke schaal, kan men zich ook de vraag stellen hoe het komt dat het ene item moeilijker is dan het andere. Dit wil zeggen dat men probeert een theorie te construeren die de verschillen in moeilijkheid tussen de items verklaart. Binnen de IRT is een benadering ontworpen die toelaat een grote klasse van deze theorieën formeel te beschrijven en statistisch te toetsen. Hoewel deze benadering in principe op elk IRT-model kan worden toegepast, heeft ze haar eerste en ook omvangrijkste uitwerking gekregen in het kader van het Raschmodel. Technisch gezien komt deze benadering neer op het opleggen van een aantal restricties aan de itemparameters. In hoofdstuk 4 is dit ook al een keer gedaan om de rationale van de LR- en de Wald-toetsen te beschrijven. Het resulterende model is minder algemeen dan het Raschmodel, en kan dus worden opgevat als een nadere specificatie ervan. Deze specificatie weerspiegelt een bepaalde theorie of hypothese

over de structuur van de moeilijkheid van de items. Een gedetailleerde uiteenzetting van dit model is het onderwerp van paragraaf 5.1

Indien het Raschmodel niet bij de data past, kan men twee standpunten innemen. Men kan items of personen verwijderen totdat de overblijvende items zich wel adequaat door het Raschmodel laten beschrijven. Daarbij kan echter de inhoudsvaliditeit van de toets of de generaliseerbaarheid naar de populatie van personen in het gedrang komen. Men kan ook proberen te achterhalen waarom het model niet past. In hoofdstuk 4 hebben we gezien dat het Raschmodel gelijke discriminatie van de items veronderstelt. Als we erachter komen, bijvoor- beeld met behulp van de M_i – toetsen, dat niet-passing toe te schrijven is aan ongelijke discri- minatie, kunnen we het Raschmodel vervangen door een algemener model dat ongelijke discriminatie toelaat, zoals het tweeparameter logistisch model dat in hoofdstuk 4 reeds kort werd besproken.

Generalisatie van het Raschmodel heeft ook nog een andere motivatie. Indien men over items beschikt met antwoordvariabelen die niet twee maar drie of meer verschillende waarden aannemen, dan komt de variant van het Raschmodel uit het vorige hoofdstuk niet in aanmerking, zodat men wel gedwongen is van een ander model gebruik te maken. Terzijde dient opgemerkt te worden dat het bespreken van IRT-modellen als generalisaties van het Raschmodel als didactisch hulpmiddel wordt gehanteerd en niet overeenkomt met de feitelijke historische ontwikkeling van de IRT: veel van de te presenteren modellen zijn eerder ontwikkeld dan het eigenlijke Raschmodel.

Paragraaf 5.2 is gewijd aan een algemene bespreking van de indelingsprincipes van IRT-modellen. In de paragrafen 5.3 en 5.4 komen unidimensionale modellen voor respectievelijk dichotome en polytome items aan de orde. Paragraaf 5.5 bespreekt enkele multidimensionale modellen.

5.1 Het lineair-logistische testmodel

Veronderstel dat de items van een toets bestaan uit wiskundige functies waarvan de afgeleide functie gevraagd wordt. Voor het nemen van afgeleiden bestaan specifieke regels, zoals:

$$\frac{d\,x^{n}}{d\,x} = n\,x^{n-1}$$

en

$$\frac{d \ln x}{d x} = \frac{1}{x}.$$

Nu is de hypothese dat de moeilijkheid van de items afhangt van de moeilijkheid van deze regels. Fischer (1973) stelde een zeer eenvoudig model voor om aan te geven hoe de itemmoeilijkheid tot stand komt. Indien in item *i* regel 1 tweemaal moet worden toegepast en regel 2 driemaal, dan is de moeilijkheid van dit item gegeven door

$$\beta_i = 2\eta_1 + 3\eta_2,$$

waarin η_1 en η_2 de moeilijkheden van de twee regels voorstellen. De coëfficiënten 2 en 3 in de gelijkheid hierboven zijn bekende constanten die volgen uit een analyse van de items. Indien we nu een toets maken met $k > 2$ items, die allemaal alleen een beroep doen op deze twee regels, dan moeten niet *k* parameters geschat worden, maar slechts 2, omdat de *k* itemparameters allemaal lineaire functies zijn van de twee η-parameters. Deze η-parameters worden aangeduid als basisparameters of elementaire parameters.

De veralgemening van bovenstaand voorbeeld is erg eenvoudig. Indien er $d < k$ basisparameters zijn, is het model gegeven door

$$\beta_i = \sum_{j=1}^{d} q_{ij} \eta_j, \qquad (i = 1,...,k). \tag{5.1}$$

De coëfficiënten q_{ij} in (5.1) zijn constanten die a priori in het model worden ingebracht en niet uit de data worden geschat. Deze coëfficiënten of gewichten zoals ze vaak worden genoemd, representeren dus de theorie van de onderzoeker. Formule (5.1) zegt dat de itemparameters lineaire combinaties zijn van *d* elementaire parameters en een dergelijke modellering wordt aangeduid als het lineair-logistische testmodel (LLTM). Dit model werd voorgesteld door Fischer (1974, 1983).

Het LLTM heeft dus twee componenten: de antwoorden op de items kunnen beschreven worden door het Raschmodel, en bovendien zijn de itemparameters specifieke lineaire combinaties van meer basale parameters η. Het schattingsprobleem zal dus bestaan uit het schatten van deze η-parameters en bij de toetsing moet de geldigheid van beide componenten van het model onderzocht worden. Schatting en toetsing worden hierna besproken.

5.1.1 Parameterschatting in het LLTM

We beginnen met een onderzoek van de aannemelijkheidsfunctie. In het Raschmodel is de aannemelijkheidsfunctie gegeven door formule (4.28), die we hier herhalen:

$$\ln L(\beta,\theta;\boldsymbol{X}) = \sum_{v=1}^{n} s_v \theta_v + \sum_{i=1}^{k} t_i(-\beta_i) - \sum_{v=1}^{n}\sum_{i=1}^{k} \ln\left[1 + \exp(\theta_v - \beta_i)\right], \tag{5.2}$$

waarin

$$s_v = \sum_{i=1}^{k} x_{vi}, \qquad t_i = \sum_{v=1}^{n} x_{vi}.$$

Substitueren we nu het rechterlid van (5.1) voor β_i in het rechterlid van (5.2), dan krijgen we:

$$\ln L(\eta,\theta;\boldsymbol{X}) = \sum_{v=1}^{n} s_v\theta_v + \sum_{j=1}^{d}(-\eta_j)\sum_{i=1}^{k} t_i q_{ij} - \sum_{v=1}^{n}\sum_{i=1}^{k} \ln\left[1 + \exp\left(\theta_v - \sum_{j=1}^{d} q_{ij}\eta_j\right)\right], \tag{5.3}$$

waarin we duidelijk de structuur van de exponentiële familie herkennen. De laatste term in het rechterlid is uitsluitend een functie van de parameters, de eerste term is onveranderd gebleven in vergelijking met (5.2), en de middelste term is een som van *d* produkten, waarvan een factor de parameter η_j is. De andere factor, $\sum_i t_i q_{ij}$, is alleen een functie van de data. Deze factor is dus een voldoende steekproefgrootheid voor de parameter η_j, en het model behoort tot de exponentiële familie. Dit is trouwens een voorbeeld van een algemeen resultaat: indien een model behoort tot de exponentiële familie, dan behoort het speciale geval van dit model dat ontstaat door lineaire restricties op de parameters aan te brengen eveneens tot de exponentiële familie.

In (5.3) is bovendien, net als in het gewone Raschmodel, de somscore de voldoende steekproefgrootheid voor de persoonsparameter. Door te conditioneren op de score kunnen we de conditionele aannemelijkheidsfunctie opstellen. Omdat het LLTM een speciaal geval is van het Raschmodel, moet de algemene formule voor de conditionele aannemelijk- heidsfunctie die in hoofdstuk 4 werd gegeven, hier ook geldig zijn. De logaritme van deze aannemelijkheidsfunctie is gegeven door formule (4.43) die we hier herhalen:

$$\ln L(\eta;\boldsymbol{x}\,|\,\boldsymbol{s}) = \sum_{i} t_i \ln \varepsilon_i - \sum_{v} \ln \gamma_{s_v}(\boldsymbol{\varepsilon}) \tag{5.4}$$

waarin

$$\varepsilon_i = \exp(-\beta_i) = \exp\left(-\sum_j^d q_{ij}\eta_j\right). \tag{5.5}$$

Substitueren we nu het rechterlid van (5.5) in (5.4), dan krijgen we:

$$\ln L(\eta; \boldsymbol{x} \mid \boldsymbol{s}) = \sum_j^d (-\eta_j) \sum_i t_i q_{ij} - \sum_v \ln \gamma_{s_v}(\boldsymbol{\varepsilon}). \tag{5.6}$$

De schattingsvergelijkingen kunnen we opstellen door van (5.6) de partiële afgeleiden naar de η-parameters gelijk te stellen aan 0, maar we kunnen ook gebruik maken van een eigen- schap van de exponentiële familie, die inhoud dat de schattingsvergelijkingen gegeven zijn door de voldoende steekproefgrootheden gelijk te stellen aan hun verwachte waarde. Dan krijgen we als schattingsvergelijkingen:

$$\begin{aligned} \sum_i q_{ij} t_i &= \mathscr{E}\left[\sum_i q_{ij} T_i \mid s_v\right] \\ &= \sum_i q_{ij} \sum_v \mathscr{E}(X_{vi} \mid s_v) \\ &= \sum_i q_{ij} \sum_v \pi_{i \mid s_v}, \qquad (j = 1, \ldots, d). \end{aligned} \tag{5.7}$$

Een vergelijking met de CML-schattingsvergelijkingen (4.45) laat meteen zien dat het gewone Raschmodel ook beschouwd kan worden als een LLTM, door de coëfficiënten q_{ij} te definiëren als

$$q_{ij} = \begin{cases} 1 \text{ indien } j = i \text{ en } i > 1, \\ 0 \text{ in andere gevallen}. \end{cases}$$

In het algemeen geldt dat in het LLTM de voldoende steekproefgrootheden gegeven zijn door d lineaire combinaties van de itemtotalen t_i en de schattingsvergelijkingen door het gelijk-stellen van die d lineaire combinaties aan hun verwachte waarde. In het gewone Raschmodel geldt natuurlijk dat $d = k - 1$.

Eén probleem dient nog even aan de orde gesteld te worden, namelijk het probleem van de normering van de basisparameters. Bij de behandeling van het Raschmodel in hoofdstuk 4 hebben we gezien dat een van de itemparameters vrij kan worden gekozen, of iets algemener uitgedrukt, dat bij elke itemparameter een willekeurige constante c kan worden opgeteld. Het LLTM is echter ook een Raschmodel en dus moet die vrijheid ook hier gelden. Dit is inderdaad zo, want de algemene vorm van het LLTM is iets algemener dan door (5.1) is aangegeven en luidt eigenlijk

$$\beta_i = \sum_{j=1}^{d} q_{ij}\eta_j + c, \qquad (i = 1,...,k), \tag{5.8}$$

waarin *c* ogenschijnlijk de status heeft van een parameter, maar niets anders is dan een willekeurige normalisatieconstante. In de afleidingen hierboven is gewerkt met (5.1) in plaats van met (5.8), doch dit is hetzelfde als de keuze $c = 0$; dat wil zeggen dat in alle afleidingen deze normering reeds was ingevoerd.

5.1.2 Het toetsen van het LLTM

Bij het toetsen van het LLTM moeten we er rekening mee houden dat het model twee componenten heeft en dat het meestal zinvol is die twee componenten afzonderlijk te toetsen. Het heeft namelijk niet veel zin de geldigheid van de restricties (5.1) te toetsen, als het Raschmodel zonder die restricties niet houdbaar is. De eerste stap in de toetsing zal er dus uit bestaan dat het Raschmodel zonder restricties getoetst wordt. Dit impliceert dat de parameters in het algemene model geschat worden, waarna een of meer toetsen die in hoofdstuk 4 besproken zijn worden toegepast. Indien deze toetsen geen aanleiding geven het algemene model te verwerpen, kunnen we het Raschmodel zonder restricties gebruiken om een LR-toets te construeren. De vector met parameters in het algemene model is gegeven door $\varphi_u = (\beta_1,...,\beta_k)$ en in het beperkte model door $\varphi_r = (\eta_1,...,\eta_d)$. De toetsings- grootheid

$$2[\ln L^*(\varphi_u; X) - \ln L^*(\varphi_r; X)],$$

waarin L^* het maximum van de conditionele aannemelijkheidsfunctie aanduidt, is asymp- totisch chi-kwadraat verdeeld met $k - 1 - d$ vrijheidsgraden. Details over de constructie van een LR-toets kan men vinden in paragraaf 4.3.3. Merk op dat (5.1) de nulhypothese is. Grote waarden van de toetsingsgrootheid geven dus aan dat de beperking van het model met de specifieke waarden q_{ij} die gebruikt zijn, niet ondersteund wordt door de observaties. De coëfficiënten q_{ij} maken dus deel uit van de nulhypothese en de reden tot verwerping van de nulhypothese zou dus kunnen zijn dat een of meer van die coëfficiënten verkeerd gespecificeerd zijn. We zullen hier een toets bespreken die gevoelig is voor zo'n verkeerde specificatie. In hoofdstuk 4 hebben we gezien dat om een LR-toets te construeren de parameters geschat moeten worden zowel in het algemene model als in het beperkte model. Bij de Wald-toetsen hoefden we maar één keer te schatten, namelijk onder het algemene model. De Wald-toetsen zijn gebaseerd op de rationele dat de restricties op de parameters in het beperkte

model ongeveer moeten gelden voor de parameterschattingen in het algemene model. Er bestaat echter ook een manier van toetsen waarbij de schatting van de parameters gebeurt onder het beperkte model. Deze toetsen staan in de literatuur bekend als Lagrange-Multiplier-toetsen (LM, Aitchison & Silvey, 1958) of efficiënte-score-toetsen (Rao, 1948). We geven hier een voorbeeld dat van toepassing is op het LLTM.

Stel dat we betwijfelen of we de coëfficiënt q_{12} wel goed gespecificeerd hebben. Als we niet echt een uitgesproken idee hebben welke waarde die coëfficiënt moet aannemen, zouden we zijn waarde uit de data kunnen schatten. Maar dat betekent dat we het getal q_{12} willen beschouwen als de waarde die een parameter, zeg κ_{12}, aanneemt. We veronderstellen dus een model dat als parameters niet alleen de d η-parameters bevat, maar ook nog de extra parameter κ_{12}. We beschouwen dit model als het algemene model en de bijbehorende parametervector is gegeven door $\varphi_u = (\eta_1,...,\eta_d,\kappa_{12})$. Het beperkte model waaronder we de schatting hebben uitgevoerd, is een restrictie op de parameterruimte, want we hebben de parameter κ_{12} gelijkgesteld aan de waarde q_{12}. Dus kunnen we schrijven: $\varphi_r = (\eta_1,...,\eta_d,q_{12})$. Het zal duidelijk zijn dat we voor een LR-toets of een Wald-toets met nulhypothese: $\kappa_{12} = q_{12}$ de parametervector φ_u moeten schatten en dat is geen eenvoudige aangelegenheid. We weten dat de CML-schatter $\hat{\kappa}_{12}$ moet voldoen aan

$$\left.\frac{\partial \ln L(\varphi_u; X \mid \boldsymbol{s})}{\partial \kappa_{12}}\right|_{\kappa_{12} = \hat{\kappa}_{12}} = 0. \tag{5.9}$$

Deze betekent dat de partiële afgeleide, geëvalueerd op het punt van de CML-schatting, gelijk moet zijn aan nul. Indien nu de nulhypothese waar is, mag de schatting $\hat{\kappa}_{12}$ niet ver afwijken van de hypothetische waarde q_{12} en moet dus gelden

$$\left.\frac{\partial \ln L(\varphi_u; X \mid \boldsymbol{s})}{\partial \kappa_{12}}\right|_{\kappa_{12} = q_{12}} \approx 0. \tag{5.10}$$

We hoeven dus de CML-schatting van κ_{12} niet te berekenen, we moeten alleen de partiële afgeleide van de log-aannemelijkheidsfunctie evalueren op het punt $\kappa_{12} = q_{12}$. Die partiële afgeleide zal echter ook een functie zijn van de η-parameters en de waarden die we voor die parameters moeten invullen is in (5.10) niet aangegeven. De waarden die men voor de η-parameters invult, zijn hun CML-schattingen $\hat{\eta}_j$, $j = 1,...,d$, onder het beperkte model. De schattingen van alle $d + 1$ parameters

onder het beperkte model kunnen we dus aangeven als $\hat{\varphi}_r = (\hat{\eta}_1,...,\hat{\eta}_d, q_{12})$. Als de nulhypothese waar is moet dus ook gelden dat

$$\left.\frac{\partial \ln L(\varphi_u; X \mid \boldsymbol{s})}{\partial \kappa_{12}}\right|_{\varphi_u = \hat{\varphi}_r} \approx 0. \tag{5.11}$$

Merk op dat per definitie geldt dat

$$\left.\frac{\partial \ln L(\varphi_u; X \mid \boldsymbol{s})}{\partial \eta_j}\right|_{\varphi_u = \hat{\varphi}_r} = 0, \quad (j = 1,...,d). \tag{5.12}$$

Als we alle partiële afgeleiden van de log-aannemelijkheidsfunctie, geëvalueerd in het punt $\hat{\varphi}_r$ verzamelen in een $d + 1$ vector $\boldsymbol{b}(\hat{\varphi}_r)$, dan zijn de eerste d elementen van die vector per definitie gelijk aan 0.

Stel dat we ook de matrix van tweede partiële afgeleiden naar alle $d + 1$ parameters van de vector φ_u bepalen en evalueren in de waarden van $\hat{\varphi}_r$. Keren we het algebraïsche teken van deze matrix om, dan krijgen we de geobserveerde informatiematrix, geëvalueerd in $\hat{\varphi}_r$. Deze matrix kunnen we dus aanduiden als $I(\hat{\varphi}_r)$. De toetsingsgrootheid $LM(q_{12})$ is dan gegeven door

$$LM(q_{12}) = \boldsymbol{b}'(\hat{\varphi}_r)\,[I(\hat{\varphi}_r)]^{-1}\,\boldsymbol{b}(\hat{\varphi}_r) \tag{5.13}$$

en is onder de nulhypothese asymptotisch chi-kwadraat verdeeld met 1 vrijheidsgraad. Het uitrekenen van (5.13) is relatief eenvoudig omdat de elementen van de $\boldsymbol{b}$-vector die overeen- komen met de η-parameters exact gelijk zijn aan nul. Op deze vereenvoudiging gaan we hier echter niet in.

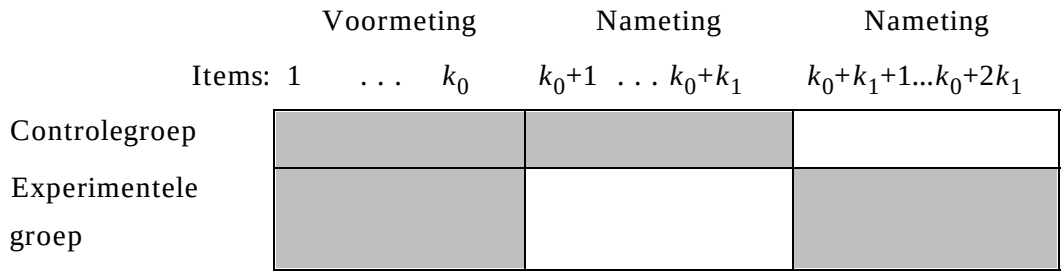
De LM-toetsen kunnen ook veralgemeend worden voor meer parameters tegelijkertijd, door in de $\boldsymbol{b}$-vector en in de informatiematrix de partiële afgeleiden op te nemen naar meerdere coëfficiënten q_{ij} die men in de toetsing van de hypothese wil betrekken.

Hoewel het gebruik van de LM-toetsen zeer aantrekkelijk is voor verfijning van het LLTM, dienen toch een kanttekening gemaakt te worden. Deze kanttekening heeft te maken met een nuancering die we impliciet in de nulhypothese hebben ingebracht. De rationale van de LM-toets hebben we beschreven alsof het hele probleem eruit bestond te weten of de restrictie $\kappa_{12} = q_{12}$ waar was en daarbij hebben we gedaan alsof het

algemene model waar was. Maar dat algemene model is heel complex, het veronderstelt het Raschmodel en de lineaire restricties waarvan de coëfficiënten, met uitzondering van q_{12}, allemaal vaste waarden hebben. Deze gespecificeerde waarden maken dus ook deel uit van het algemene model en van het beperkte model. Indien een of meer van deze gespecificeerde waarden erg afwijken van de werkelijke waarden, is het onbeperkte model niet meer juist en is de toetsingsgrootheid $LM(q_{12})$ ook niet meer chi-kwadraat verdeeld. De LM-toetsen zijn dus vooral nuttig indien de restricties die aangebracht zijn niet al te ver bezijden de werkelijkheid zijn.

5.1.3 Een toepassing van het LLTM

Een interessante toepassing van het introduceren van lineaire restricties op de itemparameters is het analyseren van gegevens die verzameld zijn in een experiment of een quasi-experiment. Stel dat in een experiment twee groepen worden onderscheiden: een experimentele groep die een behandeling krijgt en een controlegroep die geen behandeling krijgt. In beide groepen vindt een voor- en een nameting plaats. De voormeting wordt uitgevoerd met een toets van k_0 dichotome items en de nameting met een toets van k_1 items. De items in de voor- en de nameting behoeven niet dezelfde te zijn. Het is het meest voor de hand liggend om het effect van de behandeling te modelleren als een verandering in de persoonsparameters. Daar we echter gebruik willen maken van de in hoofdstuk 4 beschreven methodologische voordelen van de CML-schattingsmethode, zal in deze toepassing een verandering in de persoons- parameters vertaald worden in een verandering in de itemparameters. Met andere woorden, toename van de persoonsparameters wordt vertaald in een afname van de itemparameters. Als we aannemen dat de experimentele behandeling een positief effect heeft op de latente vaardigheid, moeten de itemparameters in de experimentele groep een kleinere waarde hebben dan in de controlegroep. Een elegante manier om dit te onderzoeken bestaat uit de volgende procedure, die logisch gezien twee stappen bevat. De eerste stap bestaat er uit, te doen alsof de oorspronkelijke k_1 items die voor de nameting worden gebruikt verdubbeld zijn, zodat er $2k_1$ items gebruikt zijn voor de nameting. Dit resulteert in een onvolledig design dat schematisch is weergegeven in figuur 5.1. De rijen in deze figuur zijn geassocieerd met groepen personen. De kolommen in de figuur zijn geassocieerd met items. Bij de voormeting hebben beide groepen dezelfde items gekregen. In de nameting is dat ook gebeurd, alleen hier wordt voorlopig even verondersteld dat de items door de experimentele manipulatie niet meer voor beide groepen hetzelfde zijn.



Figuur 5.1
Datamatrix met conceptuele items

Met andere woorden, elk 'fysiek' item in de nameting wordt gesplitst in twee 'conceptuele' items. We gaan er van uit dat de conceptuele items zo geordend zijn dat de conceptuele items $k_0 + i$ en $k_0 + k_1 + i$ naar hetzelfde fysieke item verwijzen. Deze associatie en de effecten van de behandeling worden nu gemodelleerd door het invoeren van de volgende lineaire restricties op de parameters van de conceptuele items:

$$\begin{cases} \beta_i = \eta_i, \quad (i = 1,...,k_0), \\ \beta_{k_0+i} = \eta_{k_0+i}, \quad (i = 1,...,k_1), \\ \beta_{k_0+k_1+i} = \eta_{k_0+i} + \tau, \quad (i = 1,...,k_1). \end{cases} \tag{5.14}$$

De associatie tussen de conceptuele items in de nameting komt tot uiting in de tweede en derde regel van (5.14) waar de twee conceptuele items $k_0 + i$ en $k_0 + k_1 + i$ betrokken worden op dezelfde basisparameter η_{k_0+i}. De parameter τ is de basisparameter die het effect van de experimentele behandeling weerspiegelt. Als τ positief is, worden de items moeilijker en heeft de experimentele behandeling dus een negatief effect. Bij een positief effect hoort een negatieve τ. Het algebraïsche teken van τ wordt in (5.14) niet gespecificeerd. Om duidelijk te maken dat (5.14) een speciaal geval is van (5.1), kunnen we (5.1) herschrijven als een matrixvergelijking door alle q_{ij}'s op te vatten als de elementen van een $k \times d$ gewichtenmatrix Q:

$$\beta = Q\eta. \tag{5.15}$$

Passen we (5.15) nu toe op het bovenstaande voorbeeld met $k_0 = k_1 = 2$, dan krijgen we

$$\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \beta_4\\ \beta_5\\ \beta_6\end{bmatrix} = \begin{bmatrix}1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&1&0&1\\ 0&0&0&1&1\end{bmatrix}\begin{bmatrix}\eta_1\\ \eta_2\\ \eta_3\\ \eta_4\\ \tau\end{bmatrix}. \qquad (5.16)$$

Omdat we één itemparameter vrij kunnen kiezen, kunnen we bijvoorbeeld β_1 gelijkstellen aan 0, maar omdat $\beta_1 = \eta_1$, geldt dan dat $\eta_1 = 0$. Er zijn dus niet vijf vrije basisparameters maar slechts vier. De lineaire restricties op de vrije itemparameters krijgen we dus door in (5.16) de elementen β_1 en η_1 en de eerste rij van de matrix te schrappen.

Dit model kan getoetst worden door het opstellen van een LR-toets waarbij het algemene model de geldigheid van het Raschmodel voor alle $k_0 + 2k_1$ conceptuele items veronderstelt en waarbij dus $k_0 + 2k_1 - 1$ vrije β-parameters geschat worden. In het beperkte model, waar geschat wordt onder de restricties (5.14) zijn er $k_0 + k_1$ vrije basisparameters. De LR-toets levert dus een toetsingsgrootheid op die asymptotisch chi-kwadraat verdeeld is met $k_1 - 1$ vrijheidsgraden.

Als het model geldig is, betekent dit natuurlijk niet automatisch dat het experiment effect heeft gehad. Om dit aan te tonen moeten we de nulhypothese $\tau = 0$ toetsen. Dit kan door een Wald-toets te gebruiken, waarbij de toetsingsgrootheid gegeven is door $\hat{\tau}/SE(\hat{\tau})$ en die onder de nulhypothese asymptotisch standaardnormaal verdeeld is. Het toetsen van deze nulhypothese heeft alleen zin indien het gehanteerde LLTM houdbaar blijkt. Indien dit niet het geval is, heeft een toetsing van de effectparameter geen zin.

Bij de interpretatie van de resultaten moet uiteraard rekening worden gehouden met alle aspecten van de interne validiteit in het wetenschappelijk onderzoek; het gebruik van een IRT-model maakt methodologische overwegingen niet overbodig. Voor dit soort overwegingen zij men verwezen naar Campbell en Stanley (1966), we gaan er hier nu niet verder op in.

Indien de LR-toets een significant resultaat oplevert, zou men kunnen denken dat het gehanteerde LLTM te streng is en dat het wellicht versoepeld kan worden door niet één enkele τ-parameter in het model toe te laten, maar een, mogelijk verschillende, τ_i-parameter voor elk item. Deze aanpak leidt echter tot logische problemen die verband houden met de proefopzet. Men gaat er namelijk van uit dat de hele verzameling gebruikte items aan het Raschmodel voldoen. Het Raschmodel schrijft echter voor dat de verandering in vaardigheid equivalent is met een en dezelfde verandering in de waarde van alle opgaven. Als men bij aparte items aparte effecten definieert, is het

bijvoorbeeld heel goed mogelijk dat de rangorde van de items op het latente continuüm voor de controle en de experimentele groep niet meer dezelfde is. Dit leidt dus tot een tegenspraak met de stelling dat alle opgaven aan het Raschmodel voldoen.

Tot slot van deze paragraaf nog een opmerking over de schatbaarheid van de parameters. Indien de voortoets weggelaten zou worden uit het design dat in figuur 5.1 is afgebeeld, zijn de parameters van het model, zowel met als zonder de restrictie (5.14) niet meer schatbaar. Men zou kunnen opperen dat dit rechtstreeks voortvloeit uit het in paragraaf 4.4 besproken feit dat CML-schattingen niet kunnen worden berekend uit een niet-verbonden design. Het probleem is in het algemeen echter iets gecompliceerder dan in paragraaf 4.4 werd besproken, omdat we het design moeten beschouwen in samenhang met de lineaire restricties. Zo kunnen er designs bestaan die zonder lineaire restricties niet schatbaar zijn, maar het wel worden met bepaalde lineaire restricties. De precieze condities wanneer dit het geval is, zijn gegeven in Fischer (1983). De conclusie is dus dat de voortoets niet kan worden weggelaten.

5.2 Indelingsprincipes van IRT-modellen

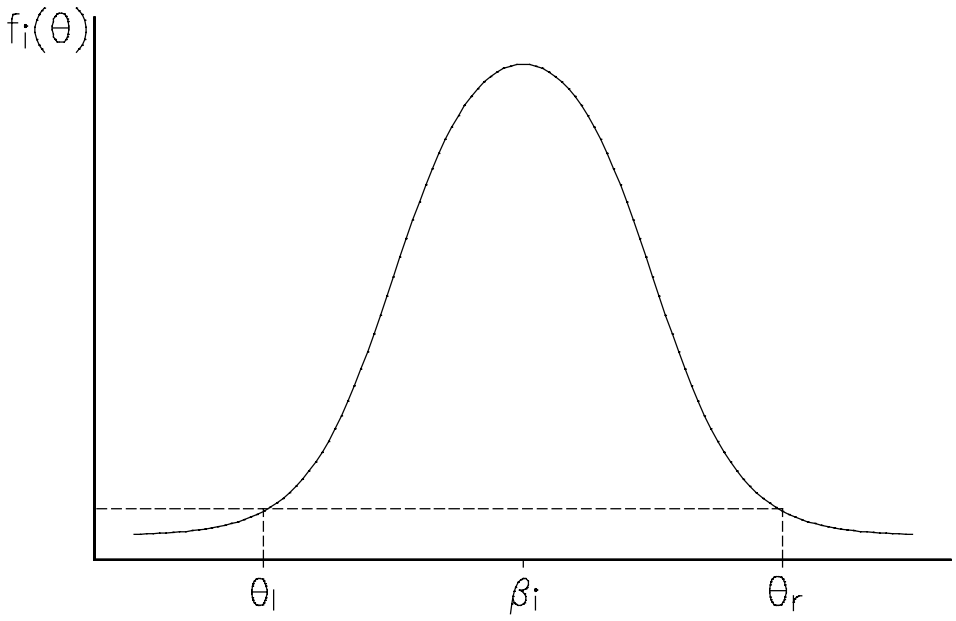
Om een inzicht te krijgen in de grote collectie IRT-modellen, zullen we drie indelingsprincipes hanteren: de algemene vorm van de itemresponsfunctie, namelijk monotoon tegenover niet-monotoon, het aantal categorieën dat de antwoordvariabele kan aannemen, namelijk twee tegenover meer dan twee, ofwel dichotoom tegenover polytoom en als derde de dimensionaliteit van de latente variabele. We becommentariëren kort deze drie principes.

In hoofdstuk 4 hebben we betoogd dat het een wenselijke eigenschap is van een IRT-model dat de itemresponsfunctie monotoon stijgend is in θ: hoe groter de vaardigheid, des te groter de kans op een juist antwoord. We kunnen echter ook modellen beschouwen waarbij de latente variabele die we wensen te meten niet adequaat aangeduid wordt met de categorie ’vaardigheid’. Beschouw het volgende item uit een fictieve vragenlijst naar politieke interesse:

"Vindt U dat Joop den Uyl een goede premier van Nederland was ?",

waarbij een positief antwoord gecodeerd wordt met 1 en een negatief antwoord met 0. Indien we veronderstellen dat het antwoord op dit item bepaald wordt door de positie van de persoon op een continuüm dat de politieke ’links-rechts’-dimensie weerspiegelt, is het niet aannemelijk dat hoe rechtser de persoon is, hoe groter de kans zal zijn dat het item bevestigend beantwoord wordt. Een veel plausibeler model is grafisch

weergegeven in figuur 5.2, waarbij β_i de positie van het item op het latente continuüm aangeeft.



Figuur 5.2
Een ééntoppige itemresponsfunctie

Deze positie weerspiegelt precies die politieke overtuiging die nodig is om de bovenstaande uitspraak met maximale kans te ondersteunen. De persoon met latente positie θ_l bevindt zich links van β_i en heeft een kleine kans om het item bevestigend te beantwoorden: Den Uyl wordt te rechts bevonden. Een persoon met positie θ_r ja zegt met een even kleine kans, maar de reden is dat Den Uyl te links bevonden wordt. Modellen met een eentoppige in plaats van een monotone itemresponsfunctie horen thuis in een domein dat doorgaans wordt aangeduid met ontvouwingstheorie. Een uiteenzetting van deze theorie kan men vinden in het werk van haar grondlegger C.H. Coombs (1964). Een goed overzicht van verschillende IRT-modellen met eentoppige itemresponsfuncties vindt men in het aan ontvouwing gewijde themanummer van het tijdschrift Kwantitatieve Methoden (Hoijtink, 1993). Deze modellen komen in dit hoofdstuk verder niet meer ter sprake.

Bij de modellen met monotone itemresponsfuncties kan men een belangrijke onderverdeling maken volgens het soort wiskundige functie dat men hanteert. In het Raschmodel is dat bijvoorbeeld de logistische functie. De grafiek van deze functie lijkt echter erg op de grafiek van de (cumulatieve) normale verdelingsfunctie. Deze laatste functie is dan ook in veel modellen gebruikt. Deze modellen staan bekend onder de algemene naam 'normaal-ogiefmodellen'. Voor een algemene inleiding en een

rechtvaardiging van het gebruik van de normale-verdelingsfunctie, verwijzen we naar hoofdstuk 16 van Lord en Novick (1968). Hoewel de logistische functie bij wiskundige afleidingen tot veel eleganter resultaten leidt dan de normale verdelingsfunctie, wordt die laatste nog steeds gebruikt, zij het niet zozeer in de literatuur die men gewoonlijk onder de benaming IRT aanduidt, maar meer in het onderzoeksdomein van de structurele modellen; zie bijvoorbeeld Muthén (1984, 1987).

Een zeer opmerkelijke klasse van modellen ontstaat indien men probeert de specifieke vorm van de itemresponsfunctie zo weinig mogelijk vast te leggen. Bij de modellen met een logistische functie of bij het normaal-ogiefmodel wordt de familie van de itemresponsfuncties zodanig gespecificeerd dat alleen nog één of meer parameters moeten worden geschat om de functies volledig te kennen. Mokken (1971) heeft een klasse van modellen gespecificeerd waarbij alleen zeer algemene kenmerken van de itemresponsfuncties worden vastgelegd, zoals monotoniciteit en dat de grafieken van de functies elkaar niet snijden. Parameters komen daarbij niet voor en deze modellen worden dan ook vaak aangeduid als niet-parametrische IRT-modellen. Mokken heeft aangetoond dat met dit soort zwakke eisen toch zinvolle uitspraken over de θ-waarde van personen kunnen worden gedaan en dat eveneens statistisch kan getoetst worden of aan deze eisen wel voldaan is. Recent onderzoek naar niet-parametrische IRT-modellen kan men vinden in Sijtsma en Molenaar (1987). Van de modellen die verder in dit hoofdstuk worden besproken, behoren de itemresponsfuncties allemaal tot de familie van de logistische functies.

Het tweede indelingsprincipe heeft betrekking op het aantal antwoordcategorieën. Indien dit aantal groter dan twee is, spreekt men niet van dichotome items maar van polytome items. Het is belangrijk op te merken dat het kenmerk dichotoom versus polytoom te maken heeft met het aantal waarden dat de antwoordvariabele X_i kan aannemen en dat dit aantal niet hetzelfde hoeft te zijn als het aantal categorieën waarin de oorspronkelijke observaties zijn ingedeeld. Een goed voorbeeld van dit onderscheid is het geval van meerkeuze-items. Stel dat een item met vier antwoordalternatieven, A, B, C en D, heeft, waarbij B het juiste antwoord is. Als we ervan uitgaan dat iedere persoon precies één van die alternatieven kiest, zijn er dus vier mogelijke antwoorden op dit item. Maar daaruit volgt niet dat we de antwoorden op dit soort items moeten analyseren met een model voor polytome items. We kunnen immers de oorspronkelijke observaties reduceren tot dichotome data door een punt toe te kennen indien het juiste alternatief gekozen is en geen punten in de andere drie gevallen. Indien we de versie van het Raschmodel uit hoofdstuk 4 gebruiken, analyseren we dichotome data en de statistische toetsen hebben alleen op deze data betrekking. Indien het model goed bij de data past, volgt daar niet uit dat deze analyse van de dichotome de enig juiste is.

Het is bijvoorbeeld mogelijk dat het kiezen van alternatief A een indicatie is van een grotere vaardigheid dan het kiezen van C of D. Indien we dit vermoeden hebben, kunnen we een analyse uitvoeren die gevoelig is voor dit onderscheid door een IRT-model voor polytome items te gebruiken. De wijze waarop de antwoorden van de personen gescoord worden, weerspiegelt een vermoeden of een hypothese en het gebruik van een formeel IRT-model is te beschouwen als een toetsing van deze hypothese. De geldigheid van een IRT-model betreft dus niet alleen de antwoorden (het gedrag) van de personen die de toets gemaakt hebben, maar ook de scoringsregel. De scoringsregel weerspiegelt een hypothese over de interpretatie die aan de responsen in de verschillende categorieën gegeven moet worden. In het boven- staande voorbeeld zouden we bijvoorbeeld 2 punten kunnen toekennen voor het antwoord B, 1 punt voor het antwoord A en 0 punten voor de antwoorden C en D, om vervolgens een model toe te passen waarbij een hogere itemscore als een indicator van een grotere vaardigheid wordt beschouwd. In dat geval zegt men dat we te doen hebben met een polytoom item met geordende antwoordcategorieën. Anderzijds zouden we ook de antwoorden A tot en met D ook kunnen omcoderen willekeurige getallen waarvan we de waarden niet wensen te interpreteren als geordende maar als nominale categorieën. Voor beide gevallen, geordende en nominale categorieën, zijn unidimensionele IRT-modellen ontwikkeld. Ze zullen behandeld worden in paragraaf 5.4.

Vooraleer we het derde indelingsprincipe bespreken, moeten we even ingaan op een complicatie die ontstaat wanneer de twee voorgaande indelingsprincipes met elkaar gecombineerd worden. Bij de bespreking van het eerste indelingsprincipe, monotone versus niet-monotone itemresponsfuncties, hebben we een terminologie gehanteerd die geschikt is voor dichotome items, maar die tekortschiet voor polytome items. Zoals we verder gedetail- leerd zullen bespreken, maar nu reeds intuïtief kunnen inzien, kunnen we voor een model met polytome items niet volstaan met een enkele itemresponsfunctie per item. We zullen een responsfunctie nodig hebben voor elke categorie van de antwoordvariabele. Daarom zullen we in het geval van polytome items ook niet meer spreken over de itemresponsfunctie maar over categorieresponsfuncties. Bovendien zal blijken dat niet alle categorieresponsfuncties van een item *i* monotoon stijgend of dalend in θ kunnen zijn. Om toch een indeling monotoon versus niet-monotoon te kunnen handhaven, zullen we de eigenschap monotoniciteit verder niet meer associëren met een categorieresponsfunctie, maar met een speciale functie die de itemregressiefunctie genoemd wordt. De regressie van de antwoordvariabele X_i op de latente variabele θ is de verwachte waarde van X_i, beschouwd als een functie van θ. In het Raschmodel is die itemregressiefunctie gegeven door:

$$\mathcal{E}(X_i \mid \theta) = 1 \times f_i(\theta) + 0 \times [1 - f_i(\theta)] = f_i(\theta). \quad (5.17)$$

Bij dichotome antwoordvariabelen valt de itemregressiefunctie samen met de itemresponsfunctie. Bij polytome items kan de itemregressiefunctie beschouwd worden als een samenvatting van alle categorieresponsfuncties. We zullen van een monotoon item spreken indien de itemregressiefunctie van de antwoordvariabele monotoon is in θ, of, iets informeler uitgedrukt, het item is monotoon als een grotere vaardigheid een grotere verwachte itemscore impliceert.

Het derde indelingsprincipe is de dimensionaliteit van de latente variabele θ. In hoofdstuk 4 is er op gewezen dat de aanname van unidimensionaliteit centraal staat in het Raschmodel. Deze aanname betekent dat alle items in een toets dezelfde vaardigheid meten. Nu is het mogelijk dat de items in een toets een beroep doen op twee verschillende vaardigheden, maar niet allemaal in dezelfde mate. Anders gezegd, alle items doen een beroep op beide vaardigheden, maar de mate waarin kan voor beide vaardigheden van item tot item verschillen. Het is bijvoorbeeld aannemelijk dat redactiesommen in een rekentoets zowel een verbale als een numerieke vaardigheid aanspreken. Als ze dat in ongelijke mate doen, zal een unidimensionaal model waarschijnlijk niet toereikend zijn om het antwoordgedrag op een dergelijke toets adequaat te beschrijven. Men kan dan proberen de oorspronkelijke toets op te splitsen in twee unidimensionale deeltoetsen, bijvoorbeeld met behulp van Martin-Löfs toets voor unidimensionaliteit (zie paragraaf 4.3.1), of men kan een model gebruiken waarin de vaardigheid meerdimensionaal is.

Op het eerste gezicht lijkt een unidimensionaal model, zoals het Raschmodel, het allereenvoudigste geval in de klasse van multidimensionale modellen. Maar het concept van een enkele dimensie betekent dat verschillende θ-waarden zinvol kunnen worden geordend. Men kan deze ordening echter ook beschouwen als een te strenge eis en proberen een model te maken waarin de verschillende θ-waarden niet geordend zijn, maar worden behandeld als nominale categorieën of klassen. Het meten is dan het toewijzen van een persoon aan een bepaalde klasse, terwijl de klassen onderling niet met elkaar in verband worden gebracht. Het model op zichzelf is uiterst eenvoudig. Stel dat er A klassen zijn. De conditionele kans op een antwoordpatroon $\boldsymbol{x}$, gegeven dat het afkomstig is van een persoon uit klasse a is gegeven door

$$\pi_{\boldsymbol{x} \mid a} = \pi_{x_1 \mid a} \pi_{x_2 \mid a} \ldots \pi_{x_k \mid a}, \quad (5.17)$$

waarin men direct een toepassing herkent van het principe van de lokale stochastische onafhankelijkheid. De data bestaan echter uit de antwoordpatronen $\boldsymbol{x}$ en het klasselidmaatschap van een persoon is niet geobserveerd. Als de kans dat een persoon

behoort tot klasse a voorgesteld wordt door π_a, $(a = 1,..., A)$, is de marginale kans op een antwoordpatroon $\boldsymbol{x}$ gegeven door

$$P(\boldsymbol{x}) = \sum_a \pi_{\boldsymbol{x}|a}\pi_a = \sum_a \pi_{x_1|a}...\pi_{x_k|a}\pi_a. \tag{5.18}$$

In het geval van dichotome items moet dus voor elk item de conditionele kans op een antwoord geschat worden gegeven de klasse a, $\pi_{x_i|a}$, en daarenboven moeten A-1 onafhankelijke kansen π_a geschat worden. Hoewel het model op zichzelf een heel eenvoudige structuur heeft, is de schatting van de parameters geen triviaal probleem. Dit model is een van de eerste IRT-modellen en werd voorgesteld door Lazarsfeld (1950). Het model kreeg van Lazarsfeld de naam latente-klassenmodel, omdat het klasselidmaatschap niet geob-serveerd, dus latent is. Lazarsfeld gebruikte trouwens niet het begrip IRT maar de algemene benaming 'Latente-structuuranalyse' om modellen met latente variabelen aan te duiden.

<table>
<tr><td colspan="3">Monotone items</td><td>Niet-monotone items</td></tr>
<tr><td rowspan="2">Unidimensionaal</td><td>Dichotoom</td><td>Hoofdst. 4 en 5.3</td><td rowspan="3">Ontvouwings modellen</td></tr>
<tr><td>Polytoom</td><td>5.4</td></tr>
<tr><td>Multidimensionaal</td><td>Dichotoom en polytoom</td><td>5.5</td></tr>
<tr><td>A-dimensionaal</td><td colspan="3">Latente-klassenmodellen</td></tr>
</table>

Figuur 5.3
Een indeling van itemresponsmodellen

In figuur 5.3 is een schematische weergave gegeven van de indeling van IRT-modellen die hiervoor werd besproken. De gearceerde oppervlakken bevatten een verwijzing naar de paragrafen in dit hoofdstuk waar een of meer modellen uit de cel van de figuur zullen worden besproken.

Het valt in figuur 5.3 op dat het onderscheid in monotone en niet-monotone items niet gehandhaafd is bij a-dimensionale gevallen. Dit kan ook niet anders, want het begrip monotoniciteit heeft geen enkele betekenis als de waarden van de latente variabele niet geordend kunnen worden. De indeling van IRT-modellen als in figuur

5.3 is voorgesteld is zeker niet de enig mogelijke. Ze is bedoeld als een handvat om enige orde te scheppen in de grote hoeveelheid modellen die in de literatuur zijn beschreven. Andere indelingen, die ook andere verbanden duidelijker belichten, zijn gegeven door Masters en Wright (1984), Thissen en Steinberg (1986) en Heinen (1993).

5.3 Unidimensionale modellen voor dichotome items

In hoofdstuk 4 is op verschillende plaatsen gewezen op een paar kwetsbare punten van het Raschmodel, namelijk de strenge eis dat alle items gelijkelijk moeten discrimineren en het feit dat het Raschmodel ongeschikt is om de relatief grote kansen op een juist antwoord te verklaren wanneer er geraden wordt bij meerkeuze-items. In de literatuur zijn modellen ontwikkeld die op het eerste gezicht een afdoend antwoord bieden op deze problemen. De meest prominente modellen zijn het twee- en het drieparameter logistisch model. Deze twee modellen worden besproken in paragraaf 5.3.1. We zullen echter zien dat het gebruik van deze modellen niet helemaal zonder problemen is omdat hierbij bepaalde aantrekkelijke eigenschappen van het Raschmodel verloren. Met name de mogelijkheid om itemparameters met de CML-methode te schatten is niet meer aanwezig. In paragraaf 5.3.2 wordt een model besproken dat de flexibiliteit van het tweeparameter logistisch model koppelt aan de theoretische voordelen van het Raschmodel. Het is het zogenaamde éénparameter logistisch model (Engels: One Parameter Logistic Model, OPLM).

In paragraaf 5.3.3 wordt ingegaan op modellen die geschikt zijn wanneer het axioma van de lokale stochastische onafhankelijkheid geschonden is. Te zelfder tijd zullen we zien dat het gebruik van deze modellen, in samenhang met de constructie van LR-toetsen, toelaat de geldigheid van dit axioma statistisch te toetsen.

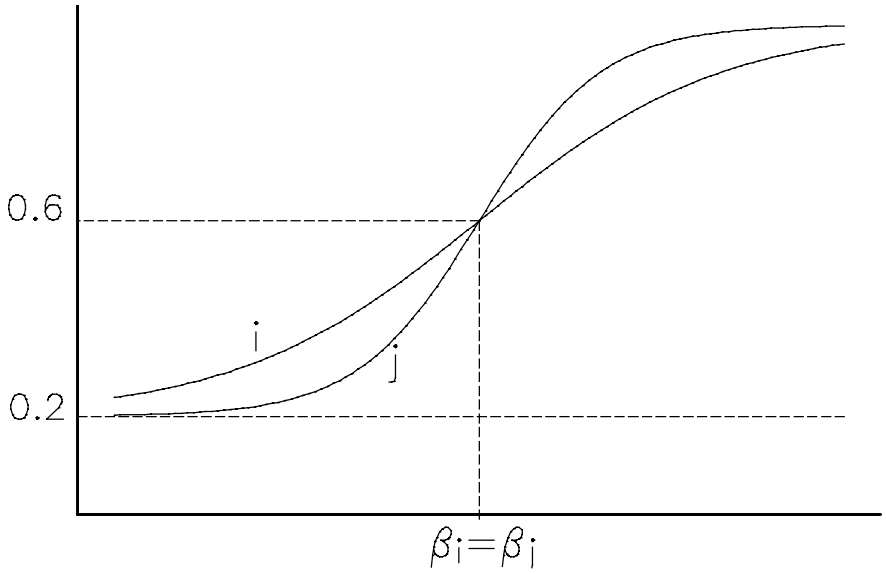
5.3.1 Het twee- en drieparameter logistisch model

Het tweeparameter logistisch model (Birnbaum, 1968) werd reeds kort besproken in hoofdstuk 4. Hier beginnen we met het drieparameter logistisch model dat eveneens door Birnbaum (1968) is beschreven. Een uitvoerige discussie over dit model kan men vinden in Lord (1980). Daarna zullen we zien dat het tweeparametermodel beschouwd kan worden als een speciaal geval van het drieparametermodel. In de literatuur worden

deze modellen vaak afgekort met 2PL en 3PL, deze afkortingen zullen we ook hier gebruiken. De itemresponsfunctie in het 3PL is gegeven door:

$$f_i(\theta) = c_i + (1 - c_i)\frac{\exp[a_i(\theta - \beta_i)]}{1 + \exp[a_i(\theta - \beta_i)]}, \quad (a_i > 0\ ;\ 0 \leq c_i < 1). \tag{5.19}$$

In figuur 5.4 staan de grafieken van twee itemresponsfuncties $f_i(\theta)$ en $f_j(\theta)$ met $\beta_i = \beta_j$, $c_i = c_j = 0.2$, $a_i = 1$ en $a_j = 2$.



Figuur 5.4
Itemresponsfuncties in het 3PL

De curve van item *j* verloopt steiler dan die van item *i*, hetgeen het effect van een grotere discriminatieparameter weerspiegeld. Het is gemakkelijk na te gaan dat in het 3PL de volgende limieten gelden

$$\lim_{\theta \to \infty} f_i(\theta) = 1$$

$$\lim_{\theta \to -\infty} f_i(\theta) = c_i$$

De parameter c_i geeft dus de kans op een juist antwoord aan indien de vaardigheid zeer klein is. Iets losser geformuleerd zou men kunnen zeggen dat c_i de kans is op een juist antwoord als men het antwoord niet 'kent'. Dit model lijkt dus geknipt te zijn voor toepassing bij meerkeuze-vragen. De parameter c_i wordt dan ook vaak aangeduid als de raadparameter. De interpretatie van deze parameter is echter ingewikkelder dan het op het eerste gezicht lijkt. In de eerste plaats is het 3PL uitsluitend gedefinieerd door

(5.19) en de bijkomende aanname van lokale stochastische onafhankelijkheid. De interpretatie van c_i als raadparameter maakt geen deel uit van het model. Indien we data hebben die uitstekend beschreven worden door het 3PL, volgt daar niet logisch uit dat er geraden is. Het zou bijvoorbeeld zo kunnen zijn dat personen die het juiste antwoord niet echt kennen, toch een of andere, verkeerde, redenering volgen die met een kans c_i in het juiste antwoord resulteert. Het is nuttig om na te gaan of we niet een model van het cognitieve functioneren kunnen opstellen dat dezelfde voorspellingen maakt als het 3PL. Daartoe definiëren we een nieuwe functie die we zullen aanduiden met het symbool h_i:

$$h_i(\theta) = \frac{\exp[a_i(\theta - \beta_i)]}{1 + \exp[a_i(\theta - \beta_i)]}. \qquad (5.20)$$

De functie $h_i(\theta)$ komt dus overeen met de breuk in het rechterlid van (5.19). Het is duidelijk dat $0 < h_i(\theta) < 1$. We interpreteren deze functie als de kans dat een persoon met vaardigheid θ het antwoord op het item kent. Voorts veronderstellen we dat, indien het juiste antwoord 'geweten' wordt, het ook daadwerkelijk gegeven wordt. Dat wil zeggen dat we hier aannemen dat de persoon zich niet kan vergissen, later zullen we onderzoeken wat er gebeurt als we deze assumptie laten vallen. Indien de persoon het antwoord niet kent, wordt er met een kans $1 - h_i(\theta)$ overgegaan op raden en het juiste antwoord wordt dan geraden met kans c_i. De verschillende gebeurtenissen en kansen zijn schematisch weergegeven in tabel 5.1.

Tabel 5.1
Een cognitief model voor het beantwoorden van meerkeuze-items

Gebeurtenis	Kans	Antwoord
Kent antwoord en vergist zich niet	$1 \times h_i(\theta) = h_i(\theta)$	Juist
Kent antwoord maar vergist zich	$0 \times h_i(\theta) = 0$	Fout
Kent antwoord niet maar raadt juist	$c_i \times [1 - h_i(\theta)]$	Juist
Kent antwoord niet en raadt verkeerd	$(1 - c_i) \times [1 - h_i(\theta)]$	Fout

De kans op een juist antwoord is dus de som van de twee kansen uit tabel 5.1 die tot een juist antwoord leiden:

$$P(X_i = 1 \mid \theta) = h_i(\theta) + c_i[1 - h_i(\theta)]$$
$$= c_i + (1 - c_i)\, h_i(\theta) = f_i(\theta).$$

Het eenvoudige cognitieve model leidt dus tot het 3PL. Binnen dit cognitieve model kunnen we dan ook de kans berekenen dat een juist antwoord door raden tot stand is gekomen:

$$P(\text{raden} \mid X_i = 1, \theta) = \frac{c_i[1 - h_i(\theta)]}{h_i(\theta) + c_i[1 - h_i(\theta)]}. \tag{5.21}$$

Het rechterlid van (5.21) is niet te vereenvoudigen, omdat de afhankelijkheid van θ er in aanwezig blijft. Dit betekent dat we geen nauwkeurige uitspraak kunnen doen over de hoeveelheid juiste antwoorden die door raden tot stand zijn gekomen in een willekeurige steekproef van antwoordpatronen. We kunnen het wel indien we de verdeling van θ kennen. Indien $g(\theta)$ de dichtheidsfunctie is van θ vinden we:

$$P(\text{raden} \mid X_i = 1) = \int_{-\infty}^{\infty} \frac{c_i[1 - h_i(\theta)]}{h_i(\theta) + c_i[1 - h_i(\theta)]}\, g(\theta)\, d\theta. \tag{5.22}$$

De dichtheidsfunctie $g(\theta)$ maakt echter geen deel uit van het 3PL, maar moet er aan toegevoegd worden.

Samenvattend kunnen we zeggen dat het cognitieve model, in de mate dat het een min of meer realistische voorstelling van cognitieve processen geeft, een rechtvaardiging is van het 3PL, maar dat het niet door het 3PL wordt geïmpliceerd. We keren nu terug naar een verdere analyse van het 3PL.

In het Raschmodel hebben we de moeilijkheidsgraad van een item omschreven als de hoeveelheid vaardigheid die nodig is om een kans te hebben van precies 0.5 om het item juist te beantwoorden. Deze interpretatie van de itemparameter geldt niet meer in het 3PL. Indien θ gelijk is aan β_i krijgen we

$$f_i(\beta_i) = c_i + (1 - c_i) \times 0.5 = 0.5 + \frac{c_i}{2}. \tag{5.23}$$

De interpretatie van β_i als moeilijkheidsparameter is dus niet zo overtuigend als in het Raschmodel, door de afhankelijkheid van c_i die in (5.23) tot uiting komt. Toch wordt de parameter β_i in de literatuur aangeduid als moeilijkheidsparameter.

Wellicht ten overvloede vermelden we nog even dat het model (5.19) niet geïdentificeerd is. Het linkerlid van (5.19) verandert niet als bij de β-parameters en bij θ een willekeurige constante c wordt opgeteld. Het nulpunt van de schaal kan dus, net

als bij het Raschmodel, vrij gekozen worden. Bovendien kunnen we zowel θ als β_i met een willekeurige positieve constante vermenigvuldigen, als we te zelfder tijd a_i door die constante delen. Dit betekent dat we de eenheid van de schaal willekeurig kunnen kiezen. Die keuze kunnen we bijvoorbeeld maken door te eisen dat $a_1 = 1$. De parameters c_i liggen op een absolute schaal en kunnen niet getransformeerd worden.

Tenslotte nog een terminologische kwestie. Het rechterlid van (5.19) kan niet teruggebracht worden tot de standaardvorm van de logistische functie. Strikt genomen is het 3PL dus geen logistisch model, maar in de literatuur wordt het wel zo genoemd. Wij zullen ons aan dit gebruik conformeren.

Het 2PL kan men opvatten als een speciaal geval van het 3PL: het is gegeven door in (5.19) de parameter c_i gelijk te stellen aan 0 voor alle items. De itemresponsfunctie in het 2PL valt dus samen met de functie $h_i(\theta)$ die in (5.20) is gedefinieerd. Wanneer we verderop het 2PL onderzoeken, zullen we echter niet het functiesymbool h gebruiken maar f om de itemresponsfunctie aan te duiden.

Parameterschatting in het 2PL en het 3PL

Bij een eerste beschouwing van (5.19) zou men de volgende redenering kunnen volgen. Het 2PL is een speciaal geval van het 3PL en het Raschmodel is op zijn beurt weer een speciaal geval van het 2PL, dat ontstaat door alle discriminatieparameters aan elkaar gelijk te stellen. Als we dus altijd werken met het 3PL, merken we vanzelf wel of de raadparameters gelijk zijn aan 0 of niet en of de discriminatieparameters gelijk zijn of ongelijk. De realiteit is niet zo eenvoudig. Het schatten van de parameters in het 2PL en het 3PL is namelijk heel wat moeilijker dan in het Raschmodel en bovendien is het uitmaken of het 2PL of het Raschmodel passende modellen zijn niet eenvoudig. Om deze moeilijkheden te illustreren zullen we ons in eerste instantie beperken tot het 2PL. Later zullen we nog enkele beschouwingen toevoegen over het 3PL.

De log-aannemelijkheidsfunctie gegeven een antwoordpatroon $\boldsymbol{x}$ voor het 2PL werd reeds besproken in hoofdstuk 4, formule (4.61). We herhalen deze formule hier:

$$\ln L(\beta, \boldsymbol{a}, \theta; \boldsymbol{x}) = \theta \sum_i a_i x_i - \sum_i x_i a_i \beta_i - \sum_i \ln\{1 + \exp[a_i(\theta - \beta_i]\}. \qquad (5.24)$$

Het is direct duidelijk dat CML als schattingsprocedure is uitgesloten. We kunnen niet conditioneren op $\sum_i a_i x_i$ omdat deze grootheid afhankelijk is van de onbekende parameters a_i. Van de schattingsmethoden die in hoofdstuk 4 werden besproken, blijven dus alleen JML en MML over. Bij de JML-methode hebben we een analoog probleem

als bij het Raschmodel. Door de aanwezigheid van de incidentele parameters θ_v kunnen we geen beroep doen op standaardresultaten uit de statistiek. Met name weten we niet of de JML-schatters wel consistent zijn. Het is niet zo dat de aanwezigheid van incidentele parameters in alle gevallen leidt tot inconsistentie van de schatters van de structurele parameters, maar als er incidentele parameters zijn en men wil toch gebruik maken van JML, dan dient men de consistentie van de schatters aan te tonen. Een dergelijk bewijs voor het 2PL is in de IRT-literatuur echter nog nooit gegeven. Hierna geven wij de schets van een bewijs dat JML in het 2PL geen consistente schatters oplevert van de β-parameters en ook niet van de discriminatieparameters. We doen dit aan de hand van het eenvoudigst mogelijke geval met $k = 2$ items.

Bij twee items zijn er maar vier mogelijke antwoordpatronen: (0 0), (0 1), (1 0) en (1 1). Bij een steekproef van n personen kunnen we de observaties dus handig samenvatten door de frequenties van die vier antwoordpatronen te hanteren. Deze frequenties worden aangeduid als respectievelijk n_{00}, n_{01}, n_{10} en n_{11}. Het aantal itemparameters dat in het 2PL moet worden geschat is $2(k-1)$, $k-1$ β-parameters en $k-1$ discriminatieparameters. Omdat we met JML werken en dus met elke persoon een parameter associëren, moeten bovendien nog n persoonsparameters geschat worden. We kiezen de normering van de schaal zo dat $\beta_1 = 0$ en $a_1 = 1$. We moeten dus $\beta_2, a_2, \theta_1, ..., \theta_n$ schatten. De schattingen kunnen we met standaardtechnieken berekenen, door de partiële afgeleiden van de log-aannemelijkheidsfunctie gelijk te stellen aan 0 en de aldus ontstane vergelijkingen op te lossen. Voor het geval $k = 2$ kan een expliciete oplossing gevonden worden. We zullen de details van de afleiding niet bespreken, maar geven alleen het resultaat. Daarbij veronderstellen we dat n_{01} en n_{10} beide van 0 verschillen.

(1) Personen met hetzelfde antwoordpatroon krijgen dezelfde schatting van θ. De schattingen van de n θ-parameters kunnen dus niet meer dan vier verschillende waarden aannemen, die we zullen aan duiden als $\hat{\theta}_{00}$, $\hat{\theta}_{01}$, $\hat{\theta}_{10}$ en $\hat{\theta}_{11}$.

(2) $\hat{\theta}_{00}$ en $\hat{\theta}_{11}$ bestaan niet. Dit wil zeggen dat er geen reële getallen bestaan die we voor die twee schatters kunnen invullen zodat aan de schattingsvergelijkingen is voldaan. Dit impliceert eigenlijk dat we het probleem iets anders moeten formuleren en zeggen dat we onze schattingen gaan baseren op de $n_{01} + n_{10}$ antwoordpatronen die precies één item juist hebben.

(3) $\hat{\theta}_{01} = \hat{\theta}_{10} = \ln(n_{10}/n_{01})$, dus alle personen met één juist antwoord krijgen dezelfde schatting van θ.

(4) $\hat{a}_2 = 1$, of iets algemener gezegd, a_2 wordt geschat op precies dezelfde waarde die we aan a_1 hebben toegekend.

(5) $\hat{\beta}_2 = 2\ln(n_{10}/n_{01})$.

Uit resultaat (4) volgt direct dat de discriminatieparameters niet consistent geschat worden: wat ook de steekproefomvang is en wat de echte waarden van de discriminatieparameters ook zijn, ze worden steeds als even groot geschat. Om de inconsistentie van de schatter van β_2 aan te tonen, beschouwen we een speciaal geval van het 2PL waar de discriminatieparameters aan elkaar gelijk zijn. Dan krijgen we voor β_2 natuurlijk dezelfde schatter die in resultaat (5) is gegeven. Maar dit speciale geval van het 2PL is niets anders dan het Raschmodel en de schatter in (5) is ook precies dezelfde als de JML-schatter van β_2 in het Raschmodel (Fischer, 1974, p. 260), waarvan is aangetoond dat hij inconsistent is. Het besluit is dus dat de itemparameters in het 2PL niet consistent geschat worden. Dit resultaat sluit niet uit dat de schatters bij een andere *k* misschien wel consistent zijn, doch dit zou dan moeten worden aangetoond.

Het niet consistent zijn van schatters heeft grote gevolgen voor de toepassingen van een model. Losweg betekent het niet-consistent zijn, dat de schattingen systematisch gaan afwijken van de werkelijke waarden en dat die systematische fout niet verholpen kan worden door de steekproef groter te maken. Dit hoeft in bepaalde opzichten niet erg te zijn. Als de systematische fout klein is, zouden we daar genoegen mee kunnen nemen. Zo blijkt in het Raschmodel bijvoorbeeld, dat de systematische fout kleiner wordt als *k* toeneemt. Bovendien kan men in het Raschmodel een correctie aanbrengen op de JML-schattingen door ze te vermenigvuldigen met $(k-1)/k$. Uit simulatiestudies blijkt dat de aldus gecorrigeerde JML-schattingen erg goed overeenkomen met de CML-schattingen die wel consistent zijn. Dit is een nuttig resultaat, maar het lost slechts een deelprobleem op. Alle theorie die in hoofdstuk 4 is behandeld over standaardfouten en de asymptotische verdeling van toetsingsgrootheden, is niet zonder meer geldig in het geval dat de ML-schatters niet consistent zijn. Men kan natuurlijk in een concrete toepassing de geobserveerde informatiematrix inverteren en de elementen op de diagonaal beschouwen als schatters van de variantie, doch men kent niet meer de eigenschappen van die schatters en die zouden wel eens erg onaantrekkelijk kunnen zijn. Het feit dat er veel publikaties zijn in de IRT-literatuur waar deze procedure wordt toegepast, kan niets veranderen aan het dubieuze karakter ervan.

Het gebruik van de MML-procedure omzeilt de problemen van de incidentele parameters. Zoals in hoofdstuk 4 reeds is benadrukt, dient men echter wel te bedenken dat MML niet alleen een procedure is, maar dat het meetmodel uitgebreid wordt met een veronderstelling over de verdeling van θ. Verder is de uiteenzetting over MML uit

hoofdstuk 4 ook van toepassing op het 2PL en het 3PL. Op de problemen van algoritmische en numerieke aard gaan we hier niet verder in. Gedetailleerde uiteenzettingen hierover kan men vinden in Bock en Aitkin (1981) en in Rigdon en Tsutakawa (1983).

Er is echter één probleem dat ogenschijnlijk veel te maken heeft met de berekening van de schattingen, maar dat een veel diepere oorzaak heeft die te maken heeft met de eigenschappen van het model. We kunnen het probleem het beste illustreren aan de hand van het 3PL. Indien we het Raschmodel toepassen, vinden we altijd dat een item met een grote *p*-waarde een kleinere geschatte moeilijkheidsparameter heeft dan een item met een kleine *p*-waarde. Men kan aantonen dat dit mathematisch noodzakelijk is, en het is ook wat we normaliter zouden verwachten. Bij het 3PL verschijnt echter een dubbelzinnigheid: een grote *p*-waarde kan wijzen op een gemakkelijk item en een kleine raadparameter maar ook op een moeilijk item met een grote raadparameter. De itemantwoorden zijn dus in zekere zin dubbelzinnig: uit de kwaliteit van het antwoord kan men de waarde van de parameters moeilijk afleiden. Of anders gezegd, de data bevatten erg weinig informatie die gebruikt kan worden om onderscheid te maken tussen moeilijkheid en raadkans. Dit heeft tot gevolg dat het vinden van het maximum van de aannemelijkheidsfunctie in het algemeen moeilijker zal zijn dan in het Raschmodel en dat de nauwkeurigheid waarmee de parameters geschat worden kleiner zal. Bovendien ontspoort de schattingsprocedure soms door een oplossing op te leveren die niet overeenkomt met het maximum van de aannemelijkheidsfunctie. Als item *i* een vierkeuze-item is, ver-wachten we dat de schatting van c_i niet al te ver zal afwijken van 0.25. Krijgen we als resultaat echter een schatting van 0.85, dan zullen we niet al te snel geneigd zijn met deze schatting genoegen te nemen. Deze problemen ontstaan dus eigenlijk omdat we de data overvragen, of vanuit een ander standpunt bekeken, omdat we te weinig informatie hebben verzameld. Indien we een betrouwbare procedure konden verzinnen waarbij de persoon bij elk itemantwoord ook aangeeft of er geraden is of niet, dan zouden we veel meer informatie hebben en we zouden ook veel nauwkeuriger kunnen schatten.

De voorgaande beschouwing geeft ook aan dat er in zekere zin grenzen zijn aan de complexiteit van IRT-modellen. Het is niet moeilijk om het cognitieve model dat in tabel 5.1 is weergegeven iets realistischer te maken, door de kans op een vergissing als men het antwoord kent niet gelijk te stellen aan 1, maar daar een nieuwe parameter d_i voor te kiezen. Dit leidt dan tot een 4PL, waarvan het in principe mogelijk is de parameters te schatten als men alleen over dichotome data beschikt. De schattingen zullen echter zo instabiel zijn dat ze in de praktijk eigenlijk niet meer bruikbaar zijn, tenzij men over gigantische steekproeven kan beschikken.

Er bestaat echter ook een andere manier om het tekort aan informatie te ondervangen, namelijk het toepassen van een schattingstechniek die afkomstig is uit de bayesiaanse statistiek. Hier voegt men zijn ongeloof dat de *c*-parameter uit het voorbeeld gelijk is aan 0.85 op een formele manier aan het model toe door middel van een a priori verdeling, die voor alle mogelijke waarden van de parameter als het ware de voorafgaande overtuiging uitdrukt dat de parameter die waarde aanneemt. Als de a priori verdeling uniform is, drukken we daarmee uit dat we eigenlijk helemaal niets weten over die parameter. Is die verdeling eentoppig met een hele kleine standaardafwijking en met modus of gemiddelde in de buurt van 0.25, dan geven we daarmee aan dat we er vrijwel zeker van zijn dat de raadkans niet ver van 0.25 zal afwijken. De observaties worden dan gebruikt om onze overtuiging te wijzigen: de gegevens en de a priori verdeling worden met elkaar gecombineerd en leveren een nieuwe verdeling van de parameter op die de a posteriori verdeling genoemd wordt en die op haar beurt weer kan fungeren als a priori verdeling voor toekomstige observaties. Als schatter van de parameter neemt men dan een of ander kenmerk van de a posteriori verdeling, zoals de modus of het gemiddelde en als maat van onzekerheid neemt men meestal de standaardafwijking van de a posteriori verdeling. Een meer technische uiteenzetting is gegeven in paragraaf 4.5 bij de behandeling van de EAP-schatter van θ in het Raschmodel. Men kan deze techniek ook toepassen bij meer parameters tegelijk, maar dan moet men een a priori verdeling specificeren voor alle parameters tegelijk. In dat geval blijkt het berekenen van de modus van de multivariate a posteriori verdeling meestal eenvoudiger te zijn dan het berekenen van het gemiddelde. Deze techniek wordt bijvoorbeeld toegepast in het computerprogramma BILOG (Mislevy & Bock, 1986) dat de parameters voor het 3PL, het 2PL en het Raschmodel schat en dat in de regel plausibele schattingen oplevert.

Hoewel het gebruiken van een bayesiaanse benadering erg elegant is en veel problemen van JML en MML omzeilt, dient men toch de nodige voorzichtigheid in acht te nemen bij het gebruik van deze techniek. Op het eerste gezicht lijkt deze benadering een element van willekeur te bevatten. Iedereen kan immers zijn eigen a priori verdeling kiezen, waardoor ook steeds, bij dezelfde data, verschillende schattingen zullen worden verkregen. De wetenschappelijke consensus zal zo ver te zoeken zijn. De bayesiaanse statistiek heeft een adequaat antwoord op dit bezwaar. Ten eerste moet de rol van de a priori verdeling niet overschat worden. Indien er maar voldoende observaties zijn, wordt de a posteriori verdeling bijna volledig bepaald door de observaties en speelt de a priori verdeling geen rol van betekenis meer. Ten tweede is de a priori verdeling bedoeld als een soort samenvatting van eerder gedane observaties en ervaringen. Als twee onderzoekers in hetzelfde domein van wetenschap actief zijn,

dezelfde literatuur lezen en vergelijkbaar onderzoek doen, kunnen hun overtuigingen in de bayesiaanse betekenis niet drastisch van elkaar verschillen. Maar dat is theorie. In de praktijk kan de misvatting optreden dat het er niet toe doet welke a priori verdeling men kiest, omdat het aantal van 200 observaties waarover men beschikt geweldig groot is vergeleken met de 25 waarop de collega of de concurrent zijn analyse uitvoerde. Of een steekproef groot genoeg is om de a priori verdeling onbelangrijk te maken, hangt af van de standaardafwijking van de a priori verdeling. Kiest men deze standaardafwijking erg klein, dan kan bij een steekproef die gevoelsmatig erg groot lijkt, de a posteriori modus zeer dicht bij de modus van de a priori verdeling liggen. Als bewijs dat men het met de a priori verdeling 'dus' bij het rechte eind had, is dit echter niet overtuigend. Men heeft bij wijze van spreken aangetoond dat men zo'n sterke overtuiging had, dat die door de 100 of 200 observaties waarover men beschikt niet wezenlijk te veranderen is. Kiest men de standaardafwijking echter te groot, dan is de a posteriori verdeling grotendeels bepaald door de observaties en gaat de schattingsprocedure erg lijken op de ML-schattingsprocedure en verliest de bayesiaanse benadering eigenlijk haar zin.

Statistische toetsen voor het 2PL en het 3PL

De behandeling van dit onderwerp kan kort zijn, om de eenvoudige reden dat er zeer weinig toetsen zijn ontwikkeld die voor deze modellen gebruikt kunnen worden. Waarom dit zo is, is niet gemakkelijk te zeggen, doch we kunnen zeker twee mogelijke redenen aangeven. De eerste reden heeft te maken met de moeilijkheid van het probleem. Alles wat in hoofdstuk 4 is gezegd over het construeren van veralgemeende X^2-toetsen had betrekking op modellen uit de exponentiële familie. Het 2PL en het 3PL behoren niet tot deze familie. Glas (1989) heeft weliswaar aangetoond dat er gelijkaardige toetsen geconstrueerd kunnen worden voor modellen buiten de exponentiële familie, zoals de R_0- en de R_{1m}-toetsen, maar de bewijsvoering is heel specifiek voor het Raschmodel en is niet zonder meer bruikbaar voor het 2PL en het 3PL.

De tweede reden heeft te maken met een verschil van instelling tussen de Europese psychometrici enerzijds en een groot gedeelte van de Amerikaanse vakgenoten. De Europese literatuur over IRT is zeer sterk beïnvloed door het werk van Rasch (1960) en Fischer (1974), waar een grote nadruk gelegd wordt op de theoretische eigenschappen die in een deugdelijk meetinstrument aanwezig moeten zijn. Dit heeft niet alleen geleid tot de prominente plaats die het Raschmodel in de IRT-literatuur

inneemt, maar ook tot een grote inspanning om statistische toetsen te ontwerpen waarmee kan worden nagegaan of aan de strenge eisen van het Raschmodel is voldaan. De Amerikaanse literatuur over IRT daarentegen is zeer sterk beïnvloed door het werk van F. Lord, die gezien zijn werkzaamheden op het toetsinstituut Educational Testing Service (ETS) een veel pragmatischer instelling had. Waar men het devies van de Europese traditie grofweg zou kunnen omschrijven als: 'maak toetsen die aan het Raschmodel voldoen', kwam Lords devies neer op: 'maak modellen die adequaat zijn voor de bestaande toetsen'. Door het wijdverspreide gebruik van meerkeuze-items is de ontwikkeling en het gebruik van het 3PL dan ook goed te begrijpen. Omdat dit model voorziet in verschillende discriminatieparameters voor de items en in een onderste asymptoot die verschillend kan zijn van 0, is er ook minder behoefte aan statistische toetsing. De twee voor de hand liggende kwetsbare plekken van het Raschmodel zijn immers modelmatig weggewerkt.

Het hierboven geschetste verschil in benadering van de IRT is natuurlijk niet absoluut en er zijn statistische toetsen ontwikkeld die van toepassing zijn voor het 2PL en het 3PL. Deze toetsen zijn besproken in paragraaf 4.3.5 als varianten van de S_i-toetsen. Bovendien is het natuurlijk mogelijk LR-toetsen te construeren waarin het 2PL of het 3PL als nulhypothese fungeert en het verzadigde multinomiale model als alternatieve hypothese. Men zou kunnen opperen dat een LR-toets waarbij het 2PL fungeert als nulhypothese en het 3PL als algemeen model of alternatieve hypothese meer onderscheidingsvermogen zal hebben. Dit is echter geen goed idee. Bij de bespreking van de LR-toetsen in hoofdstuk 4 hebben we gezien dat bij een LR-toets de parameterruimte van het beperkte model een deelruimte moet zijn van de parameterruimte in het algemene model. De eis is echter strenger. De beperkte parameter-ruimte moet helemaal binnen de algemene parameterruimte liggen. We gaan hier niet in op de precieze mathematische betekenis van 'binnen', maar we illustreren het principe met een voorbeeld. Als we het 2PL beschouwen als een speciaal geval van het 3PL, betekent dit dat we alle c_i-parameters in het 3PL fixeren op de waarde 0, maar deze waarde is de kleinste waarde die de c_i-parameters kunnen aannemen. Men zegt dat de parameters in het 2PL gefixeerd worden op de rand van de parameterruimte van het 3PL en in dit geval mag men zeker niet zonder meer aannemen dat de LR-toetsingsgrootheid chi-kwadraat verdeeld is.

5.3.2 Het éénparameter logistisch model (OPLM)

Er zijn vele varianten mogelijk op het 3PL, waarvan sommige als gevolg van moeilijkheden bij het schatten van de parameters in het algemene 3PL daadwerkelijk in de literatuur zijn toegepast. Meestal gaat het om beperkingen op de c_i-parameters. Indien in een meerkeuzetoets alle items evenveel antwoordalternatieven hebben, zou men het redelijk kunnen vinden te eisen dat alle c_i-parameters aan elkaar gelijk zijn. Deze eis komt overeen met het opleggen van k-1 lineaire restricties aan de parameters van het model, analoog aan wat gebeurt bij de moeilijkheidsparameters in het LLTM. Een verdere restrictie die soms wordt toegepast, bestaat erin die gemeenschappelijke c-parameter gelijk te stellen aan één gedeeld door het aantal antwoordalternatieven. Door deze eis verandert de status van c. Het is geen onbekende grootheid meer die uit de data moet worden geschat, maar een bekende constante. Hoewel deze twee varianten van het 3PL het schattingsprobleem sterk vereenvoudigen, is er geen mogelijkheid om CML toe te passen.

Er bestaat echter wel een mogelijkheid om dusdanige restricties op het 2PL aan te brengen dat CML wel mogelijk wordt. Indien we in (5.24) de grootheden a_i niet langer beschouwen als onbekende parameters maar als gegeven constanten, zien we dat deze speciale versie van het 2PL tot de exponentiële familie behoort en dat de gewogen score $s = \sum_i a_i x_i$ een groot- heid is die zonder meer uit de data kan worden berekend en waarop dus geconditioneerd kan worden. Hierdoor verliest a_i zijn status van parameter. Om dit essentiële onderscheid in de terminologie goed aan te geven, zullen we spreken van discriminatie-indices. Het model werd voorgesteld door Verhelst en Eggen (1989) en kreeg de naam éénparameter logistisch model (OPLM) op grond van het argument dat er per item slechts één parameter overblijft.

Bij de bespreking van het 2PL hebben we gezien dat één discriminatieparameter vrij gekozen kan worden en dat daarmee de eenheid van de schaal wordt vastgelegd. Welke waarde we kiezen doet niet ter zake. Bijgevolg is een uitspraak als: ’dit item discrimineert erg goed want zijn discriminatieparameter is gelijk aan 5’ zinloos als niet, expliciet of impliciet, gerefereerd wordt naar de eenheid van de schaal. Deze referentie is altijd aanwezig indien men verhoudingen van discriminatieparameters of -indices hanteert. Dit maakt ook duidelijk dat, indien alle discriminatie-indices met een constante worden vermenigvuldigd, het model niet verandert. Nu kunnen we die constante zo kiezen dat de resulterende indices allemaal gehele getallen zijn of willekeurig dicht door een geheel getal kunnen worden benaderd. Het houdt dus nauwelijks een beperking in als we zeggen dat de discriminatie-indices gehele getallen moeten zijn. In de verdere bespreking zullen we daar dan ook van uitgaan. Merk op dat het Raschmodel een speciaal geval is van het OPLM, waarin alle discriminatie-indices aan elkaar gelijk zijn.

Met betrekking tot de schatting van de itemparameters in het OPLM hoeven we nauwelijks iets toe te voegen aan de discussie die in hoofdstuk 4 is gewijd aan de parameterschattingen in het Raschmodel. Door een geschikte parametrisering te kiezen, blijken de formules die we gebruikt hebben bij de bespreking van het Raschmodel formeel gelijk te zijn aan de formules voor het OPLM. De conditionele aannemelijkheidsfunctie kan dus geschreven worden als:

$$\ln L(\boldsymbol{\varepsilon}; X \mid \boldsymbol{s}) = \sum_i t_i \ln \varepsilon_i - \sum_v \ln \gamma_{s_v}(\boldsymbol{\varepsilon}), \tag{5.25}$$

en die formule is precies gelijk aan (4.43). Alleen is de parameter ε_i nu gedefinieerd als

$$\varepsilon_i = \exp(-a_i \beta_i). \tag{5.26}$$

Merk op dat met s_v de gewogen score bedoeld wordt en met $t_i = \sum_v x_{vi}$ het aantal juiste antwoorden dat op item *i* is uitgebracht. De functie $\gamma_s(\boldsymbol{\varepsilon})$ is formeel gedefinieerd als

$$\gamma_s(\boldsymbol{\varepsilon}) = \sum_{\sum a_i x_i = s} \prod_i \varepsilon_i^{x_i}. \tag{5.27}$$

We geven een voorbeeld om de structuur van (5.27) te verduidelijken. Veronderstel dat $k = 4$ en de eerste drie items een discriminatie-index gelijk aan 1 hebben, maar dat $a_4 = 2$. Er zijn precies vier antwoordpatronen die een gewogen score van 2 opleveren: (1 1 0 0), (1 0 1 0), (0 1 1 0) en (0 0 0 1). De som die we nodig hebben om $\gamma_2(\boldsymbol{\varepsilon})$ uit te rekenen zal bijgevolg uit vier termen bestaan:

$$\gamma_2(\boldsymbol{\varepsilon}) = \varepsilon_1\varepsilon_2 + \varepsilon_1\varepsilon_3 + \varepsilon_2\varepsilon_3 + \varepsilon_4.$$

In tegenstelling tot de symmetrische functies die we nodig hadden bij het Raschmodel, komen in het rechterlid van bovenstaande uitdrukking niet meer alle tweetallen van parameters voor als produkt, maar alleen die combinaties van parameters die overeenkomen met een gewogen score van 2. De γ-functies zijn dus niet langer symmetrisch. Op de algoritmische problemen die opduiken bij het berekenen van die functies gaan we hier niet in. De parameterschattingen, zowel met CML als met MML, voor volledige en onvolledige designs zijn geïmplementeerd in het computerprogramma OPLM (Verhelst, Glas & Verstralen, 1993).

Voor de toetsing van het model kunnen we volstaan met een simpele verwijzing naar paragraaf 4.3: de rationale van de toetsen, maar ook hun technische uitwerking kan zonder meer toegepast worden op het meer algemene OPLM. Het is wel belangrijk, niet uit het oog te verliezen dat de vooraf gekozen discriminatie-indices deel uitmaken van het model en dus van de nulhypothese. Dit is analoog aan de situatie bij het LLTM, waar de gespecificeerde elementen van de Q-matrix eveneens deel uitmaken van de nulhypothese. De statistische toetsen hebben dus betrekking op het OPLM met de discriminatie-indices die door de gebruiker zijn gekozen. Een eventuele niet-passing van het model kan te wijten zijn aan de verkeerde specificatie van één of meer discriminatie-indices. De S_i-toetsen, maar vooral de M_i-toetsen kunnen gebruikt worden om dergelijke misspecificaties op het spoor te komen. De M_i-toetsen geven bovendien de richting aan waarin de discriminatie-index moet worden aangepast om een adequater model te krijgen. Werken met OPLM zal vaak bestaan uit het herhaaldelijk toepassen van de schattings- en toetsingsprocedures, waarbij iedere keer één of meer discriminatie-indices worden aangepast. Hoewel deze aanpassingen meestal gebeuren aan de hand van analyses op dezelfde data en er dus kanskapitalisatie kan optreden, is het belang van deze kanskapitalisatie gering als de steekproef niet te klein is. Meer beschouwingen hierover, alsook een heuristiek om plausibele waarden van de discriminatie-indices uit de data af te leiden, kan men vinden in Verhelst, Verstralen en Eggen (1991).

5.3.3 Modellen zonder de assumptie van lokale stochastische onafhankelijkheid

Overtreding van het principe van de lokale stochastische onafhankelijkheid houdt in dat de onderlinge afhankelijkheid van itemantwoorden niet verdwijnt door te conditioneren op θ. Dit betekent dat we kans op een antwoordpatroon gegeven θ niet kunnen schrijven als het produkt over items van de afzonderlijke kansen op een goed antwoord. Kelderman (1984, 1988) en Jannarone (1986) hebben een uitgebreide klasse van IRT-modellen beschreven waarin de kans op een antwoordpatroon rechtstreeks wordt gedefinieerd. We zien hier af van een complete beschrijving van deze klasse van modellen, omdat daarvoor een uitgebreid formalisme nodig is. In plaats daarvan zullen we het idee waarop een en ander gebaseerd is, toelichten aan de hand van een voorbeeld uit de klasse van modellen die door Jannarone is gedefinieerd. Stel dat een toets uit drie items bestaat. Beschouw een model waarin de kans op antwoordpatroon $\boldsymbol{x}$ gegeven θ geschreven kan worden als:

$$P(\boldsymbol{x} \mid \theta, \beta_1, \beta_2, \beta_3, \beta_{13}) = \frac{\exp\left(\sum_i x_i(\theta - \beta_i) + x_1 x_3(\theta - \beta_{13})\right)}{\sum_{\boldsymbol{y}} \exp\left(\sum_i y_i(\theta - \beta_i) + y_1 y_3(\theta - \beta_{13})\right)}, \tag{5.28}$$

waarbij het buitenste somteken in de noemer aangeeft dat de som genomen moet worden over alle mogelijke antwoordpatronen $\boldsymbol{y} = (y_1, y_2, y_3)$. In het voorbeeld heeft deze som dus acht termen. De functie van de noemer is er voor te zorgen dat de som van de kansen van alle acht antwoordpatronen gelijk is aan 1; voor de interpretatie is alleen de teller van belang. In dit model is er geen lokale stochastische onafhankelijkheid tussen de antwoordvariabelen X_1 en X_3. Dit kan formeel aangetoond worden door de formules voor $P(X_1 = 1 \mid \theta, X_3 = 1)$ en $P(X_1 = 1 \mid \theta, X_3 = 0)$ uit te schrijven zodat gedemonstreerd kan worden dat ze niet aan elkaar gelijk zijn. We kunnen echter de schending van de assumptie van lokale stochastische onafhankelijkheid ook duidelijk maken met een intuïtief argument. In de teller van (5.28) komen vier antwoordvariabelen aan bod: de drie itemantwoorden en het produkt $x_1 x_3$. Formeel kunnen we dit produkt opvatten als een vierde antwoord en dan is de teller van (5.28) niets anders dan de teller in de formule voor het Raschmodel met vier items. Doch er zijn slechts drie antwoorden geobserveerd en bijgevolg kunnen de vier itemantwoorden niet onafhankelijk zijn van elkaar. De noemer van (5.28) heeft dan ook geen 16 termen, want het produkt $y_1 y_3$ ligt volledig vast indien y_1 en y_3 gegeven zijn.

Merk op dat in dit model $\sum_i x_i + x_1 x_3$ de minimaal voldoende statistiek is voor θ. Met andere woorden, als een respondent twee items juist heeft en zowel het eerste als het derde item is goed gemaakt, is de voldoende statistiek voor de vaardigheid groter dan wanneer het eerste en het tweede item goed worden gemaakt. Het simultaan goed maken van de items een en drie levert de persoon een extra scorepunt op voor de schatting van zijn vaardigheids- parameter. De parameter β_{13} is de moeilijkheidsparameter die geassocieerd is met het be-halen van dit extra scorepunt.

Jannarone (1986) generaliseerde dit soort ideeën naar een zeer algemeen model. De parameters in dit model zijn te schatten met de CML-methode en er zijn toetsingsprocedures mogelijk die gebaseerd zijn op statistieken met een bekende asymptotische verdeling, in de lijn van de toetsingsprocedures die in hoofdstuk 4 zijn uiteengezet.

De modellen die door Kelderman (1984, 1988) zijn ontwikkeld, lijken erg veel op de modellen van Jannarone. Het essentiële verschil bestaat erin dat bij Kelderman de score gedefinieerd is als het aantal juiste itemantwoorden en niet meer afhangt van het produkt. In het voorgaande voorbeeld is de score 2 indien de persoon twee items juist heeft beantwoord, ongeacht welke twee dat zijn. Voor het voorbeeld (5.28) is de kans in Keldermans benadering gegeven door

$$P(\boldsymbol{x} \mid \theta, \beta_1, \beta_2, \beta_3, \beta_{13}) = \frac{\exp\left(\sum_i x_i(\theta - \beta_i) - x_1 x_3 \beta_{13}\right)}{\sum_{\boldsymbol{y}} \exp\left(\sum_i y_i(\theta - \beta_i) - y_1 y_3 \beta_{13}\right)}. \qquad (5.29)$$

Beide formules, (5.28) en (5.29), lijken erg op elkaar en het is ook niet zonder meer duidelijk wat de verschillen in interpretatie tussen beide benaderingen betekenen en of deze verschillen in de praktijk belangrijk zijn. De CML-procedure is in Keldermans benadering echter gemakkelijker toe te passen dan in Jannarones modellen, omdat de score onafhankelijk is van produkten van antwoordvariabelen. De klasse van modellen die Kelderman ontwikkelde is geïmplementeerd in het computerprogramma LOGIMO (Kelderman & Steen, 1988). De bestudering van Keldermans modellen is om nog een reden interessant. Kelderman bestudeerde het Raschmodel als een speciaal geval uit de klasse van de log-lineaire modellen en paste bij het schatten van de parameters ook technieken toe die veel gebruikt worden in de log-lineaire analyse.

Vooraleer we het laatste model uit deze paragraaf bespreken, moeten we nog even wat dieper ingaan op het begrip lokale stochastische onafhankelijkheid. In de definitie refereert het begrip 'lokaal' naar het feit dat er geconditioneerd wordt op de persoonsparameter θ. Op het ogenblik dat de vaardigheid van de persoon verandert gedurende het maken van de toets, bijvoorbeeld ten gevolge van een leerproces of als gevolg van vermoeidheid of verveling is niet meer duidelijk op welke manier we nog van lokale stochastische onafhankelijkheid gebruik kunnen maken. Fischer (1972) heeft een benaderingswijze voor dit probleem bedacht die veel lijkt op de benadering met fysieke en conceptuele items die in paragraaf 5.1.3 werd gehanteerd. Stel dat er na het juist beantwoorden van een item een leerproces plaatsvindt, en dat de vaardigheid toeneemt met α. Bij het beantwoorden van het zesde item beschikt persoon *v* dus over een vaardigheid $\theta_v + j\alpha$, waarin *j* het aantal correcte antwoorden is op de items 1 tot 5 en θ_v de vaardigheid bij het begin van de toetsafname. Maar dit is in de context van het Raschmodel hetzelfde als zeggen dat die persoon een vaardigheid θ_v heeft en dat het item een moeilijkheidsparameter heeft die gelijk is aan $\beta_6 + j\alpha$. We redeneren dus alsof we beschikken over zes conceptuele items in plaats van over één fysiek item. Elk conceptueel item correspondeert dus met een van de mogelijke waarden 0 tot en met 5 van *j*. Fischer heeft aangetoond dat met deze benadering geen CML-schattingen van de itemparameters en van de extra parameter α kunnen worden berekend waarna hij de hele benaderingswijze heeft opgegeven. Verhelst en Glas (1993) hebben echter aangetoond dat in het gegeven voorbeeld wel MML-schatters bestaan. Bovendien hebben zij aangetoond dat er andere situaties zijn waarin θ verandert gedurende de toetsafname, waar de CML-procedure wel kan worden toegepast.

We sluiten deze paragraaf af met een algemene beschouwing over het nut van de genoemde, misschien op het eerste gezicht nogal exotisch ogende modellen. De subtiele verschillen in interpretatie tussen de modellen van Kelderman en Jannarone kunnen de vraag doen rijzen of de vele inspanningen die onderzoekers zich getroosten om dergelijke, in het algemeen zeer ingewikkelde modellen te ontwikkelen enig praktisch nut hebben. Wij denken van wel en wel in om twee redenen.

Iedereen die enigszins bekend is met de wetenschappelijke psychologie, weet dat psychologische theorieën in elegantie en precisie niet kunnen wedijveren met bijvoorbeeld de theorieën in de natuurkunde. Een van de vele problemen waar de wetenschappelijke psychologie mee kampt, bestaat uit de vele op het eerste gezicht tegenstrijdige resultaten die in experimenten worden gevonden. De reden voor deze tegenstrijdigheden kan liggen in het gebrek aan precisie waarmee uitkomsten worden voorspeld, of in subtiele redeneringsfouten. Het construeren van formele modellen heeft het voordeel dat precieze predicties automatisch, dit wil zeggen langs wiskundige weg, uit een klein aantal veronderstellingen volgen. Het gevaar van subtiele fouten in de redenering is hierbij veel minder groot dan bij het gebruik van de natuurlijke taal.

Een tweede reden die voor de praktijk wellicht relevanter is, illustreren we met het volgende voorbeeld. Bij het construeren van examens is het in vele gevallen onvermijdelijk dat de items geformuleerd zijn als testlets, waarbij meer dan één vraag gesteld wordt bij dezelfde stam, bijvoorbeeld een inleidende tekst. De vragen worden meestal als aparte items beschouwd. Het is duidelijk dat het veel gemakkelijker is, lokale stochastische onafhankelijk- heid te realiseren tussen antwoorden op items die bij een verschillende stam behoren, dan tussen items die tot dezelfde stam horen. Het verkeerd lezen of interpreteren van de stam kan er de oorzaak van zijn dat alle items die bij die stam horen, verkeerd worden beantwoord. Daardoor is het principe van de lokale onafhankelijkheid geschonden en dat kan er de reden van zijn dat een eenvoudig IRT-model statistisch niet houdbaar is. Als men in zo'n geval toch het Raschmodel gebruikt en bijvoorbeeld de toetsscore definieert als het aantal items juist, betekent dit niet dat die scores 'waardeloos' zijn. Het kan wel betekenen dat iemand door één enkele onoplettendheid vier of vijf punten verliest, die anders wel behaald zouden zijn. Of iets algemener gezegd, de betrouwbaarheid van het resulterende meetinstrument, en dus ook de validiteit, zullen lager zijn dan wanneer een meetmodel werd gebruikt waarbij in deze afhankelijkheid werd voorzien, zoals de modellen van Jannarone en Kelderman. Vanuit deze optiek verschijnt het Raschmodel als een ideaaltype, waaraan in de praktijk vaak niet kan worden voldaan. De meer ingewikkelde modellen fungeren dan als een soort statistische correctieprocedure waarmee de vaak onvermijdelijke schendingen van het Raschmodel in de uiteindelijke meetresultaten kunnen worden

gecorrigeerd, analoog aan de manier waarop de covariantie-analyse gebruikt kan worden in quasi-experimenten, waar het ideaaltype van het gerandomiseerde experiment niet kan worden gerealiseerd.

5.4 Unidimensionale modellen voor polytome items

Dichotome items kunnen worden beschouwd als een speciaal geval van polytome items, waarbij het aantal antwoordcategorieën per item gelijk is aan twee. We kunnen dus ook het Raschmodel beschouwen als een speciaal geval van een model voor polytome items. Hoewel we in principe niets toe te voegen hebben aan de discussie over het Raschmodel die in hoofdstuk 4 is gevoerd, kunnen we bepaalde aspecten iets anders belichten, zodat de veralgemening naar modellen voor polytome items gemakkelijker wordt.

Het eerste aspect heeft te maken met het aantal responsfuncties per item dat nodig is om het model te definiëren. Omdat er twee antwoordcategorieën zijn, kunnen we in principe twee responsfuncties onderscheiden: de kans op een juist antwoord en de kans op een fout antwoord, beiden als functie van de latente variabele θ. Omdat de som van beide functies voor elke waarde van θ gelijk moet zijn aan 1, ligt de tweede functie volledig vast als de eerste gespecificeerd is. Er zijn dus wel twee functies maar er is slechts één onafhankelijke functie. Indien een item $m > 2$ antwoordcategeorieën heeft, kunnen we een responsfunctie beschouwen voor elk van de m categorieën, maar de som van deze m functies is de constante functie 1, zodat er slechts $m - 1$ onafhankelijke functies zijn. Deze functies dragen de naam categorieresponsfuncties. De itemresponsfunctie in het Raschmodel is dus de categorie- responsfunctie voor categorie 1.

Het tweede aspect betreft het aantal parameters per item. Men zou kunnen redeneren dat het natuurlijk is een parameter te associëren met elke categorie. Deze parameter zou dan als het ware de aantrekkingskracht uitdrukken die elke categorie uitoefent op de persoon die het item beantwoordt. Het is inderdaad mogelijk het Raschmodel op die manier op te schrijven:

$$P(X_i = 1|\theta) = \frac{\exp(1\,\theta - \eta_{i1})}{\exp(0\,\theta - \eta_{i0}) + \exp(1\,\theta - \eta_{i1})} = \frac{\exp(\theta - \eta_{i1})}{\exp(-\eta_{i0}) + \exp(\theta - \eta_{i1})}, \qquad (5.30)$$

waarin de coëfficiënten 1 en 0 van θ in het middelste lid van (5.30) het verschillende gewicht uitdrukken dat de twee antwoorden hebben met betrekking tot de latente

variabele θ. Het linkerlid van (5.30) blijft onveranderd indien in het rechterlid teller en noemer worden vermenigvuldigd met een constante die verschilt van nul. Kiezen we nu $\exp(\eta_{i0})$ als constante en definiëren we

$$\beta_i = \eta_{i1} - \eta_{i0}, \tag{5.31}$$

dan kunnen we (5.30) herschrijven als

$$P(X_i{=}1\,|\,\theta) = \frac{\exp[\theta - (\eta_{i1} - \eta_{i0})]}{1 + \exp[\theta - (\eta_{i1} - \eta_{i0})]} = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)}. \tag{5.32}$$

De parameter β_i kan dus geïnterpreteerd worden als het verschil tussen twee categorie-parameters. Deze parameters zelf zijn echter niet schatbaar.

Merk op dat de definitie van β_i in (5.31) niet dwingend is. We hadden net zo goed teller en noemer van het rechterlid van (5.30) kunnen vermenigvuldigen met $\exp(\eta_{i1})$ en dit resulteert in

$$P(X_i{=}1\,|\,\theta) = \frac{\exp(\theta)}{\exp(\beta_i) + \exp(\theta)}, \tag{5.33}$$

maar dit is precies hetzelfde als (5.32).

Het derde aspect is impliciet reeds aan de orde gekomen in het middelste lid van (5.30), waar we de coëfficiënten van θ expliciet hebben opgeschreven. Een antwoord $X_i = 1$ resul-teert in een coëfficiënt 1 en een antwoord $X_i = 0$ heeft coëfficiënt 0. Dat wil zeggen dat de ordening van de coëfficiënten samenvalt met de ordening van de antwoordcategorieën en dat betekent dat de categorieën als geordende categorieën worden geïnterpreteerd. Het feit dat de coëfficiënten hier gelijk zijn aan de antwoorden is een extra eis die het Raschmodel aan de data oplegt. In het 2PL of OPLM is de ordening wel bewaard, doch de gelijkheid is opgegeven.

5.4.1 Het partial credit model (PCM)

Gebruik makend van de drie voorgaande opmerkingen is de veralgemening van het Rasch-model tot een model voor polytome items voor de hand liggend. Het enige dat we moeten doen is nog een paar afspraken maken over de notatie. De categorieresponsfuncties zullen we aanduiden als $f_{ij}(\theta)$, waarbij de eerste index het item aanduidt en de tweede index de categorie. We hoeven daarbij niet aan te nemen dat elk item evenveel antwoordcategorieën heeft. Het aantal antwoordcategorieën per

item zullen we aanduiden als $m_i + 1$, waarbij de 'waarden' van de categorieën de opeenvolgende gehele getallen $0, 1, ..., m_i$ zijn. De veralgemening van (5.30) is dan gegeven door

$$f_{ij}(\theta) = P(X_i = j \mid \theta) = \frac{\exp(j\theta - \eta_{ij})}{\sum_{h=0}^{m_i} \exp(h\theta - \eta_{ih})}, \quad (j = 1, ..., m_i). \tag{5.34}$$

Voeren we nu de volgende herparametrisering in die analoog is aan (5.31):

$$\begin{aligned}
&\beta_{i0} = \eta_{i0} - \eta_{i0} = 0 \\
&\beta_{i1} = \eta_{i1} - \eta_{i0} \\
&\beta_{i2} = (\eta_{i2} - \eta_{i0}) - (\eta_{i1} - \eta_{i0}) = \eta_{i2} - \eta_{i1} \\
&. \\
&. \\
&\beta_{ij} = \eta_{ij} - \eta_{i,j-1} \\
&. \\
&. \\
&\beta_{i,m_i} = \eta_{i,m_i} - \eta_{i,m_i-1}
\end{aligned} \tag{5.35}$$

dan kan (5.34) geschreven worden als

$$f_{ij}(\theta) = \frac{\exp\left[j\theta - \sum_{g=0}^{j} \beta_{ig}\right]}{\sum_{h=0}^{m_i} \exp\left[h\theta - \sum_{g=0}^{h} \beta_{ig}\right]} = \frac{\exp\left[j\theta - \sum_{g=1}^{j} \beta_{ih}\right]}{1 + \sum_{h=1}^{m_i} \exp\left[h\theta - \sum_{g=1}^{h} \beta_{ig}\right]}, \tag{5.36}$$

waarin het rechterlid gelijk is aan het middelste lid omdat $\beta_{i0} = 0$. (De som-zonder-termen $\sum_{g=1}^{0} \beta_{ig}$ die voorkomt in geval $j = 0$, wordt daarbij gedefinieerd als 0.) Het model heeft dus maar m_i vrije parameters per item want de parameterisering is zo gekozen dat $\beta_{i0} = 0$. Het model in zijn vorm (5.34) is voorgesteld door Andersen (1977), waarbij de achterliggende gedachte het ontwikkelen was van een veralgemening van het Raschmodel waarbij de score $s = \sum_i x_i$ een voldoende steekproefgrootheid voor θ is. De equivalente vorm (5.36) is door Masters (1982) voorgesteld onder de naam

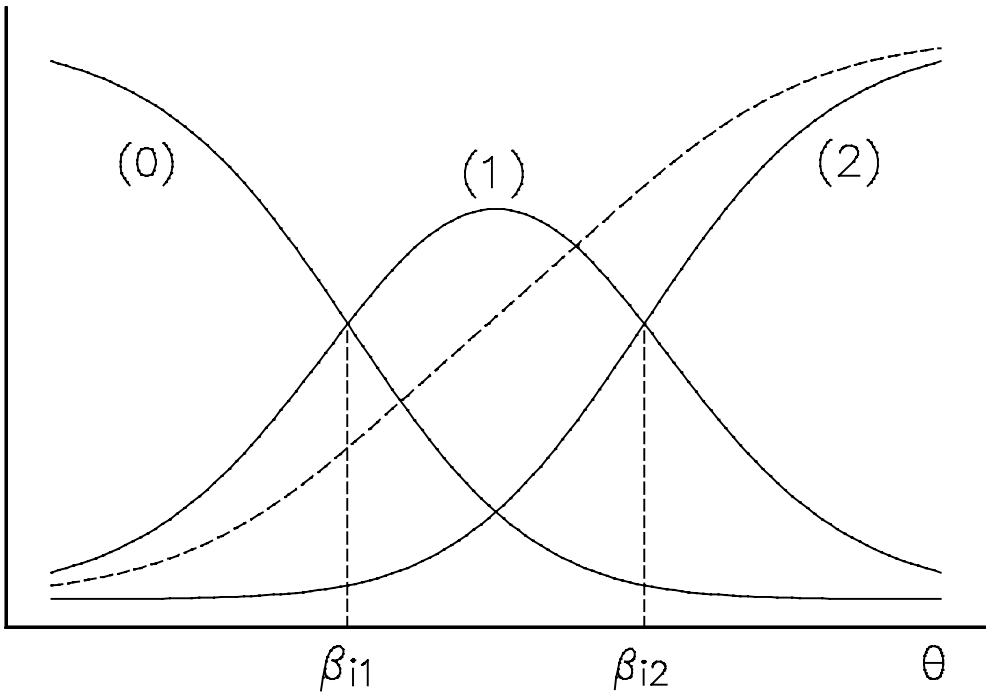
partial credit model (PCM). Om deze naam te begrijpen beschouwen we het volgende rekenitem dat ook door Masters werd gebruikt:

Bereken $\sqrt{7.5/0.3 - 16}$.

Om dit item correct op te lossen moeten drie bewerkingen in de juiste volgorde correct worden uitgevoerd, een deling, een aftrekking en een worteltrekking. De achterliggende idee was om aan elke correct uitgevoerde stap een 'partial credit' toe te kennen. Men kon dus 0, 1, 2 of 3 punten verdienen bij de beantwoording van dit item. De idee van Masters was om voor elke stap op een of andere manier het Raschmodel te gebruiken. Indien we (5.36) gebruiken om de kans $P(X_i = j \mid \theta, X_i = j \text{ of } X_i = j - 1)$ te bepalen, dan krijgen we

$$P(X_i = j \mid \theta, X_i = j \text{ of } X_i = j - 1) = \frac{\exp(\theta - \beta_{ij})}{1 + \exp(\theta - \beta_{ij})}, \quad (j = 0,...,m_i). \tag{5.37}$$

Masters vertrok van (5.37) en toonde aan dat (5.36) daaruit volgt. Hoewel de benadering van Masters elegant is, dient men zich toch te hoeden voor twee conclusies die voor de hand lijken te liggen, maar die niet gerechtvaardigd zijn. De eerste betreft de betekenis van de parameters. Men zou kunnen denken dat in het voorgaande voorbeeld de parameter β_{i2} de moeilijkheid aangeeft van de aftrekking 25-16. Deze conclusie is echter onjuist omdat de waarde van deze parameter ook beïnvloed wordt door de moeilijkheid van de daaropvolgende stap, de worteltrekking. In het algemeen kan men dus de parameters niet interpreteren als de moeilijkheid van de itemstappen. Molenaar (1983) heeft aan dit probleem een uitvoerige discussie gewijd. Een tweede misvatting ontstaat indien men denkt dat het PCM alleen geldig kan zijn bij items die in stapjes kunnen worden onderverdeeld. In feite treedt hier hetzelfde probleem op als we besproken hebben bij het 3PL. De stapjesrationale van Masters is een cognitief model dat tot het PCM leidt, maar het omgekeerde volgt niet noodzakelijk, net zo min als uit het 3PL het cognitief model volgt dat in paragraaf 5.3.1 werd besproken. Voor een voorbeeld waar de stapjesidee zeker niet van toepassing is, maar het PCM wel, zie Verhelst en Verstralen (1991). De interpretatie van de categorieparameters kunnen we het beste begrijpen aan de hand van figuur 5.5 waar de categorieresponsfuncties en de itemregressiefunctie zijn getekend voor een item i met $m_i = 2$. De categorieën zijn tussen haakjes aangeduid in de figuur.



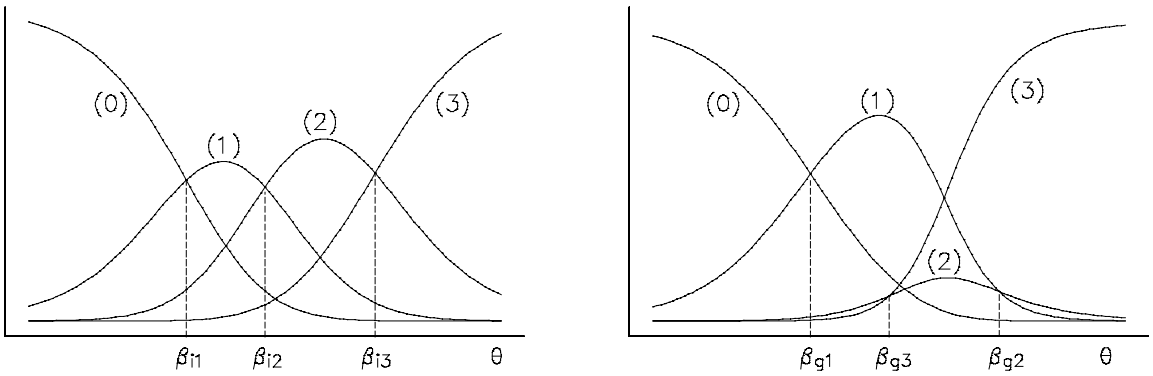
Figuur 5.5
Categorieresponsfuncties voor een item met drie antwoordcategorieën

De parameter β_{i1} geeft aan waar de responscurven voor categorie 1 en 0 elkaar snijden en de parameter β_{i2} komt overeen met het snijpunt van de categorieën 1 en 2. In het algemeen is de parameter β_{ij} die waarde van de latente variabele θ waarvoor de categorieën j en $j-1$ een even grote kans hebben gekozen te worden. Merk op dat dit ook geldt in het Raschmodel. De itemparameter β_i kunnen we ook interpreteren als de categorieparameter β_{i1}, dus als die waarde van θ waar beide categorieën een even grote kans hebben. Omdat er slechts twee categorieën zijn, is die kans gelijk aan 0.5.

De curve in stippellijnen in figuur 5.5 is een kleine modificatie van de itemregressiefunctie. Het is de curve van de functie $\mathscr{E}(X_i \mid \theta)/m_i$, die men de gestandaardiseerde itemregressie-functie kan noemen. De categorieresponscurve voor de middelste categorie is eentoppig. In het algemeen geldt in het PCM dat de curve voor categorie 0 monotoon dalend is in θ, de curve voor categorie m_i is monotoon stijgend en alle andere curven zijn eentoppig. De item-regressiefunctie echter is monotoon stijgend en dat is de reden waarom we items die aan het PCM voldoen monotone items noemen.

In figuur 5.5 is duidelijk dat categorie 1 de grootste kans heeft als θ in het interval (β_{i1}, β_{i2}) ligt. De uitspraak 'categorie j $(j = 1,..., m_i - 1)$ is de modale categorie in het interval $(\beta_{ij}, \beta_{i,j+1})$' is slechts juist indien men beseft dat dit interval alleen bestaat

indien $\beta_{ij} < \beta_{i,j+1}$ en dat deze ongelijkheid niet door het model verondersteld wordt. In figuur 5.6 zijn de categorieresponscurven afgebeeld voor twee items *i* en *g*. Voor item *i* geldt dat $\beta_{i1} < \beta_{i2} < \beta_{i3}$, maar voor item *g* geldt dat $\beta_{g2} > \beta_{g3}$.



Figuur 5.6
Geordende en niet-geordende categorieparameters

Voor item *i* geldt voor alle categorieën dat ze modaal, dat is het waarschijnlijkst, zijn in een bepaald interval van θ. Voor item *g* geldt dit niet, want categorie 2 is nooit de meest waarschijnlijke categorie. Merk op dat de waarden van θ waarvoor de categorierespons- functies van de verschillende categorieën hun grootste waarde bereiken wel degelijk geordend zijn in dezelfde volgorde als de categorieën. Zo geldt voor beide items in figuur 5.6 dat de θ-waarde waar categorie 2 haar grootste kans bereikt, groter is dan de θ-waarde waar categorie 1 haar grootste kans bereikt.

Het schatten van de parameters in het PCM kan met CML of MML gebeuren. Om de schattingsvergelijkingen op een elegante manier te kunnen opschrijven, voeren we een indicatorvector $\boldsymbol{Y}_{vi}$ in die m_i elementen bevat. Indien de antwoordvariabele X_{vi} gelijk is aan 0, zijn alle m_i elementen van $\boldsymbol{Y}_{vi}$ eveneens gelijk aan 0. Indien $X_{vi} = j$, dan is het *j*-de element van $\boldsymbol{Y}_{vi}$ gelijk aan 1, de andere elementen zijn gelijk aan 0. De vectoren $\boldsymbol{Y}_{vi}$ bevatten dus precies dezelfde informatie als de oorspronkelijke antwoordvariabelen. De elementen van de vector $\boldsymbol{Y}_{vi}$ zullen we in het algemeen aanduiden als Y_{vij}. Bijvoorbeeld, indien $m_i = 4$, dan geldt

$$X_{vi} = 3 \Leftrightarrow \boldsymbol{Y}_{vi} = (0, 0, 1, 0).$$

De geobserveerde antwoorden van persoon *v* kunnen we dus schrijven als één lange vector $\boldsymbol{Y}_v$ door alle vectoren $\boldsymbol{Y}_{vi}, (i = 1,...,k)$ gewoon achter elkaar te schrijven. De matrix *Y* van observaties krijgen we dan door de *n* vectoren $\boldsymbol{Y}_v$ in een tabel onder

elkaar te schrijven. Door gebruik te maken van het axioma van de lokale stochastische onafhankelijkheid kan de log-aannemelijkheidsfunctie gegeven één enkele vector $\boldsymbol{Y}_v$ geschreven worden als

$$\ln L(\theta_v, \beta; \boldsymbol{y}_v) = s_v\theta_v + \sum_{j=1}^{m_i} y_{vij}\left(-\sum_{g=1}^{j}\beta_{ig}\right) - \sum_{i=1}^{k}\ln\left(1 + \sum_{h=1}^{m_i}\exp[h\theta_v - \sum_{g=1}^{h}\beta_{ig}]\right), \qquad (5.38)$$

waarin

$$s_v = \sum_{i}^{k}\sum_{j}^{m_i} j\, y_{vij} = \sum_{i}^{k} x_{vi}$$

de score is van persoon *v*, dat wil zeggen het totaal aantal 'punten' dat persoon *v* behaald heeft. Definiëren we nu

$$t_{ij} = \sum_{v} y_{vij},$$

en maken we gebruik van (5.35), dan kan de log-aannemelijkheidsfunctie gegeven de antwoorden van *n* geschreven worden als

$$\ln L(\theta, \beta; Y) = \sum_{v} s_v\theta_v + \sum_{j=1}^{m_i} t_{ij}(-\eta_{ij}) - \sum_{v}\sum_{i=1}^{k}\ln\left(1 + \sum_{h=1}^{m_i}\exp(h\theta_v - \eta_{ih})\right). \qquad (5.39)$$

Het is duidelijk dat (5.39) een log-aannemelijkheidsfunctie is uit de exponentiële familie en dat bovendien kan geconditioneerd worden op de voldoende steekproefgrootheid voor θ_v. Op analoge wijze als bij het Raschmodel en bij het OPLM voor dichotome data kan de conditionele log-aannemelijkheidsfunctie geschreven worden als

$$\ln L(\boldsymbol{\varepsilon}; X \mid \boldsymbol{s}) = \sum_{i}^{k}\sum_{j}^{m_i} t_{ij}\ln\varepsilon_{ij} - \sum_{v}\ln\gamma_{s_v}(\boldsymbol{\varepsilon}), \qquad (5.40)$$

waarin

$$\varepsilon_{ij} = \exp(-\eta_{ij}) = \exp(-\sum_{g=1}^{j}\beta_{ij})$$

en

$$\gamma_s(\boldsymbol{\varepsilon}) = \sum_{\Sigma_i x_i = s}\varepsilon_{ij}^{y_{ij}}. \qquad (5.41)$$

De functie $\gamma_s(\boldsymbol{\varepsilon})$ is een veralgemening van de symmetrische basisfuncties die in het Rasch- model werden gebruikt. Het rechterlid van (5.41) geeft aan dat de som genomen moet worden over alle antwoordpatronen die de score *s* opleveren. De analogie met het Raschmodel komt verder tot uiting in de conditionele schattingsvergelijkingen die we hier zonder gedetailleerde afleiding weergeven:

$$t_{ij} = \sum_v \pi_{ij|s_v} = \sum_v \frac{\varepsilon_{ij}\gamma^{(i)}_{s_v-j}(\boldsymbol{\varepsilon})}{\gamma_{s_v}(\boldsymbol{\varepsilon})}, \tag{5.42}$$

waarin $\tau_{ij|s}$ een verkorte notatie is van $P(X_i = j \mid s)$. Het superscript (*i*) bij het functiesymbool γ geeft aan dat alle categorieparameters $\varepsilon_{ij}, (j = 1,...,m_i)$ uit de argumentvector $\boldsymbol{\varepsilon}$ moeten worden weggelaten.

De schattingsvergelijkingen voor MML zijn eveneens in analogie met het Raschmodel op te stellen. We gaan er hier niet nader op in. Zowel CML-schattingen als MML-schattingen voor de parameters in het PCM kunnen met het computerprogramma OPLM worden berekend. De statistische toetsing van het PCM wordt in de volgende paragraaf besproken.

5.4.2 Generalisaties van het partial credit model

OPLM voor polytome items

Hoewel we gezien hebben dat in het PCM het aantal categorieën per item verschillend mag zijn, levert het hanteren van verschillende aantallen bij het construeren van een toets soms moeilijkheden op. Veronderstel dat een toetsconstructeur over twee items beschikt die hij graag in eenzelfde toets wil opnemen. Het eerste item leent zich uitstekend om partieel gescoord te worden, waarbij de constructeur duidelijke voorschriften heeft wanneer een antwoord 0, 1 of 2 punten verdient. Voor het andere item ligt deze partiële scoring echter niet voor de hand, zodat alleen dichotome scoring overblijft. Binnen het PCM levert een correct antwoord op het eerste item 2 punten op, terwijl een correct antwoord op het tweede item slechts 1 punt oplevert. De twee items worden dus verschillend gewogen en deze weging volgt automatisch uit het aantal antwoordcategorieën. Dergelijke automatische koppeling kan zeer contra-intuïtief zijn en een reden waarom het PCM slechte passing geeft indien er grote variabiliteit is in het aantal antwoordcategorieën per item. Een veralgemening van het model die aan dit bezwaar tegemoetkomt ontstaat door het toevoegen van een verschillend gewicht per

item. Dit gewicht duiden we aan als a_i. De itemresponsfunctie voor deze veralgemening van het PCM is gegeven door een eenvoudige verandering van (5.34):

$$f_{ij}(\theta) = P(X_i = j \mid \theta) = \frac{\exp[a_i(j\theta - \eta_{ij})]}{\sum_{h=0}^{m_i} \exp[a_i(h\theta - \eta_{ih})]}, \quad (j = 1, ..., m_i). \tag{5.43}$$

Afhankelijk van de status die men aan de grootheid a_i toekent ontstaan polytome generalisaties van twee modellen die we reeds eerder hebben besproken. Beschouwen we de grootheden a_i als onbekende parameters die uit de data moeten worden geschat, dan is (5.43) een veralgemening van het 2PL, beschouwen we ze echter als gekende indices, dan krijgen we een polytome veralgemening van het OPLM. Willen we, zoals in het voorbeeld hierboven, alle items even zwaar laten wegen, ongeacht het aantal antwoordcategorieën, dan krijgen we een speciaal geval van het OPLM waarbij de a_i proportioneel zijn met $1/m_i$. De generalisatie (5.43) waarbij de a_i behandeld worden als te schatten parameters is in de literatuur niet beschreven als een unidimensionaal model. In paragraaf 5.5 zullen we echter zien dat het weer opduikt als een speciaal geval van een multidimensionaal model.

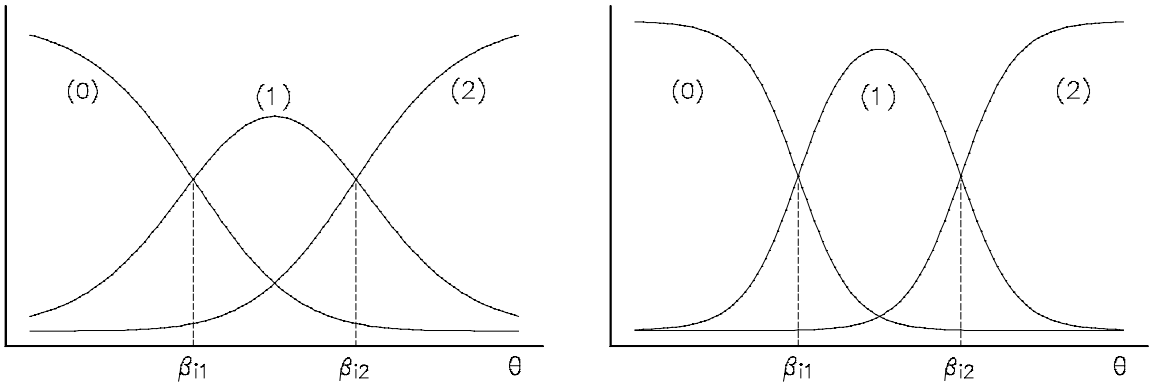
De generalisatie (5.43) waarbij de a_i bekende constanten zijn, die bovendien alleen gehele waarden aannemen, zullen we verder kortheidshalve aanduiden als het polytome OPLM. Schattingen van de parameters, zowel met CML als met MML, kunnen met het computer- programma OPLM berekend worden. Voor technische details verwijzen we naar Verhelst, Glas en Verstralen (1993).

De statistische toetsen voor het polytome OPLM en dus ook voor het PCM, zijn veralgemeningen van de statistische toetsen voor het Raschmodel en spreken meestal voor zich. Zo is bijvoorbeeld de benaderende kwadratische vorm R_{1c}^* die in (4.101) werd gegeven in de context van het Raschmodel, in het geval van het polytome OPLM gegeven door

$$R_{1c}^* = \sum_{q=1}^{r} \sum_{i=1}^{k} \sum_{j=1}^{m_i} \frac{\left[\sum_{s \in G_q} n_s(p_{ij|s} - \hat{\pi}_{ij|s})\right]^2}{\sum_{s \in G_q} n_s \hat{\pi}_{ij|s}(1 - \hat{\pi}_{ij|s})}, \tag{5.44}$$

waarin de scores worden opgedeeld in r scoregroepen G_q, $(q = 1,...,r)$. Voor de M_i- en de S_i-toetsen treedt echter een complicatie op, die onmiddellijk duidelijk wordt indien

we figuur 4.7 bekijken vanuit het standpunt van modelpassing bij polytome items. De voorspelde waarden in die figuur hebben betrekking op categorie 1 van het item *i* en een systematische onder- of overschatting van de discriminatie-index wordt onmiddellijk duidelijk uit een steiler respectievelijk vlakker verloop van de geobserveerde proporties in vergelijking met de voorspelde proporties. Deze duidelijkheid gaat echter verloren indien we analoge figuren construeren voor de middencategorieën bij polytome items. Dit is goed te zien in figuur 5.7.



Figuur 5.7
Responscurven voor een polytoom item met $a_i = 1$ (links) en $a_i = 2$ (rechts)

In de figuur rechts is de discriminatie-index twee keer zo groot als in de figuur links. Stel nu dat a_i in werkelijkheid gelijk is aan 1, doch we hebben ten onrechte gesteld dat $a_i = 2$. Als we nu, analoog aan figuur 4.7 een curve construeren waarin we $\hat{\pi}_{i1|s}$ en $p_{i1|s}$ uitzetten tegen de score *s*, dan zullen voorspelde proporties ongeveer het patroon volgen van de eentoppige curve rechts in figuur 5.7 en de geobserveerde proporties zullen het patroon volgen van de middelste curve uit het linkergedeelte van figuur 5.7. Deze beschrijving is echter nog een beetje geflatteerd omdat bij verkeerde specificatie van de discriminatie-indices ook de categorieparameters systematisch verkeerd geschat worden. Kortom, afwijkingen tussen voor- spelde en geobserveerde proporties bij de middencategorieën zijn wel systematisch, doch het is helemaal niet duidelijk hoe de scores moeten gegroepeerd worden om de statistische toetsen onderscheidend vermogen te geven tegen de verkeerde specificatie van de discriminatie-indices. In het programma OPLM is een oplossing gevonden voor dit probleem door de items na de schatting te dichotomiseren. Dichotomiseren we een item met 3 antwoordcategorieën door het antwoord 0 als lage categorie te beschouwen en de antwoorden 1 en 2 als hoge categorie, dan kunnen we voor de toetsing dezelfde rationale volgen als bij dichotome items. Definiëren we nu meer in het algemeen

$$\hat{\pi}^{*}_{ij|s} = \sum_{g=j}^{m_i} \hat{\pi}_{ig|s},$$

$$p^{*}_{ij|s} = \sum_{g=j}^{m_i} p_{ig|s},$$

dan is de veralgemening van de benaderende vorm S_i^* (formule 4.98) voor het polytome geval gegeven door

$$S^{*}_{ij} = \sum_{q=1}^{r} \frac{\left[\sum_{s\in G_q} n_s(p^{*}_{ij|s} - \hat{\pi}^{*}_{ij|s})\right]^2}{\sum_{s\in G_q} n_s \hat{\pi}^{*}_{ij|s}(1 - \hat{\pi}^{*}_{ij|s})}, \qquad (j = 1,...,m_i). \tag{5.45}$$

Per item zijn dus m_i toetsen beschikbaar, één voor elke dichotomisering van het item. Dichotomisering kan ook worden toegepast voor de M_i-toetsen. Voor toepassingen van deze toetsen zij men verwezen naar hoofdstuk 7 en hoofdstuk 9.

Terzijde kan nog worden opgemerkt dat de formules (5.44) en (5.45) geen rekening houden met de covariantie tussen de schatters van de categorieparameters. Bij parameters die tot het zelfde item behoren is de covariantie in absolute waarde heel wat groter dan bij parameters die tot verschillende items behoren. In de benaderende vormen van de toetsingsgrootheden die door het programma OPLM worden berekend, wordt alleen die laatste covariantie verwaarloosd; met de eerste wordt wel rekening gehouden. De formules worden hier niet gegeven omdat ze niet louter met sommen kunnen uitgedrukt worden.

De uitbreiding van het PCM door Wilson en Masters

De schattingsvergelijkingen (5.42) in het PCM hebben niet altijd een oplossing. Een noodzakelijke voorwaarde is dat elke categorie, inclusief de nulcategorie, van elk item in de steekproef minstens één maal geobserveerd is. Indien een categorie in de steekproef niet geobserveerd is, dan gaan Wilson en Masters (1993) het model een beetje aanpassen, om de andere parameters toch te kunnen schatten. Stel dat met item *i* bij de constructie een scoringsregel is opgesteld die resulteert in vijf geordende categorieën van 0 tot 4, doch dat in de steekproef categorie 2 niet wordt geobserveerd. Het item wordt dan omgevormd tot een item met vier antwoordcategorieën, die

respectievelijk gewicht of score 0, 1, 3 en 4 krijgen. Om te zien hoe dit probleem opgelost kan worden, herschrijven we (5.43) in een iets gewijzigde vorm:

$$f_{ij}(\theta) = \frac{\exp(j a_i \theta - a_i \eta_{ij})}{\sum_{h=0}^{m_i} \exp(h a_i \theta - a_i \eta_{ih})} = \frac{\exp(A_{ij}\theta - \delta_{ij})}{\sum_{h=0}^{m_i} (A_{ih}\theta - \delta_{ih})}. \qquad (5.46)$$

Het rechterlid van (5.46) kunnen we beschouwen als een generieke gedaante van veel unidimensionale modellen voor polytome items. We zien dat de grootheid a_i opgeslorpt is in de nieuwe categorieparameter δ_{ij}, doch dit is geen probleem want door een simpele deling krijgen we de oorspronkelijke η-parameters terug. De verschillende modellen onderscheiden zich vooral van elkaar door de structuur en de status van A_{ij}, het gewicht of de score die aan een antwoord in de *j*-de categorie op item *i* moet worden toegekend. Zo kunnen we zeggen dat de categorieresponsfuncties van het PCM gegeven zijn door het rechterlid van (5.46), met $A_{ij} = j$. In tabel 5.2 wordt een overzicht gegeven van alle unidimensionale modellen die in dit boek behandeld worden als speciale gevallen van de algemene gedaante (5.46). De enige uitzondering is het 3PL, dat niet in deze categorisering past.

Tabel 5.2
Unidimensionale modellen als speciaal geval van (5.46)

Model	A_{ij}	Opmerkingen
Raschmodel	0 en a	0 voor een fout antwoord; $a > 0$ voor een juist antwoord.
Dichotome OPLM	0 en a_i	0 voor een fout antwoord; a_i een positief geheel getal voor een juist antwoord; a_i a priori vastgelegd.
2PL	0 en a_i	0 voor een fout antwoord; $a_i > 0$, uit de data geschat.
PCM	j	$j = 0,...,m_i$
Polytome OPLM	ja_i	$j = 0,...,m_i$; a_i is een positief geheel getal, a priori vastgelegd.
Polytome 2PL	ja_i	$j = 0,...,m_i$; $a_i > 0$, uit de data geschat.

Wilson en Masters	ℓ_j	ℓ_j is een positief geheel getal a priori vastgelegd (alleen voor geobserveerde categorieën).
nominale responsmodel	a_{ij}	uit de data geschat.

De uitbreiding van het PCM die Wilson en Masters behandelen, kan ook als een speciaal geval (5.46) beschreven worden: zij kiezen voor A_{ij} van te voren, door de scoringsregel, vastgelegde gehele waarden. In het voorbeeld dat we hierboven gaven geven zij voor de vier geobserveerde categorieën respectievelijk de gewichten 0, 1, 3 en 4.

We hebben reeds eerder gezien dat het model dat door (5.46) gegeven is, niet identificeerbaar is. Als een item 5 antwoordcategorieën heeft, dan verschijnen in (5.46) ook 5 categorieparameters, η of δ, voor dat item, doch ze zijn niet allemaal schatbaar. We hebben dit probleem opgelost door in het middelste lid van (5.46) teller en noemer te vermenigvuldigen met $\exp(\eta_{i0})$ en het spreekt vanzelf dat we dezelfde techniek kunnen toepassen op het rechterlid van (5.46) door teller en noemer te vermenigvuldigen met $\exp(\delta_{i0})$. In het bovenstaande voorbeeld heeft item *i* dus vijf categorieparameters, waarbij in de toepassing van Wilson en Masters er slechts drie geschat worden. De parameter δ_{i2} wordt niet geschat omdat de tweede categorie niet geobserveerd is en de drie overige parameters die wel geschat worden zijn de verschillen $\delta_{i1}-\delta_{i0}, \delta_{i3}-\delta_{i0}$ en $\delta_{i4}-\delta_{i0}$. Het is belangrijk hierbij op te merken dat de δ-parameter die 'weggewerkt' wordt om het model identificeerbaar te maken, hier dus δ_{i0}, niet mag overeenkomen met een categorie die niet geobserveerd is. Indien categorie 0 in de steekproef niet geobserveerd is kan $\exp(\delta_{i0})$ als factor in teller en noemer in het rechterlid van (5.46) om het model te identificeren. Doch zoals we reeds eerder zagen kan een willekeurige andere parameter, waarvan de overeenkomende categorie wel is geobserveerd, gebruikt worden. Dit maakt de interpretatie van de parameters er echter niet gemakkelijker op.

Hoewel de benadering van Wilson en Masters elegant is om parameters van polytome items te schatten indien niet alle categorieën geobserveerd zijn, moet het praktische nut van hun methode niet overschat worden. Indien in de calibratiesteekproef een bepaalde categorie niet voorkomt, dan heeft men geen schatting van de bijbehorende categorieparameter. Doch dit sluit niet uit dat bij een latere toepassing die categorie wel wordt geobserveerd. Dan is het niet mogelijk uit een antwoordpatroon waar deze categorie in voorkomt θ te schatten, omdat voor een schatting van θ de ontbrekende waarde van de categorieparameter nodig is.

Het nominale responsmodel

Het rechterlid van (5.46) suggereert een verdere uitbreiding van het PCM. We kunnen namelijk het standpunt innemen dat we helemaal niets weten over de gewichten A_{ij} en ze behandelen als parameters die uit de data moeten geschat worden. Doch dit impliceert dat $A_{i,j+1}$ kleiner kan zijn dan $A_{i,j}$, dus dat een antwoord in categorie j hoger moet gewaardeerd worden dan een antwoord in categorie $j+1$. De ordening van de categorieën komt niet meer overeen met de ordening van hun gewichten. De categorienummers zijn dus gewoon labels van de categorie geworden en het resulterend model wordt dan ook het nominale responsmodel genoemd. Het werd voorgesteld door Bock (1972).

Het is niet moeilijk om uit het rechterlid van (5.46) af te leiden dat de voldoende steekproefgrootheid voor θ gegeven is door

$$\Sigma_i \Sigma_j A_{ij} y_{vij}. \tag{5.47}$$

Indien de gewichten A_{ij} a priori zijn vastgelegd zoals in het PCM, het polytome OPLM en het model van Wilson en Masters, is deze grootheid zonder meer uit de data te berekenen en kan er dus op geconditioneerd worden. In deze modellen is CML dus mogelijk. In het nominaal respons model moeten de gewichten A_{ij} geschat worden en kunnen dus niet gebruikt worden om te conditioneren. De MML-schattingsprocedure is wel mogelijk en is geïmple-menteerd in het computerprogramma MULTILOG (Thissen, 1988).

Het rating scale model

In paragraaf 5.1 hebben we gezien dat het LLTM een specificatie is van het Raschmodel die ontstaat door op de itemparameters lineaire restricties op te leggen. Dit is natuurlijk ook mogelijk bij polytome items; alleen dient men een zinvolle theorie of hypothese voor deze restricties te hebben of te construeren. We bespreken hier één voorbeeld van dergelijke restricties, het rating scale model van Andrich (1978a, 1978b).

Een rating scale is een observatie-instrument waarbij een persoon uit een aantal geordende categorieën er een uitkiest die het beste zijn mening weerspiegelt met betrekking tot een bepaalde uitspraak of een bepaald onderwerp. We geven twee voorbeelden van items die van deze techniek gebruik maken.

Item A: Den Uyl was een goede premier van Nederland.
sterk oneens oneens eens sterk eens

Item B: De colleges van prof. P. zijn interessant.
sterk oneens oneens eens sterk eens

Item A is bedoeld om de politieke attitude te meten van de persoon die het item beantwoordt en item B wordt gebruikt in een vragenlijst die bedoeld is om de attitude ten opzichte van een bepaalde onderwijsinstelling te meten. Hoewel het formaat van beide items identiek is en beide items bedoeld zijn om een attitude te meten, volgt daar niet uit dat het gedrag met betrekking tot beide items met eenzelfde soort model adequaat kan worden beschreven. Als we, net als in paragraaf 5.2, de politieke attitude interpreteren als de traditionele 'links-rechts' dimensie, ligt het voor de hand item A te interpreteren als een niet monotoon item. Personen met een ultra-linkse of ultra-rechtse overtuiging zullen het waarschijnlijk met de uitspraak in item A niet eens zijn, hoewel ze op de veronderstelde dimensie zeer ver van elkaar gelokaliseerd zijn. Voor dit item lijkt het dus redelijk een model voor niet-monotone items te gebruiken. Bij item B daarentegen lijkt het redelijk aan te nemen dat personen die het zelfde antwoord geven niet drastisch van elkaar verschillen in hun attitude. Bovendien lijkt het redelijk aan te nemen dat de categorie 'sterk eens' wijst op een positievere attitude dan de categorie 'eens' of 'oneens'. Kortom, de interpretatie van item B als een monotoon item is veel aannemelijker dan dit het geval is bij item A. Het rating scale model van Andrich is ontwikkeld als model voor items die geïnterpreteerd worden als monotone items.

Het is kenmerkend voor het gebruik van rating scales dat de antwoordcategorieën waaruit gekozen moet worden allemaal op dezelfde manier gelabeld zijn. In het model van Andrich is de kans dat een persoon *v* op item *i* met categorie *j* antwoordt, afhankelijk van de latente attitude θ_v van die persoon, van de 'moeilijkheid' van het item *i* en van de 'moeilijkheid' van antwoordcategorie *j*. Om een goed begrip te hebben van het onderscheid tussen beide moeilijkheden beschouwen we nog een ander item uit de schoolattitudevragenlijst:

Item C: Prof. P. is de ideale lesgever.
sterk oneens oneens eens sterk eens

Een persoon die het sterk eens is met de uitspraak in item B hoeft het niet sterk eens te zijn met de uitspraak in item C. Met andere woorden item C is 'moeilijker' dan item

B. We hadden natuurlijk ook een vragenlijst kunnen construeren waarin we dezelfde uitspraken gebruikten als in de items B en C, maar de antwoordcategorieën formuleerden als: 'nee' en 'ja'. Het zal wel duidelijk zijn dat er een positievere attitude vereist is om het antwoord 'sterk eens' te kiezen dan het veel minder sterk gekleurde antwoord 'ja'. De categorie 'ja' impliceert een lagere drempel dan de categorie 'sterk eens'.

Het rating scale model van Andrich is een speciaal geval van het PCM waar de categorieparameter β_{ij} uit formule (5.36) geschreven wordt als

$$\beta_{ij} = \gamma_i + \tau_j, \quad (i = 1,...,k\,;\, j = 1,...,m), \tag{5.48}$$

waarin m het gemeenschappelijke aantal antwoordcategorieën is, γ_i de itemparameter van item i en τ_j de parameter van antwoordcategorie j. De parameters γ en τ kunnen we dus opvatten als basisparameters; de categorie-parameters β_{ij} van het PCM zijn dus lineaire combinaties van de basisparameters.

Naast het rating scale model van Andrich bestaan er nog andere interessante modellen, die kunnen geschreven worden als restricties op de PCM-parameters β_{ij}, doch in die gevallen gaat het niet meer om lineaire restricties. Details over deze modellen kan men vinden in Masters en Wright (1984).

5.5 Multidimensionale IRT-modellen

Het begrip unidimensionaliteit dat tot hier toe is gehanteerd, is redelijk eenduidig; het begrip multidimensionaliteit heeft vele betekenissen. Vooraleer we specifieke modellen aan de orde stellen, geven we een overzicht van de verschillende betekenissen van het begrip.

Grosso modo kunnen we twee klassen van multidimensionale benaderingen binnen de IRT onderscheiden. De eerste klasse betreft modellen die een beperkt probleem oplossen. De verzameling items die moet worden geanalyseerd is reeds opgedeeld in een aantal groepen items en voor elk van die groepen weet of veronderstelt men dat ze geschaald kunnen worden met een unidimensionaal IRT-model, bijvoorbeeld met het Raschmodel. Bij de tweede klasse van modellen weet men dit niet, of wenst men die veronderstelling niet te maken. Modellen die tot die klasse behoren zijn bedoeld om de multidimensionale structuur van de items te ontrafelen. Deze vage noties worden nu explicieter gemaakt.

Veronderstel dat men de beschikking heeft over een aantal toetsen, zeg Q, die elk adequaat beschreven kunnen worden door een unidimensionaal IRT-model. Elk van deze toetsen is dus geschikt om een latente eigenschap $\theta_q, (q = 1,..., Q)$, te meten. De vraag die men zich kan stellen is of deze Q eigenschappen iets met elkaar te maken hebben, hoe groot bijvoorbeeld de correlatie tussen die eigenschappen is in een bepaalde populatie. Een voorbeeld van deze benadering wordt besproken in paragraaf 5.5.1.

In de tweede klasse van modellen wordt er van uitgegaan dat elk item een beroep doet op twee of meer latente vaardigheden. Deze modellen zijn bedoeld om na te gaan in welke mate elk item uit een toets een beroep doet op elke vaardigheid. Een mogelijke situatie is dat een gedeelte van de items uitsluitend een beroep doet op één vaardigheid en de overige items uitsluitend een andere vaardigheid aanspreken. Het zou echter ook kunnen zijn dat alle items op alle vaardigheden in verschillende aanspreken. Het is echter niet zonder meer duidelijk wat bedoeld wordt met uitdrukkingen als: ’een beroep doen op’ of ’aanspreken’. Deze begrippen dekken een heel complexe lading, die we met enkele voorbeelden zullen toelichten.

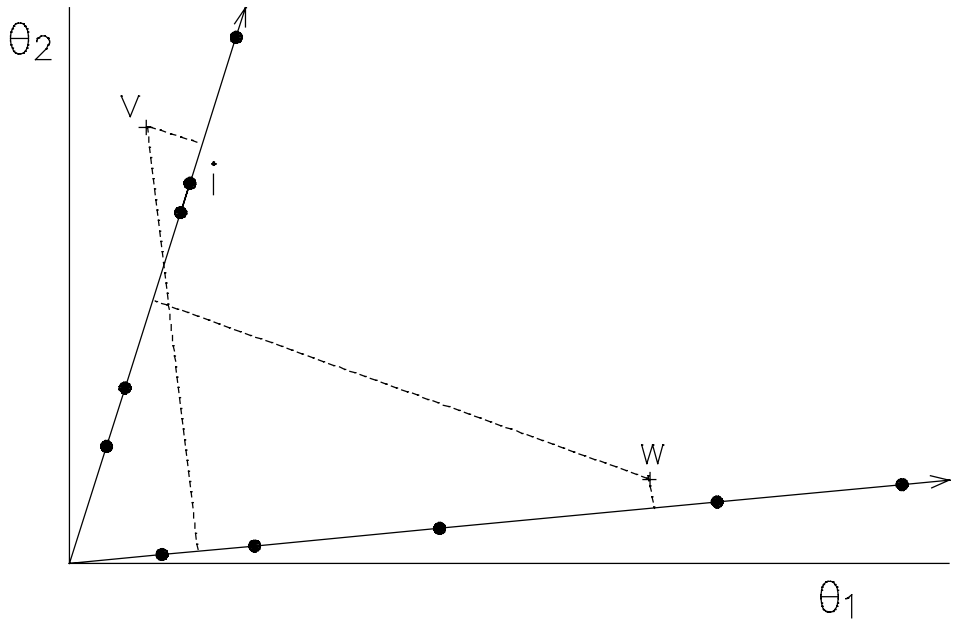
In de psychologie wordt soms gebruikt gemaakt van de Rorschachtest. Daarbij moet de persoon bij tien plaatjes waarop een ongestructureerde inktvlek staat aangeven wat hij of zij in die inktvlek ziet. De antwoorden worden op grond van een theorie uit de persoonlijkheidsleer gecategoriseerd in een aantal categorieën, waarbij ervan wordt uitgegaan dat elke categorie wijst op een bepaalde persoonseigenschap. De kans dat een persoon bij een plaatje een antwoord geeft in een bepaalde categorie zal dus afhangen van de mate waarin deze persoon over de overeenkomstige eigenschap beschikt en van de mate waarin het plaatje een bepaalde categorie van antwoorden uitlokt. Als we de plaatjes beschouwen als items, kunnen we dus stellen dat elk item verschillende latente eigenschappen aanspreekt. Een IRT-model dat het gedrag bij de Rorschachtest adequaat beschrijft, zal dus een multidimensionaal model zijn. In paragraaf 5.5.2 wordt zo’n model besproken.

Een heel andere betekenis van het begrip multidimensionaliteit kan geïllustreerd worden met het volgende voorbeeld. In veel schoolse situaties worden belangrijke beslissingen genomen aan de hand van een enkel rapportcijfer, dat meestal een gewogen gemiddelde is van verschillende proefwerkcijfers. Deze praktijk weerspiegelt de assumptie dat het algemene cijfer, een unidimensionale grootheid, een adequate beslissingsgrond biedt, hoewel niemand zal beweren dat twee leerlingen met hetzelfde cijfer op alle vakken even goed of even slecht zijn. Een slecht cijfer voor wiskunde kan gecompenseerd worden door een goed cijfer voor taal en omgekeerd. Een soortgelijke gedachte kan men van toepassing achten op itemniveau. Als een item een beroep doet

op twee vaardigheden kan een bepaalde kans op een juist antwoord van bijvoorbeeld 0.5 tot stand komen omdat men in beide vaardigheden middelmatig is, maar ook omdat men in de ene vaardigheid erg laag scoort, maar dit tekort kan compenseren omdat men excelleert in de andere vaardigheid. Modellen die dit soort mechanisme veronderstellen worden soms aangeduid als compensatorische modellen. De structuur van deze modellen komt in paragraaf 5.5.3 aan de orde.

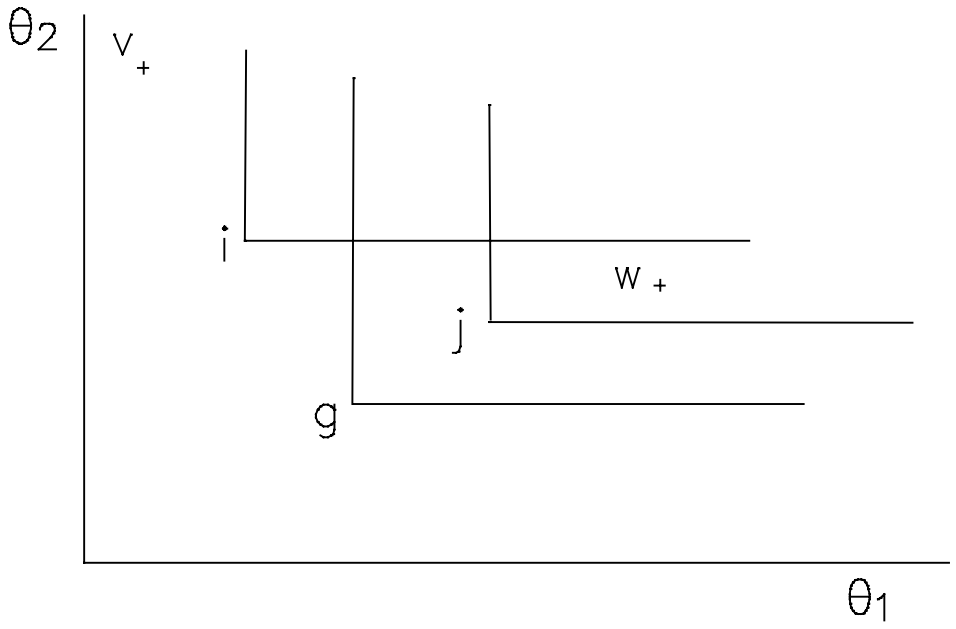
Het voorbeeld van de schoolcijfers is niet helemaal realistisch. De meeste schoolreglementen staan niet toe dat een 1 voor wiskunde gecompenseerd kan worden door een 10 voor taal. Men bouwt dus een mechanisme in de beslissingsregel in, dat bepaalt dat zowel op wiskunde als op taal een bepaald minimum cijfer behaald dient te worden. Dit soort regels kan men ook van toepassing achten op itemniveau. Of een persoon een item juist kan beantwoorden, hangt dan af of een bepaald niveau bereikt is op alle vaardigheden waarop dit item een beroep doet. Modellen die een dergelijk mechanisme veronderstellen worden conjunctieve modellen genoemd. In paragraaf 5.5.4 gaan we op deze modellen in.

De figuren 5.8 en 5.9 zijn een grafisch hulpmiddel om het onderscheid tussen compensatorische en conjunctieve modellen te verduidelijken. Figuur 5.8 is een voorstelling van een compensatorisch model waarbij alle items in de figuur voorgesteld met stippen een beroep doen op de vaardigheden θ_1 en θ_2.



Figuur 5.8
Een compensatorisch model

Vijf items liggen op een lijn die bijna verticaal staat, waarmee wordt aangegeven dat deze vijf items op dezelfde manier een beroep doen op de twee vaardigheden; ze doen echter meer een beroep op θ_2 dan op θ_1, want de hoek die de lijn vormt met de verticale as is kleiner dan de hoek met de horizontale as.



Figuur 5.9
Een conjunctief model

Deze vijf items samen meten dus een unidimensionale vaardigheid, die een bepaald mengsel is van de beide vaardigheden θ_1 en θ_2. De pijl die bij de lijn getekend is geeft de richting van de toenemende vaardigheid aan. Mutatis mutandis geldt dit ook voor de andere vijf items. De tien items samen meten echter niet een unidimensionale vaardigheid, omdat het mengsel van vaardigheden waarop ze een beroep doen niet voor alle items hetzelfde is. De positie van de *letter v* in de figuur geeft aan dat persoon *v* over een hoge mate van vaardigheid θ_2 beschikt, maar over een lage mate van vaardigheid θ_1. We verwachten dus dat die persoon het goed zal doen op items die vooral een beroep doen op θ_2 en minder goed op items die vooral θ_1 aanspreken. Het omgekeerde geldt voor persoon *w*. Om te weten of persoon *v* het goed zal doen bij de beantwoording van item *i*, nemen we de projectie van het punt dat zijn vaardigheid voorstelt op de lijn die de schaal voorstelt waarop het item ligt. We kunnen dit op een analoge manier doen voor de tweede schaal, en ook voor persoon *w*. Deze projecties zijn aangegeven als de eindpunten van de stippellijnen. Met een deterministische interpretatie zouden we kunnen zeggen dat persoon *v* over meer van de gecombineerde vaardigheid beschikt dan item *i* vereist, en dat deze persoon item *i* dus correct zal

beantwoorden. Met deze interpretatie is gemakkelijk uit de figuur af te leiden dat de personen *v* en *w* elk vijf van de tien items juist zullen beantwoorden. Hun scores zijn dus gelijk, hoewel hun begaafdheden drastisch verschillen. Ze hebben beide op een verschillende manier hun tekort op de ene vaardigheid gecompenseerd door een grote mate van de andere vaardigheid.

In figuur 5.9 is een voorstelling van een conjunctief model gegeven. De positie van de items valt samen met het snijpunt van een horizontaal en een verticaal lijnstuk. In een deterministische interpretatie stelt de hoogte van het horizontale lijnstuk de minimale hoeveelheid vaardigheid θ_2 voor die nodig is om het item correct te beantwoorden. Het verticale lijnstuk geeft de minimale hoeveelheid van vaardigheid θ_1 aan. Men kan een item alleen dan juist beantwoorden als men zich rechts boven het punt bevindt dat het item voorstelt. Persoon *v* zal dus geen enkel item juist beantwoorden, en persoon *w* zal een juist antwoord geven op de items *j* en *g*. Hoewel persoon *v* duidelijk over meer vaardigheid θ_2 beschikt dan persoon *w*, helpt dat niet om het tekort aan vaardigheid θ_1 te compenseren.

5.5.1 Een OPLM met een multivariate vaardigheidsverdeling

Indien een unidimensionaal OPLM geen goede passing oplevert, kan men op zoek gaan naar een opdeling van de items in deelverzamelingen die wel goed te beschrijven zijn met een unidimensionaal model. Het zoeken naar zo'n opdeling is geen triviaal probleem en het kan op verschillende manieren gebeuren. Men kan bijvoorbeeld gebruik maken van de toets voor unidimensionaliteit die door Martin-Löf ontwikkeld is (zie hoofdstuk 4), of een factoranalyse uitvoeren op de matrix van interitemcorrelaties (Bol & Verhelst, 1985). Wij gaan niet op dit probleem in. Indien men zo'n opdeling heeft, rijst de vraag hoe de vaardigheden die door de verschillende deeltoetsen worden gemeten met elkaar in verband staan. Een elegante manier om dit probleem aan te pakken, is een multivariate normale verdeling te veronderstellen voor de vaardigheid $\theta = (\theta_1,...,\theta_q,...,\theta_Q)$. Een multivariaat normale verdeling is net als de gewone normale verdeling, eigenlijk een familie van verdelingen, en een lid van deze familie wordt gespecificeerd door de waarden van de parameters vast te leggen. Deze parameters zijn de vector van gemiddelden $\boldsymbol{\mu} = (\mu_1,...,\mu_Q)$ en de covariantiematrix Σ, waarin niet alleen de variantie van elk van de afzonderlijke θ-variabelen wordt gespecificeerd maar ook hun covarianties. Bij een *Q*-variate normale verdeling zijn er dus $Q + Q(Q+1)/2$ parameters. Indien de oorspronkelijke *k* items zijn opgedeeld in *Q* deelverzamelingen, kan men het nulpunt van de *Q* schalen vrij kiezen, door

bijvoorbeeld alle gemiddelden gelijk te stellen aan 0. In totaal moeten er dus $k + Q(Q+1)/2$ parameters geschat worden.

Als we het antwoordpatroon op de q-de deeltoets aanduiden als $\boldsymbol{x}^{(q)}$, en het antwoord-patroon voor alle k items als $\boldsymbol{x} = (\boldsymbol{x}^{(1)},..,\boldsymbol{x}^{(Q)})$, kunnen we de aannemelijkheidsfunctie schrijven als

$$\begin{aligned} L(\beta,\Sigma;\boldsymbol{x}) &= \int\!\!\ldots\!\int P(\boldsymbol{x}\,|\,\theta;\beta)\,g(\theta;\Sigma)\,d\theta \\ &= \int\!\!\ldots\!\int \prod_{q=1}^{Q} P(\boldsymbol{x}^{(q)}\,|\,\theta_q;\beta^{(q)})\,g(\theta;\Sigma)\,d\theta_1\ldots d\theta_Q\,, \end{aligned} \tag{5.49}$$

waarin $\beta^{(q)}$ de vector met itemparameters is voor de items in de q-de subtoets. De aannemelijkheidsfunctie gegeven de observaties van verschillende personen is dan gewoon het produkt van uitdrukkingen als het rechterlid van (5.49). Merk op dat (5.49) de multivariate versie is van de aannemelijkheidsfunctie die we in hoofdstuk 4 opgesteld hebben bij de bespreking van de MML-schattingsmethode. In deze context is dit heel natuurlijk, want de toevoeging van een veronderstelling over de verdeling van de vaardigheid in de populatie is een essentieel onderdeel van het model. Glas (1989, 1992) bespreekt de details van de schattingsprocedure en geeft ook aan hoe het model statistisch kan worden getoetst.

Een eenvoudiger versie van dit model werd eerder voorgesteld door Andersen (1985). Bij de toepassing die Andersen bespreekt, levert de opdeling van de items in subtoetsen geen enkel probleem op. Indien dezelfde toets op twee verschillende tijdstippen aan dezelfde personen wordt afgenomen, kan men proberen te achterhalen of en hoe de vaardigheid in de tussentijd is veranderd. Door te veronderstellen dat de verdelingen van θ op de twee tijdstippen gezamenlijk een bivariaat normale verdeling vormen, krijgt men direct een speciaal geval van het model dat hierboven werd besproken met Q=2. Andersen veronderstelde bovendien dat de itemparameters bekend zijn, bijvoorbeeld uit een voorafgaande calibratie. De waarden van de itemparameters op de twee tijdstippen zijn dus exact gelijk. Daarmee liggen de nulpunten van de twee schalen vast, en moeten de gemiddelden μ_1 en μ_2 geschat worden, evenals de twee varianties en de covariantie. Het verschil $\mu_2 - \mu_1$ geeft de gemiddelde toename in vaardigheid, maar het model laat toe dat de twee varianties verschillend kunnen zijn, en dat de correlatie tussen θ_1 en θ_2 ongelijk is aan 1. Men zou kunnen opmerken dat er nooit een correlatie van 1 gevonden wordt tussen twee metingen. Dit is zo, als het gaat over correlaties tussen geobserveerde variabelen die

altijd een zekere mate van onbetrouwbaarheid bevatten waardoor de correlatie niet 1 kan zijn. Hier gaat het echter om de correlatie tussen latente variabelen, die per definitie geen meetfout bevatten. De hoogte van de correlatie geeft een aanduiding van de stabiliteit in de tijd van de latente vaardigheid.

5.5.2 Het multidimensionale model van Rasch

Rasch heeft niet alleen het zeer bekende Raschmodel voor dichotome items ontwikkeld. Hij heeft ook aandacht besteed aan polytome items. In zijn bekommernis om modellen te ontwikkelen waarbij de eigenschappen van items, de itemparameters, bepaald kunnen worden onafhankelijk van wie de items heeft beantwoord, en omgekeerd, waar de eigenschappen van personen gemeten kunnen worden, onafhankelijk van welke items men daar voor gebruikt, kwam Rasch (1961) tot een merkwaardig resultaat: indien de antwoorden op de items in *m* verschillende categorieën kunnen worden ondergebracht, dan hebben we een *m*-dimensionaal model nodig, waarbij de categorieresponscurven gegeven zijn door:

$$P(X_i = j \mid \xi_v) = \frac{\exp(\xi_v^{(j)} - \eta_{ij})}{\sum_{h=1}^{m} \exp(\xi_v^{(h)} - \eta_{ih})}, \quad (j = 1,...,m) \tag{5.50}$$

waarin $\xi_v = (\xi_v^{(1)},...,\xi_v^{(m)})$ en $\xi_v^{(j)}$ geïnterpreteerd kan worden als de mate waarin persoon *v* de neiging heeft om een antwoord in categorie j te geven. Denk hierbij aan de toepassing over de Rorschachtest die we eerder bespraken. De parameter η_{ij} kan dan geïnterpreteerd worden als de mate waarin item *i* een antwoord in categorie *j* uitlokt.

Het model dat in (5.50) is gegeven is echter niet geïdentificeerd, omdat er twee soorten transformaties zijn die we op het rechterlid van (5.50) kunnen uitvoeren, zonder dat het linkerlid verandert. Vermenigvuldigen we teller en noemer van (5.50) met $\exp(\eta_{i1} - \xi_v^{(1)})$ en definiëren we

$$\theta_v^{(j)} = \xi_v^{(j)} - \xi_v^{(1)}, \quad (j = 1,...,m), \tag{5.51}$$

$$\beta_{ij} = \eta_{ij} - \eta_{i1}, \quad (j = 1,...,m\,;\, i = 1,...,k), \tag{5.52}$$

dan kan (5.50) herschreven worden als

$$P(X_i = j \mid \theta_v) = \frac{\exp(\theta_v^{(j)} - \beta_{ij})}{1 + \sum_{h=2}^{m} \exp(\theta_v^{(h)} - \beta_{ih})}, \quad (j = 2,...,m) \tag{5.53}$$

en voor het geval $j = 1$ als

$$P(X_i = j \mid \theta_v) = \frac{1}{1 + \sum_{h=2}^{m} \exp(\theta_v^{(h)} - \beta_{ih})}, \quad (j = 2,...,m). \tag{5.54}$$

De 1 in de formules (5.53) en (5.54) verschijnt dus als gevolg van de transformaties (5.51) en (5.52), waaruit direct volgt dat $\theta_v^{(1)} = \beta_{i1} = 0$ voor alle personen v en alle items i. Dit betekent dat de neiging om in een bepaalde categorie te antwoorden niet in absolute zin kan worden bepaald. De parameter $\theta_v^{(j)}$ moet dus geïnterpreteerd worden als de sterkte van de neiging om met categorie j te antwoorden vergeleken met de neiging om met categorie 1 te antwoorden. Categorie 1 heet de referentiecategorie. Het blijkt dus dat er maar m-1 onafhankelijke dimensies zijn. Stellen we m gelijk aan 2, dan resulteert een unidimensionaal geval, en het is gemakkelijk na te gaan dat in dat geval de formules (5.53) en (5.54) equivalent zijn met de formules voor het unidimensionale Raschmodel dat in hoofdstuk 4 werd behandeld. Merk op dat in dit geval het foute antwoord fungeert als referentiecategorie.

De tweede onbepaaldheid kennen we reeds uit het unidimensionale geval. Indien bij $\theta_v^{(j)}$ en β_{ij} een constante c_j opgeteld wordt, verandert hun verschil niet. Dit betekent dat we het nulpunt op elk van de m-1 vrije dimensies vrij kunnen kiezen, bijvoorbeeld door β_{1j} gelijk te stellen aan 0. Het totale aantal vrije parameters in het model is dus gelijk aan (k-1)(m-1). Hoewel meestal erg makkelijk gedaan wordt over normalisaties, moet men hier toch goed uitkijken, omdat niet alle vergelijkingen van parameters zinvol zijn. De vraag of persoon v meer geneigd is om met categorie j te antwoorden dan persoon w, kan men zinvol beantwoorden door het verschil

$$\theta_v^{(j)} - \theta_w^{(j)} = \xi_v^{(j)} - \xi_w^{(j)}$$

te beschouwen. De vraag of persoon *v* meer geneigd is om met categorie *j* te antwoorden dan met categorie *g*, is niet zinvol te beantwoorden, omdat het verschil

$$\theta_v^{(j)} - \theta_v^{(g)}, \quad (j \neq g),$$

volstrekt willekeurig is: de normalisaties van beide dimensies kunnen vrij gekozen worden. Soortgelijke argumenten gelden natuurlijk ook bij het vergelijken van categorieparameters.

Hoewel dit model heel wat eigenschappen heeft die theoretisch zeer aantrekkelijk zijn, waaronder de mogelijkheid om de categorieparameters te schatten met CML, is het bedenken van interessante toepassingsmogelijkheden niet zo eenvoudig. Bovendien is het afleiden van de schattingsvergelijkingen heel wat complexer dan bij het dichotome Raschmodel. De geïnteresseerde lezer kan een gedetailleerde bespreking van de CML-schattingsprocedure vinden in Fischer (1974), waar ook het voorbeeld van de Rorschachtest wordt besproken. Een afleiding van het model vanuit de eis van het bestaan van voldoende steekproefgrootheden voor de persoonsparameters kan men vinden in Andersen (1973c).

5.5.3 Compensatorische IRT-modellen

Uit figuur 5.8 is het vrij gemakkelijk te begrijpen hoe de meeste compensatorische modellen in elkaar zitten. Om de uiteenzetting niet nodeloos ingewikkeld te maken, zullen we de bespreking beperken tot het geval van dichotome items. De gerichte lijn waarop in figuur 5.8 item *i* is afgebeeld kunnen we beschouwen als de reële-getallenas. Het punt dat item *i* voorstelt kan dus geïnterpreteerd worden als een getal, dat we β_i zullen noemen. De richting van de lijn is volledig bepaald door de hoeken die de lijn maakt met de twee assen van het assenstelsel, en dus ook door de cosinussen van die hoeken. We duiden die twee cosinussen aan met respectievelijk a_{i1} en a_{i2}. Het punt in de tweedimensionale ruimte dat de vaardigheid van persoon *v* aanduidt kunnen we nauwkeurig beschrijven met de twee coördinaten van dat punt, θ_{v1} en θ_{v2}. De projectie van dit punt op de lijn waarop item *i* ligt is gegeven door

$$a_{i1}\theta_{v1} + a_{i2}\theta_{v2}$$

en dit getal is groter dan β_i. In de deterministische interpretatie die we eerder gaven, leidde dit positieve verschil tot een juist antwoord. In een kansmodel zullen we zeggen

dat hoe groter dit verschil is, des te groter de kans is op een juist antwoord. Als we gebruik maken van een logistische responsfunctie krijgen we dus automatisch als model:

$$P(X_i = 1 \mid \theta_{v1}, \theta_{v2}) = \frac{\exp(a_{i1}\theta_{v1} + a_{i2}\theta_{v2} - \beta_i)}{1 + \exp(a_{i1}\theta_{v1} + a_{i2}\theta_{v2} - \beta_i)}. \tag{5.55}$$

De generalisatie tot Q dimensies is dan voor de hand liggend:

$$P(X_i = 1 \mid \theta_{v1}, ..., \theta_{vQ}) = \frac{\exp\left(\sum_{q=1}^{Q} a_{iq}\theta_{vq} - \beta_i\right)}{1 + \exp\left(\sum_{q=1}^{Q} a_{iq}\theta_{vq} - \beta_i\right)}. \tag{5.56}$$

Er is echter een eigenschap van het besproken model die nog niet aan de orde is geweest, namelijk dat de som van de kwadraten van de cosinussen a_{i1} en a_{i2} gelijk is aan 1. Deze regel geldt ook indien er meer dan twee dimensies zijn. Dus:

$$\sum_{q=1}^{Q} a_{iq}^2 = 1, \qquad (i = 1, ..., k). \tag{5.57}$$

Uit figuur 5.8 is duidelijk dat, indien we dit model toepassen op de items die allemaal op dezelfde lijn liggen als item *i*, het unidimensionale Raschmodel moet gelden. Dus kan het model dat gedefinieerd is door (5.56) samen met de restrictie (5.57) beschouwd worden als een multidimensionaal compensatorisch Raschmodel. Dit model is in de literatuur echter nog nooit beschreven en bestudeerd. De variant die wel beschreven is, is gegeven door (5.56) waarbij de restrictie (5.57) niet wordt opgelegd (McKinley & Reckase, 1982). De geometrische interpretatie van dit model is iets gecompliceerder dan aangegeven in figuur 4.8, en we gaan er hier niet verder op in; er wordt een interpretatie gegeven in Bol en Verhelst (1985). Als de restrictie (5.57) niet wordt opgelegd, ontstaat een compensatorische generali-satie van het 2PL. Dit is gemakkelijk te zien door in (5.57) Q gelijk te stellen aan 1.

Omdat de gewichten a_{iq} in (5.57) niet bekend zijn, zijn er geen voldoende steekproef-grootheden voor de persoonsparameters, en is CML dus onmogelijk. De schatting van

de parameters gebeurt dan ook meestal met MML, waarbij de veronderstelling gemaakt wordt dat θ Q-variaat normaal verdeeld is. Het computerprogramma MAXLOG (McKinley & Reckase, 1983) kan gebruikt worden om de parameters van dit model te schatten.

Lezers die enigszins bekend zijn met factoranalyse, zullen in figuur 5.8 en in de wijze waarop het model is opgebouwd zeker overeenkomsten gezien hebben met de factoranalyse. Als in plaats van de logistische functie, de (cumulatieve) normale verdelingsfunctie als responsfunctie wordt gebruikt en tevens de multivariaat normale verdeling van de vaardigheden, kan aangetoond worden dat het model een uitbreiding is van een factoranalytisch model dat vaak gehanteerd wordt, namelijk het model waarbij de factoren multivariaat normaal verdeeld zijn. Het is een uitbreiding omdat in de factoranalyse alleen de parameters a_{iq} geschat worden, die daar de naam factorlading krijgen, en niet de β-parameters. Bovendien is er een interessant contrast in de manier van parameterschattingen: binnen de traditie van de factoranalyse gebruikt men de correlatiematrix om de parameters te schatten. Indien de variabelen dichotoom zijn, kan deze methode echter tot problemen leiden (zie hoofdstuk 15 van Lord & Novick, 1968). Men kan echter ook de parameters van het model schatten door de aannemelijkheidsfunctie van de geobserveerde antwoordpatronen te maximaliseren, waarbij men meer informatie gebruikt dan aanwezig is in de interitemcorrelatiematrix. De variant van (5.56), waar de normale verdelingsfunctie is gebruikt in plaats van de logistische functie wordt dan ook, met een impliciete referentie naar de schattingsmethode, aangeduid als 'full information factor analysis' (Bock, Gibbons & Muraki, 1988). Het programma TESTFACT (Wilson, Wood & Gibbons, 1991) kan gebruikt worden om de parameters te schatten. Een algemeen overzicht van compensatorische IRT-modellen kan men vinden in Knol (1986).

Tot slot van deze paragraaf komen we nog even terug op een opmerking die in hoofdstuk 4 werd gemaakt, waarin werd betoogd dat het goed mogelijk is dat een unidimensionaal Raschmodel meerdere vaardigheden aanspreekt. Stel dat in figuur 5.8 θ_1 verbale vaardigheid voorstelt, en θ_2 numerieke vaardigheid. Uit de figuur is duidelijk dat alle items beide vaardigheden aanspreken. Als we in een model al deze items betrekken, hebben we inderdaad twee dimensies nodig. Beperken we het model echter tot de items die op dezelfde lijn liggen als item *i*, dan zijn die twee vaardigheden nog wel vereist om deze items te beantwoorden, maar een analyse van de antwoorden zal aanduiden dat we genoeg hebben aan 1 dimensie. Met andere woorden, het 'mengsel' van beide vaardigheden is voor alle items hetzelfde, en we zijn niet meer in staat beide vaardigheden van elkaar te onderscheiden.

5.5.4 Conjunctieve IRT-modellen

Het idee van het stellen van minimumeisen voor verschillende aspecten van een taak is reeds oud (Johnson, 1935), maar in de toegepaste psychometrie zijn de middelen schaars om dit algemene idee op een rationele manier toe te passen. Coombs (1964) heeft er uitvoerig aandacht aan besteed, doch het is pas recent dat er formele modellen zijn ontwikkeld die in de praktijk goed bruikbaar zijn. We bespreken hier kort een model dat door Maris (1992) is ontwikkeld. De deterministische interpretatie van Maris' model is als volgt. Indien aan twee minimumeisen moet worden voldaan, kunnen we ons voorstellen dat er impliciet twee vragen worden gesteld, en het antwoord op het item als geheel is alleen juist indien het antwoord op beide impliciete vragen juist is. Deze impliciete vragen worden natuurlijk niet echt gesteld, en de antwoorden erop zijn dan ook niet observeerbaar. Daarom worden ze latente antwoorden genoemd. Als er Q dimensies zijn, zijn er dus Q latente antwoorden die we zullen aanduiden als $Y_{i1},...,Y_{iQ}$, en die alle de waarden 1 of 0 kunnen aannemen. Het geobserveerde antwoord X_i is alleen gelijk aan 1 indien alle latente antwoorden juist zijn. Het deterministische model kan dus geschreven worden als

$$X_i = \prod_{q=1}^{Q} Y_{iq}. \tag{5.58}$$

Een analyse in het deterministische model komt er dus op neer de items op de Q dimensies zo te ordenen dat alle geobserveerde antwoordpatronen overeenkomen met een gebied in de multidimensionale ruimte dat, onder een conjunctieve interpretatie, met die antwoordpatronen overeenkomt. Zo is er in figuur 5.9 geen plaats voor een antwoordpatroon waarbij alleen item j juist werd beantwoord. Een deterministische oplossing vinden is meestal niet zo eenvoudig, en de reden is, dat het lastig is om te bepalen wat de waarde van Q moet zijn om alle geobserveerde antwoorpatronen hun plaats in de multidimensionale ruimte te geven (Koppen, 1987).

Bij een kansmodel loopt dit iets soepeler omdat in theorie elk antwoordpatroon onder elk model kan voorkomen. Maris construeerde zijn model door aan te nemen dat de latente antwoorden van eenzelfde persoon stochastisch onafhankelijk zijn van elkaar, waardoor we onmiddellijk de probabilistische versie van (5.58) kunnen opschrijven:

$$P(X_i = 1 \mid \theta_1,...,\theta_Q.) = \prod_{q=1}^{Q} P(Y_{iQ} = 1 \mid \theta_q). \tag{5.59}$$

Het model wordt dan gecompleteerd door voor elk latent antwoord het Raschmodel aan te nemen, zodat het model geschreven kan worden als

$$P(X_i = 1 \mid \theta_1,...,\theta_Q) = \prod_{q=1}^{Q} \frac{\exp(\theta_q - \beta_{iq})}{1 + \exp(\theta_q - \beta_{iq})}. \qquad (5.60)$$

Het model is dus een multidimensionaal conjunctief Raschmodel, en we zien dat het unidimensionale Raschmodel resulteert indien $Q = 1$.

De problemen met de parameterschatting en de statistische toetsing van het model zijn zeker niet allemaal opgelost. Zo past Maris alleen de JML-schattingsmethode toe die waarschijnlijk geen consistente schattingen oplevert. Hij beschrijft wel de MML-methode, maar de toepassing ervan brengt vele numerieke problemen met zich mee. Een variant van Maris' model, waarbij wel de MML-methode is gebruikt, kan men vinden in Van Leeuwe (1990).

5.6 Nabeschouwing

De grote weelde aan IRT-modellen die in dit hoofdstuk aan bod is gekomen, zal bij de lezer misschien de indruk wekken van wildgroei, zeker als men beseft dat er maar een kleine selectie van de bestaande modellen de revue is gepasseerd. Zie bijvoorbeeld de grote witte oppervlakte rechts en beneden in figuur 5.3. Het grote bos dat men door de vele bomen uit het oog dreigt te verliezen, is bovendien overwoekerd door veel stekelig struikgewas, zoals problemen van statistische, numerieke en algoritmische aard. Het feit dat er een groot aanbod is aan computerprogramma's biedt natuurlijk comfort, doch het zou een misvatting zijn te denken dat de psychometrie bestaat uit een aantal ingewikkelde rekensommen die nu dank zij het beschikbaar zijn van snelle rekenapparatuur gemakkelijk kunnen worden uitgevoerd. De strategie 'ik probeer ze allemaal en ik zie wel welk model het beste past' is een heilloze weg die de verwarring alleen maar groter kan maken. Het toepassen van een psychometrisch model is het toetsen van een hypothese aan de werkelijkheid en deze hypothese dient inhoudelijk zinvol te zijn. Ze probeert de verbanden tussen verschillende gedragingen te formuleren en zo zuinig en accuraat mogelijk te beschrijven. Zie bijvoorbeeld Roskam (1982). De keuze tussen, bijvoorbeeld, een compensatorisch en een conjunctief model moet men niet aan een computerprogramma overlaten, maar baseren op een analyse van het

gedragsdomein dat men wenst te analyseren. De wetenschap dat er goed uitgewerkte psychometrische formaliseringen en bijbehorende computerprogramma's bestaan, wordt dan een bron van welbevinden in plaats van verwarring.

6

Itemresponstheorie en onvolledige gegevens

In onderzoek in de gedragswetenschappen komt het veelvuldig voor dat men niet alle gegevens bij alle personen die aan een onderzoek meedoen heeft kunnen of willen verzamelen. Onderzoek waarbij de itemresponstheorie (IRT) wordt toegepast, vormt hierop geen uitzondering. Het ontbreken van gegevens of data kunnen we ons in deze situatie als volgt voorstellen. Als we de antwoorden van personen op items of vragen weergeven in een datamatrix en als we aannemen dat in totaal n personen en k items in het onderzoek betrokken zijn, dan zal een aantal van de in totaal $n \times k$ cellen van deze matrix leeg zijn. De lege cellen vertegenwoordigen de ontbrekende gegevens of 'missing data' in het onderzoek. De redenen voor het ontbreken van gegevens kunnen van onderzoek tot onderzoek sterk variëren maar zijn globaal in te delen in drie categorieën. Het criterium voor deze indeling is de mate waarin de onderzoeker zelf het optreden van de ontbrekende gegevens onder controle heeft. De eerste categorie die we onderscheiden is dat de onderzoeker van te voren vastlegt aan welke (groep) respondenten welke items worden voorgelegd en van te voren dus ook weet waar de lege cellen in de matrix zullen zitten. Een voorbeeld hiervan is dat bij een enquête de getrokken steekproef van respondenten vanwege de lengte van de vragenlijst beurtelings het eerste deel, met algemene vragen, en het tweede deel van een vragenlijst wordt voorgelegd, dan wel het eerste en het derde en laatste deel van de lijst. De tweede categorie is dat de onderzoeker vastgelegd heeft volgens welke procedure lege cellen in de datamatrix kunnen ontstaan, maar van te voren niet exact kan voorspellen waar de cellen precies leeg zullen zijn. In het hetzelfde voorbeeld van een enquête zou dit het geval zijn als we niet beurtelings, maar op grond van de uitkomst van een worp met een munt of bijvoor-beeld op grond van de leeftijd van de respondent zouden bepalen wie welk deel van de vragenlijst gaat beantwoorden. De derde en laatste categorie van het optreden van ont-brekende gegevens is dat zonder dat de onderzoeker daar enige invloed op heeft gegevens ontbreken. Bij een enquête is dit bijvoorbeeld het geval als een respondent weigert op een bepaalde vraag antwoord te geven.

De eerste twee categorieën van ontbrekende gegevens noemt men wel structureel onvolledig, de laatste categorie ontstaat spontaan tijdens het waarnemen en zijn vanuit het gezichtspunt van de onderzoeker doorgaans ongewenst en storend. Bij de laatste categorie kan de analyse van de gegevens vaak alleen maar goed plaatsvinden als we aannames doen omtrent de mechanismen die de ontbrekende gegevens veroorzaken. Meestal zijn deze aannames niet of heel moeilijk toetsbaar. Met structureel onvolledige gegevens kennen we deze mechanismen en kunnen we in de analyse doorgaans veel beter uit de voeten. In dit hoofdstuk zullen wij ons bezighouden met structureel onvolledige designs. In de itemresponstheorie wordt namelijk met modellen gewerkt die onder bepaalde voorwaarden erg goed structureel onvolledige gegevens kunnen analyseren. Ook de niet structureel ontbrekende gegevens komen in de psychometrische praktijk voor. Denk hierbij aan ontbrekende gegevens die ontstaan doordat leerlingen opgaven in een toets overslaan of ook wel de situatie waarin de toets een zodanige lengte heeft dat sommige leerlingen bepaalde opgaven niet bereiken. We zullen deze onder-werpen niet bespreken. Voor voorbeelden van modellen die rekening houden met een tijdslimiet op de toetsafname verwijzen we naar Verhelst, Verstralen en Jansen (1993).

In het hiernavolgende zullen we eerst de relatie tussen IRT en onvolledige gegevens in het algemeen bespreken. Daarna wordt een overzicht gegeven van de in de praktijk veel voor-komende designs. In paragraaf 6.2 doen we dit door middel van het beschrijven van de datamatrices in onvolledige designs. In paragraaf 6.3 gebeurt dit aan de hand van het stochastische mechanisme dat de onvolledige gegevens veroorzaakt. Wij bespreken daarbij de drie in de praktijk meest gebruikte stochastische designtypen. Als we IRT toepassen beginnen we met het calibratie-onderzoek, het schatten van de itemparameters. Daarom zullen we hierna uitvoerig ingaan op de mogelijkheden en voorwaarden voor calibratie in onvolledige designs. Beide schattingsmethoden uit hoofdstuk 4, met behulp van de marginale aannemelijkheidsfunctie (MML) en met behulp van de conditionele aannemelijkheidsfunctie (CML) worden behandeld. In paragraaf 6.4 bespreken we de algemene voorwaarden, terwijl in 6.5 uitgebreid de mogelijkheden in de stochastische designs aan de orde komen. In paragraaf 6.6. zullen we tenslotte nog kort ingaan op het schatten van persoonsparameters in onvolledige designs.

6.1 De relatie tussen onvolledige gegevens en IRT

Alhoewel de itemresponstheorie in het algemeen een aantal voordelen heeft boven de klassieke testtheorie (zie hoofdstuk 4), komen deze voordelen vooral goed tot uitdrukking als we IRT gaan toepassen in problemen waarbij er sprake is van onvolledige gegevens. Anderzijds is het zo, dat veel van de specifieke toepassingen van IRT alleen maar mogelijk zijn omdat onvolledige gegevens analyseerbaar zijn. In zekere zin is het dus zo dat IRT en onvolledige gegevens elkaar nodig hebben. Wij gaan hier aan de hand van enkele voorbeelden nader op in.

Een veel genoemde en geroemde eigenschap van IRT is dat personen met verschillende opgaven op dezelfde schaal gemeten kunnen worden. Ofwel iets nauwkeuriger geformuleerd, indien het IRT-model geldt voor een verzameling items in een of andere goed gedefinieerde populatie dan is het mogelijk de vaardigheid van personen uit deze populatie te schatten op dezelfde schaal op basis van antwoorden van verschillende deelverzamelingen items. Deze eigenschap maakt het bijvoorbeeld mogelijk om van twee verschillende toetsen met ver-schillende opgaven de resultaten op dezelfde schaal te vergelijken. Als de itemparameters van de items bekend verondersteld kunnen worden, dan kunnen we nagaan of verschillen in prestaties tussen bijvoorbeeld jaargroepen echte verschillen zijn zonder dezelfde opgaven te laten maken. Daarbij kunnen we een mogelijke alternatieve verklaring voor verschillen tussen groepen, dat de opgaven qua moeilijkheid verschillen, zoals die onder het klassieke testmodel mogelijk is, uitsluiten. Op de mogelijkheden en technieken om deze zogenaamde geëquiva-leerde toetsen te verkrijgen wordt in hoofdstuk 9 uitvoerig ingegaan. Hier wordt het slechts als voorbeeld genoemd van een toepassing van IRT die de analyse van een onvolledige data-matrix nodig heeft: twee groepen personen maken elk slechts een deel van de totale verzameling opgaven.

Een tweede algemeen genoemd voordeel van IRT is dat de itemparameters van IRT-modellen in meer of in mindere mate onafhankelijk van de getrokken steekproef geschat kunnen worden. Indien conditionele schattingsmethoden voor de itemparameters toepasbaar zijn, zoals in het Raschmodel en in het OPLM model (zie hoofdstuk 4 en 5), behoeven er zelfs in het geheel geen aannames te worden gedaan omtrent de verdeling van de vaardigheid van de steekproef waarmee de itemparameters geschat worden. Van deze eigenschap maken we natuurlijk gebruik als we van grotere verzamelingen items de parameterwaarden op dezelfde schaal willen hebben. Dit zogenaamde calibreren van de items gebeurt vaak op basis van gegevens uit onvolledige designs. Met name is dit het geval als we itembanken, hoofdstuk 1, gaan opbouwen met gecalibreerde opgaven. Het is in calibratie-onderzoek vaak alleen al praktisch

onmogelijk, vanwege de beschikbare testtijd, om alle opgaven aan alle leerlingen in de steekproef voor te leggen. Vanwege de genoemde eigenschap van de steekproef-onafhankelijkheid van de itemparameterschattingen is dit in IRT-modellen ook niet nodig.

6.1.1 Efficiëntie van de schattingen

Zijn er enerzijds vaak praktische redenen aanwezig die noodzaken tot onvolledige designs, in toepassingen van IRT zijn het doorgaans overwegingen van efficiëntie die leiden tot het gebruik van onvolledige designs. Met efficiëntie wordt hier bedoeld de statistische efficiëntie van de schattingen van de parameters.

We zullen aan de hand van een voorbeeld illustreren, dat de standaardfout van de itemparameterschattingen kleiner is naarmate de vaardigheid van de steekproef op basis waarvan de parameters worden geschat meer overeenkomt met de moeilijkheid van de items. In dit voorbeeld gebruiken we drie gesimuleerde dataverzamelingen. Deze dataverzamelingen hebben gemeenschappelijk dat ze elk uit 1000 antwoorden op 10 items bestaan. Verder is gemeenschappelijk dat alle items in elke dataverzameling het Raschmodel volgen:

$$P(X_i = 1 \mid \theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)}. \qquad (6.1)$$

Dat wil zeggen: item i, met moeilijkheid β_i, wordt door een persoon met vaardigheid θ, met de in (6.1) gegeven kans goed ($\{X_i = 1\}$) gemaakt. Tenslotte is gemeenschappelijk dat bij elke dataverzameling de vaardigheid van de personen aselect getrokken wordt uit de normale verdeling met gemiddelde 0 en variantie 1: θ is $N(0,1)$ verdeeld. De drie simulaties onderscheiden zich doordat de itemmoeilijkheden, waarmee de antwoorden volgens model (6.1) gegenereerd werden, verschilden. In de eerste simulatie was $\beta = 0$, in de tweede $\beta = 1$ en in de laatste $\beta = 2$. Dus steeds waren alle items in een simulatie even moeilijk, maar in de achtereenvolgende nam de moeilijkheid telkens met 1 toe en daarmee nam de overeen-stemming tussen de gemiddelde vaardigheid (0) en de itemmoeilijkheden per simulatie af.

Tabel 6.1

Geschatte itemmoeilijkheden, standaardfouten, p-waarden gesimuleerde gegevens, waarbij de afstand tussen het gemiddelde van de vaardigheid en de moeilijkheid toeneemt per simulatie

	simulatie 1			simulatie 2			simulatie 3		
item	$\hat{\beta}$	$SE(\hat{\beta})$	p	$\hat{\beta}$	$SE(\hat{\beta})$	p	$\hat{\beta}$	$SE(\hat{\beta})$	p
1	-0.120	.066	.528	0.072	.072	.281	-0.051	.086	.166
2	-0.076	.066	.519	-0.108	.070	.313	0.059	.088	.153
3	0.056	.066	.492	-0.080	.070	.308	0.016	.087	.158
4	-0.022	.066	.508	0.060	.072	.283	-0.170	.084	.181
5	-0.018	.066	.507	-0.047	.071	.302	0.033	.088	.156
6	0.031	.066	.497	-0.019	.071	.297	-0.035	.086	.164
7	0.046	.066	.494	-0.008	.071	.295	0.024	.088	.157
8	0.002	.066	.503	0.196	.073	.260	-0.010	.087	.161
9	-0.037	.066	.511	-0.058	.071	.304	0.050	.088	.154
10	0.139	.066	.475	-0.008	.071	.295	0.085	.089	.150

Het resultaat van de itemparameterschattingen met de standaardfouten en de klassieke p-waarden van de aldus gegenereerde antwoorden, bepaald met het programma OPLM (Verhelst, Glas & Verstralen, 1993), staan in tabel 6.1.We zien duidelijk het effect, dat de standaardfouten van de itemparameters kleiner zijn naarmate de vaardigheid van de steekproef beter in overeenstemming is met de moeilijkheid van de items, hoewel het aantal waarnemingen voor alle items 1000 is. De itemmoeilijkheden in de eerste simulatie worden het nauwkeurigste geschat. Naarmate de gemiddelde vaardigheid verder afligt van de moeilijkheid van de items wordt de standaardfout groter. Opgemerkt kan nog worden dat de standaardfouten van de items per simulatie ook enigszins verschillen, hetgeen veroorzaakt wordt doordat ook de SE's geschat worden (zie hoofdstuk 4).

Dit eenvoudige voorbeeld moge duidelijk maken dat de efficiëntie van de itemparameter-schattingen in het algemeen verhoogd kan worden door moeilijkheid en vaardigheid op elkaar af te stemmen. De efficiëntie van statistische schattingen wordt doorgaans uitgedrukt in het verschil of in de verhoudingen tussen de zogenaamde statistische informatie (zie hoofdstuk 4) die in een gegevensverzameling met betrekking tot een parameter aanwezig is. Voor een kwantificering van de informatiewinst met betrekking tot de itemparameterschattingen bij bepaalde onvolledige designs verwijzen wij naar Verhelst (1989). Het zal duidelijk zijn dat principieel dezelfde argumentatie geldt voor de schatting van de persoonsparameters en of van de kenmerken van de populatie personen: deze schattingen zullen efficiënter zijn naarmate de moeilijkheid van de voorgelegde items beter is afgestemd op de vaardigheid. In praktijk-toepassingen zijn, in tegenstelling tot het hiervoor geschetste voorbeeld, de items niet even moeilijk en hebben de personen niet dezelfde vaardigheid. We kunnen dus aan efficiëntie

winnen door de moeilijkste items aan de meest vaardige personen voor te leggen en de gemakkelijkste aan de minst vaardige. Dit resulteert uiteraard in een onvolledig design.

6.1.2 Calibratie in onvolledige designs en linken

Met name in de Amerikaanse psychometrische literatuur, bijvoorbeeld Hambleton en Swaminathan (1985), wordt calibreren in onvolledige designs vaak beschreven als een activiteit die in twee fasen uiteenvalt. De eerste is het calibreren in volledige deeldesigns, waarna in de tweede fase de parameters, om onderling vergelijkbaar te kunnen zijn, via het zogenaamde 'linken' op dezelfde schaal worden gebracht. Men noemt dit ook wel het equivaleren van de itemparameters.

Zoals bekend (hoofdstuk 4) wordt tijdens het calibratieproces de schaal op enigszins arbitraire wijze gefixeerd. We fixeren de schaal tijdens de calibratie, als we met de CML-schattingsmethode werken, zoals in het Raschmodel en het OPLM model vaak door de som van de geschatte itemmoeilijkheden (en dus ook het gemiddelde) op 0 te stellen: $\sum_{i=1}^{k} \hat{\beta}_i = 0$. Een andere mogelijkheid die veelal wordt toegepast bij calibratie met MML is de schaal te fixeren zodanig dat het gemiddelde van de steekproefverdeling van de vaardigheid θ vastgelegd wordt op 0 en de variantie van deze verdeling op 1. In het algemeen is het echter zo dat we de gekozen schaal op willekeurige wijze lineair kunnen transformeren. Zoals uiteengezet in hoofdstuk 4 veranderen we daardoor slechts het willekeurig te kiezen nulpunt en de eenheid van de schaal.

Als voorbeeld hiervan blikken we even terug op de resultaten van tabel 6.1 Daar zien we dat de geschatte moeilijkheden tussen de simulaties nauwelijks verschillen, ondanks dat we weten dat er wel verschillen zijn. Duidelijk is dat te zien in tabel 6.1 aan de klassieke p-waarden. Waaruit volgt dat per calibratie de schaal op dezelfde willekeurige wijze gefixeerd is en dat de waarden van de itemparameters per simulatie op een andere niet vergelijkbare schaal liggen. Om de moeilijkheidsschattingen van de items in de drie simulaties te kunnen vergelijken zullen er nog transformaties nodig zijn die de parameterschattingen op dezelfde schaal brengen.

Hoe dit in zijn werk zou kunnen gaan, zullen we toelichten met een ander voorbeeld. In dit voorbeeld hebben we een onvolledig design en wordt in twee aparte calibraties de schaal gefixeerd, waarna er bij het verbinden van de schalen ervoor gezorgd wordt dat de itemparameters van beide groepen items op dezelfde schaal komen te liggen. Dit komt neer op het vinden van een transformatie van een van de, of eventueel van beide, gecalibreerde schalen. Zo'n transformatie kan op verschillende manieren worden

bepaald en uitgevoerd. Een ervan zullen we met ons voorbeeld toelichten. We beschouwen een design met twee groepen van tien items en twee groepen personen. Hierbij zijn item 1 tot en met 5 gemaakt door de eerste groep, de items 6 tot en met 10 alleen door tweede en de items 11 tot en met 15 door beide groepen. Om zeker te zijn de items aan een IRT-model voldoen, zijn antwoorden op de items conform het Raschmodel (6.1) gegenereerd. In beide groepen werden 1000 antwoordpatronen gegenereerd. De calibratie van de items in beide groepen apart, dat wil zeggen per volledig deeldesign, met de CML-schattingsmethode van het programma OPLM leverde de in tabel 6.2 gegeven schattingen van de moeilijkheid op.

We zien in tabel 6.2 dat voor item 11 tot en met 15 ondanks dat het dezelfde items zijn en ondanks dat we weten zeker weten dat het Raschmodel geldt de geschatte moeilijkheden tussen de calibraties nogal verschillen. Deze verschillen kunnen twee oorzaken hebben. Kleinere fluctuaties kunnen veroorzaakt worden door de steekproef, want de steekproeven zijn eindig. Systematische verschillen worden echter veroorzaakt doordat in beide calibraties op een arbitraire wijze het nulpunt van de schaal is vastgelegd, zodanig dat de gemiddelde moeilijkheid in de te calibreren toets 0 is. De eenheid van de schaal is in dit voorbeeld van het Raschmodel op dezelfde wijze vastgelegd: alle discriminatie-indices zijn in beide calibraties gelijk aan 1 gekozen. Een manier, zie bijvoorbeeld ook Wright en Stone (1979), om alle itemparameters vergelijkbaar en dus op één schaal te krijgen is de volgende.

Tabel 6.2

Geschatte itemmoeilijkheden in een onvolledig design met overlappende items per volledig deeldesign met de verschillen tussen de gemeenschappelijke items

Item	Calibratie 1 $\hat{\beta}^{(1)}$	Calibratie 2 $\hat{\beta}^{(2)}$	$\hat{\beta}^{(2)} - \hat{\beta}^{(1)}$
1	-2.041		
2	-0.927		
3	0.093		
4	0.976		
5	1.919		
6		-0.533	
7		-0.489	
8		-0.445	
9		-0.430	
10		-0.626	
11	0.026	0.481	.455
12	-0.051	0.545	.596

13	-0.109	0.453	.562
14	0.035	0.527	.492
15	0.079	0.516	.437
Gem.	0.000	0.000	.508

Bepaal in eerste instantie de verschillen tussen moeilijkheidsschattingen van de gemeenschappelijk items. Het resultaat staat in de vierde kolom van tabel 6.2. Het gemiddelde verschil per item in geschatte moeilijkheid tussen beide calibraties is 2.542/5 = .508. Een manier om de itemparameters van de eerste calibratie op de schaal van tweede calibratie te krijgen is simpel het optellen van dit gemiddelde verschil bij alle geschatte moeilijkheden van de eerste calibratie. Het resultaat staat in tabel 6.3. Omdat we nu voor de gemeenschappelijke items 11 tot en met 15 beschikken over twee schattingen van de moeilijkheid, die variëren door statistische variatie, zouden we als uiteindelijk schattingen voor deze items het gemiddelde kunnen nemen. Het resultaat van de op deze wijze op dezelfde schaal gebrachte schattingen van de itemparameters staat in de vierde kolom van tabel 6.3. We zien dat het gemiddelde van de geschatte moeilijkheden op deze schaal 2.560/15 = .171 bedraagt.

Tabel 6.3

Het op dezelfde schaal brengen van in volledige deeldesigns geschatte itemmoeilijkheden het resultaat van een simulatie calibratie

Item	Calibratie 1 $\hat{\beta}^{(1)} + .508$	Calibratie 2 $\hat{\beta}^{(2)}$	Calibratie	Calibratie gem .00	Calibratie simultaan
1	-1.533		-1.533	-1.704	-1.703
2	-0.418		-0.418	-0.589	-0.589
3	0.601		0.601	0.430	0.431
4	1.484		1.484	1.313	1.314
5	2.427		2.427	2.256	2.256
6		-0.533	-0.533	-0.704	-0.704
7		-0.489	-0.489	-0.660	-0.660
8		-0.445	-0.445	-0.616	-0.616
9		-0.430	-0.430	-0.601	-0.601
10		-0.626	-0.626	-0.797	-0.797
11	0.534	0.481	0.508	0.337	0.339
12	0.457	0.545	0.501	0.330	0.326
13	0.399	0.453	0.426	0.255	0.253
14	0.543	0.527	0.535	0.364	0.366
15	0.587	0.516	0.552	0.381	0.384
Gem.	0.508	0.000	0.171	0.000	0.000

Daarmee hebben we dus bereikt dat de moeilijkheidsparameters van alle items op dezelfde schaal zijn gebracht en daardoor onderling vergelijkbaar zijn. Tenslotte kunnen we voor de totale itemverzameling op gebruikelijke wijze de schaal fixeren, zodanig dat gemiddelde moeilijkheid over alle items 0.000 wordt. Dit bereiken we eenvoudig door van alle geschatte moeilijkheden 0.171 af te trekken. Het resultaat staat in de vijfde kolom van tabel 6.3.

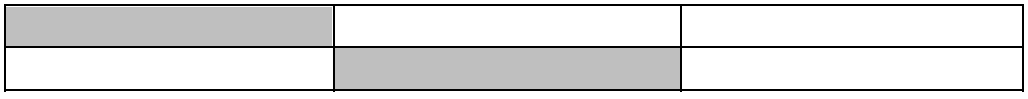
Wij zullen niet nader op ingaan op de verschillende andere manieren, die in de psycho-metrische literatuur zijn voorgesteld om in verschillende onvolledige designs een 'linktransformatie' te bepalen om parameters op één schaal te brengen. De reden hiervoor is dat het calibreren in een onvolledige gegevensverzameling ook beschouwd kan worden als een simultaan proces, waarin naast het schatten van de parameters deze tevens op dezelfde schaal worden afgebeeld. Het onderscheid in fasen, calibreren in volledige deeldesigns en vervolgens linken, dat in de literatuur vaak wordt gemaakt, is historisch ontstaan en is eigenlijk niet meer functioneel. De schattings- en toetsingstheorie voor IRT-modellen is in eerste instantie ontwikkeld voor volledige designs. En oudere computerprogrammatuur voor de calibratie kon dan ook alleen

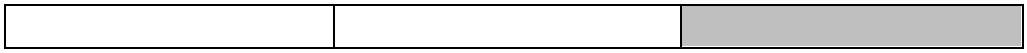
maar volledige designs analyseren en daarom moest het proces in twee fasen verlopen. Tegenwoordig is echter de theorie voor het schatten en toetsen in onvolledige designs zo ver ontwikkeld, dat ze geïmplementeerd is in programmatuur (bijvoorbeeld OPLM) zodat de traditionele omweg niet meer noodzakelijk is: calibratie vindt plaats in onvolledige designs, waarbij de itemparameters op dezelfde schaal komen te liggen door gebruik te maken van de gemeenschappelijk elementen in de deeldesigns, en de schaal wordt in een keer voor de totale gegevensverzameling gefixeerd. Ter illustratie zijn in de laatste kolom van tabel 6.3 de resultaten van de simultane calibratie van alle opgaven in de onvolledige gegevensverzameling met OPLM opgenomen. Zoals het resultaat laat zien, is er nauwelijks sprake van verschillen in de geschatte moeilijkheden. Merk echter op dat de standaardfouten van de itemparameterschattingen bij simultane calibratie kleiner worden dan bij combinatie van afzonderlijke calibraties. Zie hiervoor Vale (1986) en Verhelst (1993). Het calibreren in volledige deeldesigns en daarna de parameters op dezelfde schaal brengen of equivaleren moet dus zo mogelijk vervangen worden door simultane calibratie in een onvolledig design.

Of we in een keer in een onvolledig design de schaal fixeren, dan wel in fasen, er zal altijd tussen de volledige deeldesigns iets gemeenschappelijks moeten zijn, dat er voor kan zorgen dat de parameters op dezelfde schaal kunnen worden gebracht. De gemeenschappelijkheid kan liggen in de personen die verschillende items maken, dan wel in de items die door personen worden gemaakt. Voor deze zogenaamde ankering zijn verschillende mogelijkheden die we in de volgende paragraaf zullen bespreken. Het anker zorgt ervoor dat er een basis is voor de vergelijking tussen verschillende calibraties, dan wel dat in een calibratie de schaal eenduidig kan worden gefixeerd.

6.2 De datamatrices van structureel onvolledige designs

In deze paragraaf zullen we de in praktijk meest voorkomende structureel onvolledige designs beschrijven. We zullen dat doen door aan te geven hoe de uiteindelijk te analyseren datamatrix eruit ziet. In de figuren waarin de designs zijn gegeven, staan steeds verticaal personen en horizontaal items. Door arcering is aangegeven welke (groepen) personen welke (groepen) items hebben gemaakt. De niet-gearceerde gebieden geven de ontbrekende gegevens. Steeds zullen we aangeven hoe de in het voorgaande genoemde ankering plaatsvindt.





Figuur 6.1
Niet-verbonden of niet-geankerd design

In figuur 6.1 is schematisch een voorbeeld datamatrix weergegeven waarbij er geen overlap is tussen de drie toetsen en evenmin overlap tussen de drie groepen personen die de toetsen maken. Dit design wordt in de praktijk veel toegepast ondanks dat dit een design is, waarbij de wijze van ankering tussen de delen niet aan de datamatrix is te zien. Als de parameters van de opgaven in deze toetsen op dezelfde schaal moeten worden gebracht, zal het duidelijk zijn dat dit niet via gemeenschappelijke items of personen kan. Dus de gegevens zullen op een speciale manier verzameld moeten zijn, dan wel zullen er extra aannames nodig zijn, omtrent de wijze waarop de gegevensverzameling tot stand is gekomen om de onderdelen aan elkaar te verbinden. Een veel gebruikte opzet hierbij is dat statistisch equivalente groepen de verschillende toetsen maken, hetgeen in de praktijk goed gerealiseerd kan worden door leerlingen die aselect zijn getrokken uit een populatie aselect toe te wijzen aan de toetsen. Dit wordt dan het ’random group design’ genoemd. Omgekeerd zou men op soortgelijke wijze kunnen veronderstellen of bewerkstelligen dat er equivalente toetsen zijn. Dit komt in de praktijk minder vaak voor.

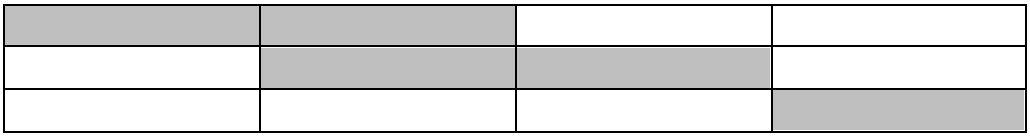
Hoe het ook zij, het niet-geankerd design, waarbij de groepen proefpersonen even groot zijn, heeft in het algemeen als voordeel dat elk item in principe even vaak wordt afgenomen. Alhoewel er geen garantie is dat een gelijk aantal afnames per item tot even nauwkeurige schattingen van de itemparameters zal leiden, geeft dit zolang er geen a priori informatie over de itemparameters of de vaardigheid van de steekproeven leerlingen aanwezig is, de beste gelegenheid voor even precieze schattingen van alle items. Bovendien is het niet-geankerde design in sommige toepassingen het enig mogelijke design. Bijvoorbeeld bij examens waarbij geheimhouding van de opgaven een belangrijke rol speelt.

Figuur 6.2
Ankeritemsdesign

Het in de praktijk traditioneel meest voorkomende design is in figuur 6.2 in zijn meest simpele vorm weergeven. In dit design met ankeritems, ’common items design’

of soms ook wel ankertoets design genoemd, wordt een deelverzameling van de items door beide onderscheiden groepen personen gemaakt. De itemparameters worden in de calibratie op een schaal gebracht via de items die gemeenschappelijk zijn afgenomen. Het zal duidelijk zijn dat dit design eenvoudig naar meer groepen items en personen kan worden gegeneraliseerd. Het belangrijkste voordeel van dit design is dat in de analyse noch de equivalentie van de groepen personen, noch van de groepen items verondersteld hoeft te worden. Een mogelijk nadeel is dat de parameters van de gemeenschappelijke items in het design nauwkeuriger geschat zullen worden dan de items die slechts in een toets voorkomen, want de gemeenschappelijke items worden door meer personen beantwoord.

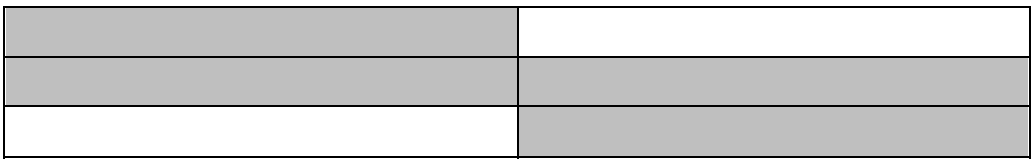
De designs die hier worden besproken komen in de praktijk om diverse redenen ook in allerlei combinaties voor. Een voorbeeld hiervan staat in figuur 6.3.



Figuur 6.3
Gedeeltelijk verbonden design

Figuur 6.3 geeft een slechts gedeeltelijk verbonden design. De items van toets 1 en toets 2 zijn via een ankertoets wel verbonden, terwijl de items van toets 3 niet verbonden zijn met de items van toets 1 of toets 2. Dit design heeft de voor- en nadelen van de basisdesigns waaruit het is samengesteld.

Een variant op het klassieke ankeritemsdesign of ankertoets design is het ankergroep-design. Zie figuur 6.4.



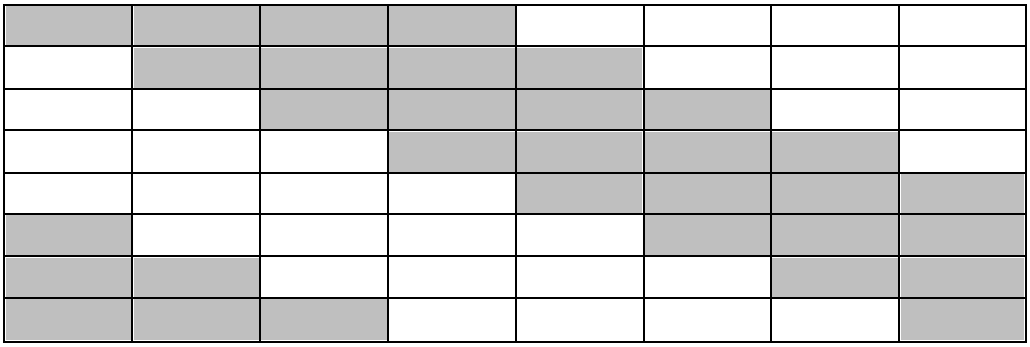
Figuur 6.4
Ankergroepdesign

Het ankergroepdesign, 'common person design,' is eigenlijk het gespiegelde van het ankeritemsdesign. De itemparameters worden op de gemeenschappelijke schaal geplaatst door de vaardigheden van de leerlingen die in dit voorbeeld de beide toetsen maken. Ook in dit design is het niet nodig aan te nemen dat groepen items of

leerlingen equivalent zijn. Alle opgaven worden in principe even nauwkeurig geschat echter ten koste van de ongelijkheid van de nauwkeurigheid waarmee personen kunnen worden geschat. Verder is een praktisch nadeel van dit design dat het moeilijk kan zijn om een groep leerlingen te vinden die alle opgaven kan maken.

Uiteraard kan men het ankergroepdesign en het ankeritemsdesign ook weer combineren en een dubbel anker leggen, zowel over personen als over groepen. Veel voordelen heeft zo'n design echter niet, men houdt namelijk het ongelijke aantal waarnemingen per item en per persoon.

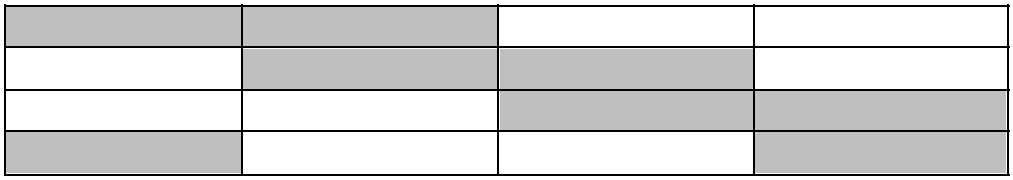
Het nadeel van het ongelijke aantal waarnemingen per opgave en per persoon wordt opgelost in zogenaamde ineengestrengelde of kettingdesigns, 'interlaced design', Vale (1986). In zijn meest extreme vorm heeft zo'n design evenveel verschillende toetsen als er opgaven of items zijn. In figuur 6.5 is een voorbeeldje gegeven met in totaal acht items waarbij elke toets bestaat uit vier items.



Figuur 6.5
Ineengestrengeld of kettingdesign: een item per blokje

De eerste toets begint met item 1 en bestaat verder uit de daaropvolgende items totdat de toets zijn vastgelegde lengte bereikt. De tweede toets begint met tweede item. Enzovoort, totdat elk item eenmaal het eerste item in een toets is geweest. Een voordeel van dit design is dat er duidelijk een ankeritem effect wordt bereikt, terwijl toch het aantal afnames per item en de toetslengte per persoon in totaal gelijk is. Indien de aldus ontstane toetsen aselect over de groepen worden verdeeld zijn ook de groepen statistisch equivalent. Het nadeel van dit design is praktisch van aard: er moeten net zoveel boekjes gedrukt als er items zijn. Dit design zal dus in toepassingen met grotere aantallen items alleen gerealiseerd kunnen worden als de items via de computer worden aangeboden. Zolang de toetsen op papier worden gedrukt is een praktische bruikbare en zeer aantrekkelijke variant van het volledig ineengestrengelde design het geblokt kettingdesign. In figuur 6.6 is daarvan een voorbeeld gegeven. De

blokken bevatten hierbij meerdere items. Als we, als in figuur 6.5, in totaal acht items hebben, bestaat elk blokje in figuur 6.6 dus uit twee items.



Figuur 6.6
Geblokt kettingdesign

In dit design zal het equivalente groepen effect wellicht minder bereikt, echter de voordelen van het design zijn evident: er zijn slechts een beperkt aantal fysieke toetsboekjes nodig en alle items worden in dit design ook weer even vaak afgenomen.

6.3 De stochastische structuur van structureel onvolledige designs

In deze paragraaf zullen we nader ingaan op de verschillende soorten structureel onvolledige gegevens design die in de IRT veel gebruikt worden. Wij onderscheiden de drie designtypen, die in de praktijk het meest voorkomen. De designs onderscheiden zich van elkaar door het mechanisme of procedure waardoor de ontbrekende gegevens in het design, lege cellen in de datamatrix, ontstaan. Dit mechanisme zullen we beschrijven als een toevalsmechanisme: door middel van kansen of verdelingen is aan te geven dat bepaalde waarnemingen wel of niet zullen voorkomen in de datamatrix. Vandaar dat we spreken over de stochastische structuur van de designs. In de paragrafen 6.5 en 6.6 zullen we bekijken in welke omstandigheden bij het schatten van de modelparameters rekening gehouden moet worden met het toevalsmechanisme dat de lege cellen in datamatrices veroorzaakt. Voor de goede orde wijzen wij erop dat bij de designs, die hierna worden beschreven, in principe alle in paragraaf 6.2 beschreven datamatrices kunnen voorkomen.

Voor de beschrijving spreken we eerst wat notatie af. In totaal beschouwen we een verzameling van k items. Hieruit worden B toetsboekjes samengesteld, geïndexeerd met $b = 1,...,B$. Elk boekje bevat $k_b, b = 1,...,B$ items, die elkaar, eventueel deels, over-lappen. Elke persoon maakt de items uit slechts één boekje. Voor elke persoon $v, v = 1,...,n$ definiëren we een zogenaamde itemindicator variabele. Deze variabele is een vector, die evenveel elementen bevat als het totaal aantal, k, opgaven: $\boldsymbol{R}_v = (R_{v1},...,R_{vk})$. Elk element van de itemindicator vector kan de waarde 1 of 0

aannemen al naar gelang de persoon het betreffende item maakt of niet. De itemindicator vector kan B verschillende waarden aannemen, net zoveel als er verschillende toetsboekjes zijn. De waarde voor bijvoorbeeld toetsboekje 1 bestaat uit een vector met een lengte van k met daarin k_1 enen en $k - k_1$ nullen op de plaatsen, die de items uit de totale verzameling indiceren, respectievelijk voor items die in het toetsboekje zitten en voor items die er niet in zitten. In het algemeen neemt de itemindicator de waarden $\boldsymbol{r}_b$ aan die staan voor een permutatie van k_b, het aantal items in toetsboekje b, enen en $k - k_b$ nullen voor, $b = 1,...,B$. Dat wil zeggen dat van een persoon v de itemindicator $\boldsymbol{R}_v$ de waarde $\boldsymbol{r}_b$ heeft, als deze persoon boekje b heeft gemaakt. In de hiernavolgende bespreking van de drie meest voorkomende stochastische designs zal steeds worden aangegeven wat de verdeling is van deze itemindicator.

6.3.1 Gerandomiseerd onvolledig design

In gerandomiseerde ofwel volledig door het toeval bepaalde designs, 'randomized-incompletedesign', besluit een onderzoeker zonder gebruik te maken van a priori kennis van de vaardigheid van de persoon met een van te voren bekende kans een van de B toetsboekjes aan een persoon toe te wijzen. In de praktijk worden in deze designs vaak uit de beschikbare itemverzameling B boekjes samengesteld, die een even groot aantal items bevatten en vaak ook nog nominaal parallel zijn, dat wil zeggen, gelijk qua inhoudelijke samenstelling en qua ingeschatte moeilijkheid. De toewijzing van een boekje aan een persoon kan natuurlijk echt aselect geschieden: elke persoon krijgt met een even grote kans, en wel $1/B$, een bepaald boekje te maken. Meer algemeen krijgt een persoon een boekje met bekende kans ϕ_b, zodanig dat $\sum_{b=1}^{B}\phi_b = 1$. In het algemeen wordt de verdeling van de itemindicator gegeven door

$$P(\boldsymbol{R}_v = \boldsymbol{r}_b) = \phi_b. \qquad (6.2)$$

Dit geldt voor alle personen $v = 1,...,n$ en alle toetsboekjes $b = 1,...,B$. De belangrijkste reden om gerandomiseerde designs in IRT calibratie-onderzoek te gebruiken is dat het doorgaans fysiek onmogelijk is om leerlingen alle opgaven uit de verzameling te calibreren opgaven te laten maken. Zolang men bij de opzet geen gebruik kan of wil maken van a priori kennis over de vaardigheid van de leerlingen en of de moeilijkheid van de opgaven, zijn gerandomiseerde designs het meest praktisch en naar verwachting het meest efficiënt voor de calibratie van alle opgaven.

Een bijzonder geval van gerandomiseerde onvolledige designs zijn de in de praktijk vaak voorkomende a priori gefixeerde onvolledige designs. Dat zijn designs waarin de verdeling van de itemindicator gegeven wordt door

$$P(\boldsymbol{R}_v = \boldsymbol{r}_b) = 0 \quad \text{of} \quad 1. \tag{6.3}$$

Met andere woorden, van te voren is met kans 1 bepaald wie welk toetsboekje krijgt. Van belang hierbij is op te merken, dat in de toekenning van een toetsboekje aan een persoon de kenmerken van de persoon ook geen rol spelen. Als dat het geval zou zijn dan hebben we een designtype dat in paragraaf 6.3.3 wordt besproken. Gefixeerde onvolledige designs zijn de designs die in de inleiding van dit hoofdstuk beschreven werden als de structureel onvolledige die in de eerste categorie vallen. De categorie waarbij de onderzoeker volledig onder controle heeft waar de lege cellen in datamatrix zullen zitten. De gerandomiseerde designs in het algemeen en ook de designs die hierna worden beschreven vallen onder de tweede categorie: slechts de procedure volgens welke de ontbrekende gegevens ontstaan, staat onder controle van de onderzoeker.

6.3.2 Meerfasen onvolledig design

In meerfasen designs, 'multistage testing design', is de toewijzing van items aan personen mede afhankelijk van de resultaten die de personen op een deel van de items halen. In de eerste fase krijgen bijvoorbeeld alle personen dezelfde deelverzameling items, meestal van middelmatige moeilijkheid, uit de totale itemverzameling te maken. Op grond van de scores op deze eerste groep items, die je de sorteertoets zou kunnen noemen, maken de personen in fase twee verschillende items. Bijvoorbeeld personen met hoge scores op de sorteertoets maken in fase twee een deelverzameling items uit de totale itemverzameling die van te voren wat moeilijker ingeschat wordt, terwijl personen met lage scores een verzameling gemakkelijker geachte items maken. Een simpel voorbeeld met een totale itemverzameling bestaande uit twintig items. Tien (nummers 1 tot 10) zijn er middelmatig moeilijk, vijf (item 11 tot 15) worden redelijk gemakkelijk geacht, en de laatste vijf zijn items waarvan de geschatte moeilijkheid wat hoger ligt (item 16 tot 20). Een tweefasen design zou er dan uit kunnen zien als in figuur 6.7 is aangegeven.

	Items		
Leerlingen	1 t/m 10	11 t/m 15	16 t/m 20
	$0 \leq s \leq 5$		
	$6 \leq s \leq 10$		
	Fase 1	Fase 2	

Figuur 6.7
Tweefasen design

De sorteertoets bestaat uit de middelmatig moeilijke items, is de somscore s hierop meer dan 5 dan maakt de persoon in fase twee de moeilijker ingeschatte items (16 tot 20), anders de gemakkelijker items.

Het zal duidelijk zijn dat dit sorteerproces in principe ook in een tweede fase kan worden voortgezet en in een derde fase of nog verder. Het sorteren op grond van een verzameling items hoeft natuurlijk niet plaats te vinden in twee groepen, maar evengoed kunnen meerdere groepen worden onderscheiden, die evenveel verschillende trajecten starten in de item-verzameling. Essentieel voor meerfasen toetsen is dat de selectie items die een persoon uiteindelijk maakt direct afhankelijk is van de score op items die eerder door deze persoon zijn gemaakt.

De uiteindelijke verzameling items die een persoon maakt duiden we, als eerder, weer aan met boekje b. De verdeling van de itemindicator voor een persoon in meerfasen toetsen wordt dan gegeven door de kans dat een bepaald boekje wordt gemaakt. Deze kans is 0 of 1 afhankelijk van het wel of niet voldaan zijn aan de criteria die gesteld worden aan de geobserveerde itemscores om een bepaald boekje te krijgen. In het voorbeeld uit figuur 6.7 krijgt men met kans 1 boekje 1 als $s_v = \sum_{i=1}^{10} x_{vi} \leq 5$, waarin x_{vi} de score is van persoon v op item i ,en met kans 0 boekje 2; als $s_v \geq 6$ is de kans op boekje 1 gelijk aan 0 en op boekje 2 gelijk aan 1. Algemener geldt natuurlijk ook dat als we alle itemscores van een persoon gegeven hebben, de kans op een bepaald boekje ook 0 of 1 is. Als we de vector van de van persoon v geobserveerde itemscores schrijven als $\boldsymbol{X}_{obs,v}$, met obs,v de verzameling van alle itemnummers of indexen die deze persoon maakt, dan geldt

$$P(\boldsymbol{R}_v = \boldsymbol{r}_b \mid \boldsymbol{x}_{obs,v}) = 0 \quad \text{of} \quad 1. \tag{6.4}$$

Dit geldt weer voor alle personen $v = 1,...,n$ en alle toetsboekjes $b = 1,...,B$.

Het idee achter meerfasen toetsen is dat daarmee de efficiëntie van de schattingen kan worden verhoogd, doordat met de toewijzing van de items aan persoon afstemming

plaats vindt tussen de van te voren ingeschatte moeilijkheid van de items en de tussentijds ingeschatte vaardigheid van de personen. Het zal duidelijk zijn dat naarmate er meer fasen worden onderscheiden in principe het afstemmen van moeilijkheid op vaardigheid nauwkeuriger kan gebeuren. Meerfasen designs vinden toepassing bij zowel calibratie-onderzoek als in situaties waarin we bijvoorbeeld met behulp van een gecalibreerde item-verzameling persoonsparameters willen schatten. Adaptief toetsen is eigenlijk een limietgeval van meerfasen toetsen; daarbij zijn er voor elke persoon evenveel fases als hij of zij items maakt. Het aantal items zal hierbij per persoon in het algemeen verschillen. Na elke itemafname wordt op grond van een voorlopige schatting van de vaardigheid, gebaseerd op de tot dan toe gemaakte items, een nieuw item gekozen waarvan de moeilijkheid het best in overeenstemming met deze vaardigheid. Gestopt wordt met toetsen, zodra de vaardigheid van de persoon met vooraf vastgestelde nauwkeurigheid kan worden geschat. Adaptief toetsen wordt in calibratie opzetten niet toegepast omdat criteria om het beste item uit een verzameling beschikbare te kiezen eigenlijk alleen met bekend (veronderstelde) itemparameters goed gekwantificeerd kunnen worden. Als het gaat om de vaardigheid van personen te schatten is adaptief toetsen de meest efficiënte vorm van toetsen.

6.3.3 Groepsgericht onvolledig design

In groepsgerichte designs, 'targeted testing design', wordt de toewijzing van de items aan de personen bepaald op basis van te voren bekende achtergrondinformatie van de persoon. Die achtergrondinformatie kunnen we uitdrukken door de waarden die een toevalsvariabele Y aanneemt. Dan hangt Y doorgaans positief samen met de vaardigheid van de leerlingen. Groepsgerichte designs zien er dan zo uit dat de gemakkelijker geachte boekje(s) gemaakt worden leerlingen met waarden van Y die naar verwachting samengaan met een geringere vaardigheid; leerlingen met Y waarden die duiden op een hogere vaardigheid maken de naar verwachting moeilijke boekje(s). Efficiëntie winst in de schatting door betere afstemming van de vaardigheden op de moeilijkheden wordt hierbij weer verwacht. Zonder dat dit de algemeenheid beperkt, nemen we aan dat we van de achtergrondvariabele Y evenveel waarden onderscheiden als verschillende toetsboekjes (B) in het design. Die waarden zijn dus in het algemeen $y_1,...,y_B$. Bij elke waarde y_b wordt een ander boekje b gemaakt. Dit boekje bestaat uit een deelverzameling items uit de totale itemverzameling. De waarde van de itemindicator van een persoon die dit boekje maakt is $\boldsymbol{r}_b$. Dan kunnen we als voorheen de verdeling van de itemindicator in groepsgerichte designs schrijven als:

$$P(\boldsymbol{R}_v = \boldsymbol{r}_b \mid Y_v = y_b) = 1,$$
$$P(\boldsymbol{R}_v = \boldsymbol{r}_b \mid Y_v \neq y_b) = 0, \tag{6.5}$$

voor alle personen $v = 1,...,n$ en voor alle te onderscheiden waarden van de achtergrond-variabele $b = 1,...,B$.

Bij groepsgerichte designs zijn twee situaties te onderscheiden met betrekking tot de rol die de achtergrondvariabele in de analyse en eventueel in de steekproeftrekking speelt. In de eerste is de rol van de achtergrondvariabele zeer beperkt: hij wordt alleen maar gebruikt om de efficiëntie van de schattingen te verhogen en zijn we niet geïnteresseerd in de resultaten van leerlingen met bepaalde waarden van de achtergrondvariabele. De tweede en in de praktijk meest voorkomende rol van de achtergrondvariabele is dat we ook in de vaardigheids-verdelingen bij verschillende waarden van achtergrondvariabele geïnteresseerd zijn. De totale populatie wordt door de achtergrondvariabele opgedeeld in een aantal subpopulaties die ons interesseren.

Een concreet voorbeeld van de eerste situatie deed zich voor bij het Periodiek Peilings Onderzoek (PPON) in het basisonderwijs (Verhelst & Eggen, 1989), waarbij het geschatte niveau van de leerling door de leerkracht bepaalde welke toets de leerling maakte. Dit voorbeeld wordt uitgebreid besproken in paragraaf 7.1. Hier zij slechts vermeld dat in dit onderzoek het leerkrachtoordeel gebruikt werd om de efficiëntie van het design te verhogen, zonder dat men geïnteresseerd in de variabele zelf.

De tweede situatie komt in de praktijk regelmatig voor. Behalve in de itemparameters zijn we ook geïnteresseerd in de vaardigheidsverdelingen van de onderscheiden groepen. Stel dat we bijvoorbeeld een verzameling items die luistervaardigheid meten, willen calibreren voor de populatie van leerlingen uit het derde leerjaar van het VBO en het MAVO. In dat geval zal de verdeling van de vaardigheid in de subpopulaties VBO en MAVO zeker interessant zijn. In de praktijk komt de interesse in de verschillende vaardigheidsverdelingen daarbij vaak expliciet naar voren als men ten behoeve van het calibratie-onderzoek geen aselecte steekproef uit de totale populatie van derde klassers VBO en MAVO trekt, maar een gestratificeerde steekproef: per schooltype trekt men een aselecte steekproef. Om er zeker van te zijn dat per subpopulatie de vaardigheidsverdelingen even nauwkeurig kunnen worden geschat, zijn de aantallen leerlingen uit de subpopulaties in de steekproef vaak even groot, maar de proporties uit de verschillende subpopulaties niet noodzakelijk gelijk aan de proporties in de totale populatie. Zodat we niet meer beschikken over een aselecte steekproef uit de totale populatie.

6.4 Algemene voorwaarden voor calibratie in onvolledige designs

In deze paragraaf zullen we ingaan op de algemene voorwaarden die moeten gelden voor het bestaan van eindige en unieke itemparameterschattingen voor zowel de CML- als de MML-methode in onvolledige designs. We bespreken hier in feite alleen de voorwaarden die moeten gelden in gefixeerde onvolledige designs, waarbij de onderzoeker het ontstaan van de onvolledige gegevens volledig onder controle heeft. Zie de itemindicator verdeling (6.3). In paragraaf 6.5 gaan we dan in op de nadere voorwaarden die gesteld moeten worden aan een calibratiemethode bij stochastische designs.

In gefixeerde onvolledige designs geldt voor de calibratie, met welke methode dan ook, dat het in ieder geval noodzakelijk is dat er tussen de verschillende te onderscheiden volledige deeldesigns iets gemeenschappelijk is. In paragraaf 6.1 werd al aangegeven dat dit nodig is om in een onvolledig design de itemparameters op één schaal te kunnen brengen. Om ervan verzekerd te zijn voor alle parameters unieke schattingen te krijgen moet deze voorwaarde nog iets worden aangescherpt. In de psychometrische literatuur zijn de voorwaarden voor het bestaan van en het uniek zijn van CML-schattingen in gefixeerde onvolledige designs in het Raschmodel exact uitgewerkt door Fischer (1981). Omdat de voorwaarden aan het design voor het bestaan van CML-schattingen strenger zijn dan voor het bestaan van MML-schattingen, zullen we deze hierna kort schetsen. Over de minder strenge condities aan het design bij MML zullen we daarna enkele opmerkingen maken.

Fischer (1981) toont in eerste instantie aan onder welke voorwaarden er eindige en unieke CML-schattingen voor de itemparameters in volledige designs bestaan, waarna hij zijn resultaten generaliseert naar het bestaan en uniek zijn van de schattingen in onvolledige designs. We geven nu, zonder op details in te gaan, een beschrijving van deze voorwaarden. In volledige designs worden Fischers voorwaarden gesteld aan de datamatrix van alle itemantwoorden:

$$\boldsymbol{x} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & \cdots & x_{nk} \end{bmatrix}.$$

De rij-index van deze matrix geeft een persoon aan, de kolom-index een item. Om itemparameterschattingen te verkrijgen is het noodzakelijk dat de kolomsommen uit

deze matrix $t_j = \sum_{v=1}^{n} x_{vj}$ niet gelijk zijn aan 0, iedereen maakt de opgave fout, of aan n, iedereen maakt de opgave goed. Zoals we in hoofdstuk 4 zagen bereikt de aannemelijkheidsfunctie voor zo'n item zijn maximum bij respectievelijk $-\infty$ en ∞ en bestaat er dus geen eindige schatting van de itemparameter voor dat item. Aan deze voorwaarde moet voor elk item $j = 1,\ldots,k$ voldaan zijn. Fischer geeft aan dat voor de gehele datamatrix $\boldsymbol{x}$ nog iets meer moet gelden: het mag niet zo zijn dat deze uiteenvalt in twee delen die geen verbinding met elkaar hebben. Hij definieert daarvoor het begrip 'goed geconditioneerd' zijn van de datamatrix en toont aan dat het goed geconditioneerd zijn van de datamatrix de voorwaarde is voor het bestaan van unieke schattingen van de itemparameters. Een datamatrix is goed geconditioneerd als in elke mogelijke opdeling van de items in twee niet-lege deelverzamelingen I_1 en I_2 er minstens één persoon is die een item uit I_1 goed heeft en een item uit I_2 fout heeft. Anders heet de datamatrix 'slecht geconditioneerd'.

Stel we hebben een opdeling van de items, I_1 en I_2. Dan kunnen we de personen proberen op te delen in drie groepen: P_1 bestaat uit de personen die alle items uit deelverzameling I_2 goed hebben; P_2 bestaat uit alle personen die alle items uit deelverzameling I_1 fout hebben met uitzondering van de personen die al in groep P_1 zitten; de groep personen P_3 zijn alle personen die niet in groep P_1 of P_2 zitten. Dan kunnen we door permutaties van rijen en kolommen de datamatrix altijd schrijven als

$$\boldsymbol{x} = \begin{bmatrix} [x^1] & [x^2] \\ [x^3] & [x^4] \\ [x^5] & [x^6] \end{bmatrix} = \begin{array}{c} \\ \\ P_1 \\ \\ P_2 \\ \\ P_3 \end{array} \begin{array}{c} I_1 \qquad\quad I_2 \\ \begin{bmatrix} & & 1 & \ldots & 1 \\ & [x^1] & \ldots & \ldots & \ldots \\ & & 1 & \ldots & 1 \\ 0 & \ldots & 0 & & \\ \ldots & \ldots & \ldots & [x^4] & \\ 0 & \ldots & 0 & & \\ & [x^5] & & [x^6] & \end{bmatrix} \end{array} .$$

Hierin staan de zes submatrices $[x^l], l = 1,\ldots,6$, de niet gespecificeerde matrices bevatten in principe rijen en kolommen waarin niet alleen 0 of alleen 1 staat. Fischer toont aan dat als er voor een datamatrix een opdeling van de items bestaat waarvoor de submatrices $[x^5]$ en $[x^6]$ leeg zijn, ofwel dat er voor de datamatrix voor die

opdeling van de items geen enkele persoon in groep P_3 zit, dan is de datamatrix slecht geconditioneerd. De datamatrix is goed geconditioneerd als er voor elke opdeling in de items I_1 en I_2 er op zijn minst een persoon in groep P_3 zit. Dat willen zeggen in $\left[x^5\right]$ zit minstens een rij met niet alleen 0 en in $\left[x^6\right]$ en rij met niet alleen 1. Het formele bewijs van Fischer zullen we hier niet reproduceren. Echter het idee achter het goed geconditioneerd moeten zijn van de matrix voor de schatting van de parameters en dus dat het voor een datamatrix noodzakelijk is dat de derde groep P_3 bestaat, is als volgt. Zou de derde groep niet bestaan dan kan aangetoond worden dat de aannemelijkheidsfunctie blijft stijgen als de parameterwaarden van de items in I_2 steeds groter worden; voor de items in I_1 is dat het geval als de parameterwaarden steeds kleiner worden. Er bestaan dan met andere woorden geen eindige schatters. Het bestaan van P_3 brengt de noodzakelijke verbinding in de datamatrix tot stand die dit voorkomt.

De voorwaarden voor het eindig en uniek zijn van CML-schattingen in onvolledige designs in het Raschmodel zijn hetzelfde (Fischer, 1981) met dien verstande dat de submatrices $[x^2]$ en $\left[x^3\right]$ behalve respectievelijk enen en nullen ook lege cellen mag bevatten. De lege cellen duiden dan de ontbrekende itemantwoorden aan. Op analoge wijze kan dan goed geconditioneerd zijn van de datamatrix gedefinieerd worden en kan worden aangetoond dat dit ook de voorwaarde voor het eindig en uniek zijn van de schattingen is. Fischer (1981) geeft een eenvoudige algoritme om de vervulling van deze conditie na te gaan. Tenslotte zij nog opgemerkt dat in de praktijk doorgaans aan de voorwaarden is voldaan als een anker bestaat uit een tiental niet te extreme opgaven.

Als bij het Raschmodel aan de CML-voorwaarden aan de datamatrix is voldaan dan leert de praktijk dat dan tevens aan de voorwaarden voor het bestaan van de parameterschattingen bij MML is voldaan. Hierbij moeten we echter wel bedenken dat bij CML (zie hoofdstuk 4) geen enkele aanname behoeft te worden gedaan omtrent de vaardigheid van de steekproeven leerlingen waarmee we items calibreren. Bij MML echter hebben we expliciet de aanname nodig dat de steekproef waarmee we onze items calibreren, een aselecte is uit één en dezelfde gespecificeerde verdeling, waarvan we de parameters gelijk met de itemparameters schatten. Dan wel dat we aselecte steekproeven hebben uit meerdere verdelingen, waarbij we van elke verdeling parameters schatten samen met de itemparameters (zie paragraaf 4.4). Als aan deze extra aanname is voldaan dan behoeft de verbondenheidsvoorwaarde bij MML niet meer te gelden. De verbinding kan dan worden gevonden in de equivalente groepen personen die verschillende items maken.

Over de toepasbaarheid van de CML- en de MML-schattingsmethode in de bij onvolledige designs behorende datamatrices, zoals die in paragraaf 6.2 besproken zijn,

kunnen we op basis van het bovenstaande in het algemeen het volgende concluderen. De datamatrices van het niet-verbonden design en het gedeeltelijk verbonden design kunnen niet gecalibreerd worden met de CML-methode en eventueel (met de extra aanname) wel met de MML-methode. De overige matrices komen in principe voor beide in aanmerking.

Tenslotte zij opgemerkt dat de bestaansvoorwaarden voor CML- en MML-schattingen in onvolledige designs, zoals hiervoor besproken slechts betrekking hebben op het Raschmodel. Voor uitgebreidere modellen, zoals het OPLM en voor modellen met polytome items, zijn de voorwaarden uiteraard complexer. Generaliseringen van het voorgaande voor deze modellen zijn mogelijk, maar deze zullen we niet bespreken.

6.5 Voorwaarden voor calibratie in stochastische designs

In deze paragraaf gaan we ervan uit dat aan de algemene voorwaarden uit paragraaf 6.4 is voldaan en zullen we beschrijven aan welke extra voorwaarden moet worden voldaan voor calibratie van de items in gerandomiseerde, in meerfasen en in groepsgerichte designs. We zullen daarbij opnieuw onderscheid maken tussen CML en MML als calibratie methode. In onze voorbeelden beperken we ons hierbij opnieuw tot het Raschmodel, echter de principes die besproken worden, kunnen ook op de in hoofdstuk 5 besproken uitgebreidere modellen worden toegepast.

De eerste centrale vraag die we bij alle stochastische designs moeten beantwoorden is: moeten we bij de analyse van de gegevens altijd rekening met het stochastische karakter van de designvariabele zelf of kunnen we in de analyse de designvariabele evengoed negeren, zonder dat dit gevolgen heeft voor de analyse. Voor de goede orde zij opgemerkt, dat we met het negeren van de designvariabele in de analyse bedoelen dat het stochastische karakter ervan in de analyse buiten beschouwing wordt gelaten; de informatie wie welke items heeft gemaakt kan natuurlijk nooit worden genegeerd. Het is voor te stellen dat de mogelijkheid om de designvariabele buiten de analyse te houden de analyse soms veel simpeler kan maken. Als we rekening moeten houden met de toevalsstructuur van het design, dan hebben we niet alleen de itemantwoorden X_{vi} als toevalsvariabelen, maar ook het al of niet hebben van dat antwoord. Of anders geformuleerd, als we bijvoorbeeld een aannemelijkheidsfunctie beschouwen dan kijken we bij het negeren van de designvariabele slechts naar de verdeling van alle geobserveerde itemantwoorden $\boldsymbol{X}_{obs}$, terwijl we bij het meenemen van de design-variabele de simultane verdeling van $(\boldsymbol{X}_{obs}, \boldsymbol{R})$ zullen moeten beschouwen. Door Rubin (1976) is een algemene theorie ontwikkeld met betrekking tot de analyse met

ontbrekende gegevens, waarin het eventueel negeren van de designvariabele centraal staat. Zijn begrippenkader, dat later met meer voorbeelden is uitgewerkt in Little en Rubin (1987), is in de itemresponstheorie onder meer door Mislevy en Whu (1988), Mislevy en Sheenan (1989) en door Eggen en Verhelst (1992) gehanteerd om analyse mogelijkheden in stochastische designs te beschrijven. De laatsten geven zowel voor de CML- als de MML-methode de voorwaarden voor calibratie in de drie genoemde designs.

In het hiernavolgende zullen we voornamelijk de resultaten van Eggen en Verhelst (1992) samenvatten en met voorbeelden illustreren. Alvorens dit te doen zullen we echter twee onderwerpen nog nader moeten bespreken. Het betreft allereerst het begrippenkader van Rubin (1976) en vervolgens de voor de calibratie in onvolledige designs essentiële verschillen tussen de CML- en de MML-schattingsmethode. Eerst echter een opmerking over het grote praktische belang van de mogelijkheid het design te negeren in de IRT. Belangrijk is dat in de IRT de standaardprogrammatuur die ontwikkeld is voor zowel de CML- als MML-analyse impliciet uitgaat van het negeren van de designvariabele in de analyse. Data afkomstig uit niet-negeerbare designs kunnen dus niet geanalyseerd worden met standaardprogrammatuur. In de praktijk is het echter zo dat aan de data niet te 'zien' is uit welk design ze komen. Dat wil zeggen, de programmatuur behandelt ze alsof ze uit negeerbare designs komen en levert in het geval het design niet negeerbaar is onjuiste uitkomsten. Het belang van het voldaan zijn aan de voorwaarden voor het negeren van het stochastische karakter van het design is daarom evident om foute resultaten te voorkomen.

Rubins theorie

Rubin introduceert het zogenaamde 'ignorability' principe. Dit principe wordt onder andere gedefinieerd voor statistische analyse met de grootste-aannemelijkheid ofwel ML-methode (Maximum Likelihood). Omdat de calibratie van items, en trouwens ook het schatten van persoonsparameters, in IRT plaatsvindt met deze methode zullen we de voorwaarden voor correct toepassen van dit principe hiertoe beperken. Dit principe houdt in dat we ons voor de analyse van gegevens kunnen beperken tot slechts de resultaten op waargenomen variabelen, zonder dat we in de procedure ook informatie over het design moeten meenemen. Het design wordt genegeerd. In het algemeen beschouwen we in een analyse een vector toevalsvariabele $\boldsymbol{U} = (U_1,...,U_m)$ met verdeling $f_\tau(\boldsymbol{u})$. De parametervector τ bevat de parameters die we willen schatten. Om de gedachten te bepalen is het voor te stellen dat $m = n.k$, met k het aantal

variabelen en n het aantal personen dat in de analyse wordt beschouwd. Als er ontbrekende gegevens zijn, definiëren we een 'missing data indicator' $\boldsymbol{M} = (M_1,...,M_m)$, die aangeeft of een variabele U_j daadwerkelijk geobserveerd is, $m_j = 1$, of niet, $m_j = 0$. Dus $\boldsymbol{M}$ is op dezelfde wijze gedefinieerd als de itemindicator variabele $\boldsymbol{R}$ in paragraaf 6.3. $\boldsymbol{M}$ wordt echter, zoals verderop duidelijk zal worden algemener gebruikt dan alleen als itemindicator $\boldsymbol{R}$. De missing data indicator partitioneert $\boldsymbol{U}$ en zijn geobserveerde waarde $\boldsymbol{u}$ in

$$\boldsymbol{U} = (\boldsymbol{U}_{obs}, \boldsymbol{U}_{mis}) \text{ en } \boldsymbol{u} = (\boldsymbol{u}_{obs}, \boldsymbol{u}_{mis}). \tag{6.6}$$

De verzameling *obs* bevat de indexen van waargenomen variabelen, dat wil zeggen, elke j waarvoor $m_j = 1$, en *mis* is de verzameling van indexen van de niet waargenomen variabelen $(m_j = 0)$. $\boldsymbol{U}_{obs}$ en $\boldsymbol{u}_{obs}$ zijn respectievelijk de toevalsvariabele en de realisatie van de waargenomen variabelen. $\boldsymbol{U}_{mis}$ de toevalsvariabele en $\boldsymbol{u}_{mis}$ de waarden die we geobserveerd zouden hebben, als we dat gewild of gekund hadden, van de niet waargenomen variabelen. In een analyse met de grootste-aannemelijkheidsmethode zouden we ons moeten baseren op de gezamenlijke verdeling $g_{\tau,\phi}$ van alle waargenomen toevalsvariabelen, dat wil zeggen van $\boldsymbol{U}_{obs}$ en $\boldsymbol{M}$:

$$g_{\tau,\phi}(\boldsymbol{u}_{obs}, \boldsymbol{m}) = \int_{\boldsymbol{u}_{mis}} g_{\tau,\phi}(\boldsymbol{u}_{obs}, \boldsymbol{u}_{mis}, \boldsymbol{m}) \quad d\boldsymbol{u}_{mis}. \tag{6.7}$$

We merken op dat we in het hoofdstuk een uitdrukking als (6.7) zowel voor een verdeling van toevalsvariabele gebruiken als voor een aannemelijkheidsfunctie, zonder dat laatste expliciet als functie van de parameter(s) te schrijven. In (6.7) staat ϕ voor een mogelijke parameter van de verdeling van de missing data indicator $\boldsymbol{M}$. Bij n personen en experimentele onafhankelijkheid (zie hoofdstuk 4) is dit ook te schrijven als:

$$\int_{\boldsymbol{u}_{mis}} g_{\tau,\phi}(\boldsymbol{u}_{obs}, \boldsymbol{u}_{mis}, \boldsymbol{m})\, d\boldsymbol{u}_{mis} = \prod_{v=1}^{n} \int_{\boldsymbol{u}_{mis,v}} g_{\tau,\phi}(\boldsymbol{u}_{obs,v}, \boldsymbol{u}_{mis,v}, \boldsymbol{m}_v)\, d\boldsymbol{u}_{mis,v}. \tag{6.8}$$

We zien dat (6.8) zowel afhangt van de verdeling van $\boldsymbol{M}$, met parameter ϕ, als van de variabele $\boldsymbol{U}$, met parameter τ, waarin we geïnteresseerd zijn. Als we in plaats van (6.8)

$$\int_{\boldsymbol{u}_{mis}} f_\tau(\boldsymbol{u})\, d\boldsymbol{u}_{mis} \;=$$

$$\int_{\boldsymbol{u}_{mis}} f_\tau(\boldsymbol{u}_{obs}, \boldsymbol{u}_{mis})\, d\boldsymbol{u}_{mis} \;= \tag{6.9}$$

$$\prod_{v=1}^{n} \int_{\boldsymbol{u}_{mis,v}} f_\tau(\boldsymbol{u}_{obs,v}, \boldsymbol{u}_{mis,v})\, d\boldsymbol{u}_{mis,v}\,,$$

zouden toepassen, dan negeren we de designvariabele in de analyse. We hebben dan een eenvoudiger uitdrukking die alleen afhangt van de verdeling van de variabelen die ons interesseren, met parameter τ. Als het geoorloofd is, dat wil zeggen niet tot fouten leidt, (6.9) in plaats van (6.8) in de analyse toe te passen dan geldt het 'ignorability' principe. Zonder fouten te maken nemen we dan aan dat de observaties van $\boldsymbol{U}$ uit de marginale verdeling van alleen de waargenomen variabelen $\boldsymbol{U}_{obs}$ komen en we negeren de designvariabele. De rechtvaardiging hiervan hangt af van de eigenschappen die de verdeling van de missing data indicator heeft, of zoals Rubin het noemt: van de eigenschappen van "the proces that causes missing data". Dit proces wordt door Rubin beschreven met de voorwaardelijke verdeling van de missing data indicator gegeven de data: $h_\phi(\boldsymbol{m} \mid \boldsymbol{u})$. Als voor deze verdeling de eigenschap geldt dat

$$h_\phi(\boldsymbol{m} \mid \boldsymbol{u}_{obs}, \boldsymbol{u}_{mis}) = h_\phi(\boldsymbol{m} \mid \boldsymbol{u}_{obs}) \text{ voor alle } \boldsymbol{u}_{mis}, \tag{6.10}$$

dan is het gerechtvaardigd het design in de ML-analyse te negeren. Ofwel de kansen op het ontbreken van de gegevens hangen niet af van de waarden van de gegevens die niet zijn waargenomen, maar hangen mogelijkerwijs uitsluitend af van wel waargenomen gegevens. Rubin noemt de situatie waarin dit geldt MAR, 'missing at random'. We tonen nu aan dat als aan de MAR-voorwaarde (6.10) voldaan is, we in de ML-analyse evengoed uit kunnen gaan van de eenvoudiger verdeling (6.9) als van (6.8). Het rechterlid van (6.8) kunnen we in het algemeen herschrijven als:

$$\prod_{v=1}^{n} \int_{\boldsymbol{u}_{mis,v}} g_{\tau,\phi}(\boldsymbol{u}_{obs,v}, \boldsymbol{u}_{mis,v}, \boldsymbol{m}_v)\, d\boldsymbol{u}_{mis,v} =$$

(6.11)

$$\prod_{v=1}^{n} \int_{\boldsymbol{u}_{mis,v}} h_\phi(\boldsymbol{m}_v \mid \boldsymbol{u}_{obs,v}, \boldsymbol{u}_{mis,v}) \cdot f_\tau(\boldsymbol{u}_{obs,v}, \boldsymbol{u}_{mis,v})\, d\boldsymbol{u}_{mis,v} =$$

$$\prod_{v=1}^{n} \int_{\boldsymbol{u}_{mis,v}} h_\phi(\boldsymbol{m}_v \mid \boldsymbol{u}_{obs,v}) \cdot f_\tau(\boldsymbol{u}_{obs,v}, \boldsymbol{u}_{mis,v})\, d\boldsymbol{u}_{mis,v} =$$

$$\prod_{v=1}^{n} h_\phi(\boldsymbol{m}_v \mid \boldsymbol{u}_{obs,v}) \cdot \prod_{v=1}^{n} \int_{\boldsymbol{u}_{mis,v}} f_\tau(\boldsymbol{u}_{obs,v}, \boldsymbol{u}_{mis,v})\, d\boldsymbol{u}_{mis,v}$$

In (6.11) maken we in de eerste gelijkheid gebruik van de eigenschappen van voorwaardelijke kansen: de gezamenlijke verdeling $g_{\tau,\phi}$ wordt geschreven als het produkt van de voor-waardelijke verdeling h_ϕ van de missing data indicator en de verdeling van dat deel waarop geconditioneerd wordt. Deze laatste verdeling is de verdeling f_τ van de variabelen $\boldsymbol{u} = (\boldsymbol{u}_{obs}, \boldsymbol{u}_{mis})$. In de volgende gelijkheid wordt gebruik gemaakt van de MAR-eigenschap (6.10) van de verdeling van de designvariabele. Omdat tenslotte $h_\phi(\boldsymbol{m}_v \mid \boldsymbol{u}_{obs,v})$ onafhankelijk is van $\boldsymbol{u}_{mis,v}$ kan deze term buiten de integraal worden gehaald. Het resultaat is dat de aannemelijkheidsfunctie (6.8) uiteenvalt in twee termen, waarvan de tweede term gelijk is aan de eenvoudigere aannemelijkheidsfunctie (6.9) en een eerste term die onafhankelijk is van de parameter τ waarnaar we de aannemelijkheidsfunctie moeten maxi-maliseren. Het zal duidelijk zijn dat we bij het maximaliseren naar τ deze eerste term evengoed kunnen weglaten. Voor de goede orde zij vermeld dat naast de MAR-voorwaarde ook nog voldaan moet zijn aan een voorwaarde, die betrekking heeft op de mogelijke waarden die de te schatten parameters τ en eventuele parameters ϕ van de verdeling van de missing data indicator kunnen aannemen. Aangezien aan deze voorwaarde in onze toepassing altijd voldaan is, zullen we hieraan geen aandacht besteden. Aldus hebben we gezien dat het voldoen aan de MAR-voorwaarde voldoende is voor het negeren van het design in de analyse.

Soms geldt dat de ontbrekende gegevens MCAR, 'missing completely at random', zijn, hetgeen betekent dat

$$h_\phi(\boldsymbol{m} \mid \boldsymbol{u}_{obs}, \boldsymbol{u}_{mis}) = h_\phi(\boldsymbol{m}) \text{ voor alle } \boldsymbol{u}_{mis} \text{ en } \boldsymbol{u}_{mis}. \tag{6.12}$$

Dat wil zeggen de kans op het ontbreken van gegevens hangt noch van de waargenomen noch de niet waargenomen gegevens af. Het zal duidelijk zijn dat als aan de sterkere MCAR voorwaarde is voldaan automatisch ook voldaan aan de MAR-voorwaarde.

Verschil designvariabele bij CML en MML

In de analyse in onvolledige designs verschillen de CML- en MML-schattingsmethode op een essentieel punt van elkaar. De reden voor het onderscheid tussen CML en MML is dat in genoemde designs het mechanisme dat verantwoordelijk voor het ontbreken van gegevens een toevalsproces is en dat bij de calibratie met de CML- en MML-methode er in principe uitgegaan wordt van een verschillend toevalsproces dat de itemantwoorden genereert. Bij CML worden alleen de itemantwoorden $X_{vi}, v = 1,...,n; i = 1,...,k$ als toevalsvariabelen beschouwd, terwijl bij MML naast deze itemantwoorden ook de vaardigheden van de personen die items maken $\theta_v, v = 1,...,n$ expliciet als toevalsvariabelen worden beschouwd. De consequentie hiervan is dat de algemene missing data indicator voor een persoon $\boldsymbol{M}_v$ bij MML altijd één element meer bevat dan bij CML. Als de totale itemverzameling bijvoorbeeld vijf items bevat, waarvan een bepaald persoon v, volgens een of ander stochastisch design uit paragraaf 6.3, het eerste, het derde en het vierde item wel maakt en de andere twee items niet dan heeft de missing data indicator bij een CML-analyse dezelfde waarde als de itemindicator $\boldsymbol{m}_v = \boldsymbol{r}_v = (1,0,1,1,0)$. In de MML-analyse daarentegen is $\boldsymbol{m}_v = (\boldsymbol{r}_v, 0) = (1,0,1,1,0,0)$, waarin de laatste 0 het niet waarnemen van de variabele θ_v indiceert.

In Eggen en Verhelst (1992) is uiteengezet, dat Rubins voorwaarden voor het negeren van de designvariabelen in de analyse bij de MML-methode onverkort toepasbaar zijn. Het controleren van Rubins voorwaarden geeft uitsluitsel over de mogelijkheid de designvariabele te negeren in de analyse. In paragraaf 6.5.1, zullen we dit voor de stochastische designs uit paragraaf 6.3 bespreken. Bij CML blijken Rubins voorwaarden niet beslissend te zijn. De mogelijkheid van toepassing van CML in stochastische designs blijkt in de eerste plaats af te hangen van dat deel van de aannemelijkheidsfunctie dat we in de CML-analyse buiten beschouwing laten. In paragraaf 6.5.2 zullen we dat uitwerken. In deze paragrafen zullen wij als in hoofdstuk 4 een deel van de uitwerkingen alleen geven voor het Raschmodel, de principes zijn echter evenzeer toepasbaar voor de uitgebreidere modellen die in hoofdstuk 5 zijn behandeld. De verdeling van de itemantwoorden, ook als we deze als aannemelijkheids-functie beschouwen, zullen we daarbij steeds aangeven met $P_{\cdot\cdot}(...)$.

6.5.1 MML in stochastische designs

Aansluitend bij de notatie in hoofdstuk 4 en uit de vorige paragraaf hebben we in een MML-analyse te maken met de toevalsvariabele

$$\boldsymbol{U} = (\boldsymbol{X}, \theta) = (\boldsymbol{X}_1, \theta_1, ..., \boldsymbol{X}_n, \theta_n). \tag{6.13}$$

Met θ_v de vaardigheid van persoon v, $v = 1,...,n$ en $\boldsymbol{X}_v = (X_{v1},..., X_{vk})$ de antwoorden van deze personen op de k items, die eventueel niet allemaal zijn geobserveerd. De parametervector die we willen schatten is $\tau = (\beta, \mu, \sigma^2)$, met $\beta = (\beta_1,..., \beta_k)$ de vector van alle k moeilijkheidsparameters en μ en σ^2, respectievelijk het gemiddelde en de variantie van de normale vaardigheidsverdeling $g_{\mu,\sigma^2}(\theta)$ (zie formule 4.55).

MML in gerandomiseerde onvolledige designs.

In deze designs is de verdeling van de missing data indicator gelijk aan de verdeling van de itemindicator (zie (6.2)), omdat de vaardigheid θ_v nooit wordt waargenomen geldt:

$$P(\boldsymbol{M}_v = (\boldsymbol{r}_b, 0)) = P(\boldsymbol{R}_v = \boldsymbol{r}_b) = \phi_b. \tag{6.14}$$

Hierin is $\boldsymbol{r}_b$ zoals eerder de vector met lengte k met een 1 op de plaatsen die de items indiceren die in boekje b zitten en een 0 op de overige plaatsen. Deze formule geldt uiteraard weer voor alle personen: $v = 1,...,n$ en alle boekjes $b = 1,...,B$.

Als we kijken waarin de totale verzameling van toevalsvariabelen $\boldsymbol{U}$ (6.13) uiteenvalt door de missing data indicator $\boldsymbol{M}_v$ volgens (6.6), dan is eenvoudig na te gaan dat in dit geval voor elke persoon v geldt:

$$\left.\begin{aligned} \boldsymbol{U}_{obs,v} &= \boldsymbol{X}_{obs,v} \\ \boldsymbol{U}_{mis,v} &= (\boldsymbol{X}_{mis,v}, \theta_v) \end{aligned}\right\}, \quad (v = 1,...,n). \tag{6.15}$$

In (6.14) zien we dat de verdeling van de missing data indicator noch van de waarden van de niet waargenomen data noch van de waargenomen data afhangt. De ontbrekende data zijn in gerandomiseerde designs dus MCAR, formule (6.12) is geldig, en duidelijk is dat aan Rubins voorwaarden voor het negeren van het design is voldaan. Het bewijs hiervan, een toepassing van (6.11) laten we aan de lezer over. We kunnen dus de marginale verdeling van de observaties $\boldsymbol{X}_{obs}$ als basis voor de analyse

gebruiken. De aannemelijkheidsfunctie wordt dan gegeven door het in (6.9) invullen van de specificatie (6.15):

$$\prod_{v} \int_{x_{mis,v}} \int_{\theta_v} f_\tau(\boldsymbol{x}_{obs,v}, \boldsymbol{x}_{mis,v}, \theta_v)\, d\theta_v\, d\boldsymbol{x}_{mis,v} =$$

$$\prod_{v} \int_{x_{mis,v}} \int_{\theta_v} P_\beta(\boldsymbol{x}_{obs,v}, \boldsymbol{x}_{mis,v} \mid \theta_v).g_{\mu,\sigma^2}(\theta_v)\, d\theta_v\, d\boldsymbol{x}_{mis,v} = \tag{6.16}$$

$$\prod_{v} \int_{\theta_v} P_\beta(\boldsymbol{x}_{obs,v} \mid \theta_v).g_{\mu,\sigma^2}(\theta_v)\, d\theta_v \ .$$

In (6.16) volgt de eerste gelijkheid uit de eigenschappen van voorwaardelijke kansen, zoals we die eerder bij de afleiding van de marginale aannemelijkheidsfunctie, formule (4.49), zagen. De tweede gelijkheid volgt uit de lokale stochastische onafhankelijkheid van de itemantwoorden en het uitintegreren van $\boldsymbol{x}_{mis,v}, v = 1,...,n$. De aannemelijkheidsfunctie (6.16) lijkt uiteindelijke dus zeer veel op de marginale aannemelijkheidsfunctie voor volledige gegevens (formule 4.57). Het verschil zit er slechts in dat per persoon v slechts de kansen op de waargenomen responsen worden meegenomen en dat per persoon alleen de itemparameters van de waargenomen items in de aannemelijkheidsfunctie meedoen. De relatie met de volledige data MML-analyse wordt duidelijk gemaakt als we met n_b het aantal personen noteren dat boekje b maakt, dan geldt dat $\sum_{b=1}^{B} n_b = n$, het totaal aantal personen. Als we verder $\beta_{(b)}$ definiëren als de k_b-vector van de itemparameters van de items in boekje b, dan kunnen we (6.16) herschrijven als

$$\prod_{v=1}^{n} \int_{\theta_v} P_\beta(\boldsymbol{x}_{obs,v} \mid \theta_v).g_{\mu,\sigma^2}(\theta_v)\, d\theta_v =$$

$$\prod_{b=1}^{B} \prod_{v=1}^{n_b} \int_{\theta_v} P_{\beta_{(b)}}(\boldsymbol{x}_{obs,v} \mid \theta_v).g_{\mu,\sigma^2}(\theta_v)\, d\theta_v . \tag{6.17}$$

We zien in (6.17) dus dat we de marginale aannemelijkheidsfunctie in onvolledige designs kunnen schrijven als een produkt van B marginale aannemelijkheidsfuncties, evenveel als er verschillende toetsboekjes zijn, voor volledige gegevens. Vergelijk formule (4.113).

MML in meerfasen onvolledige designs

In meerfasen designs is de opdeling door de missing data indicator in geobserveerde en niet geobserveerde variabelen hetzelfde als bij gerandomiseerde designs (zie (6.15)). De verdeling van de missing data indicator volgt op dezelfde wijze als bij gerandomiseerde designs nu echter met de itemindicator van meerfasen designs (6.4) als basis:

$$P(\boldsymbol{M}_v = (\boldsymbol{r}_b, 0) \mid \boldsymbol{x}_{obs,v}) = P(\boldsymbol{R}_v = \boldsymbol{r}_b \mid \boldsymbol{x}_{obs,v}) = 0 \quad \text{of} \quad 1. \tag{6.18}$$

Formule (6.18) geldt voor elke persoon $v = 1,...,n$ en elk boekje $b = 1,...,B$. Eenvoudig is in te zien dat de verdeling van de missing data indicator voldoet aan de voorwaarde (6.10), dat wil zeggen de missing data zijn MAR. De designverdeling hangt immers alleen af van de geobserveerde waarden en niet van de niet geobserveerde. Volgens het ignorability principe is het dus gerechtvaardigd het design in de analyse te negeren. De algemene uitdrukking voor de marginale aannemelijkheidsfunctie is in dit geval identiek aan de marginale aannemelijk-heidsfunctie bij gerandomiseerde designs (6.16) of (6.17).

In paragraaf 6.5.2 zullen we in tabel 6.6. een voorbeeld van een MML-analyse in een meerfasen design geven en de resultaten vergelijken met een CML-analyse.

MML in groepsgerichte designs

In groepsgerichte calibratiedesigns hebben we in paragraaf 6.3.3 twee situaties onderscheiden. In de eerste hebben wij een achtergrondvariabele Y die slechts een rol speelt in de toewijzing van boekjes aan leerlingen en zijn we niet geïnteresseerd in de verschillende vaardigheids-verdelingen. In de tweede zijn we behalve in de itemparameters ook geïnteresseerd in de parameters van de in totaal B vaardigheidsverdelingen voor de verschillende niveaus van de achtergrondvariabele: we kunnen B subpopulaties onderscheiden in de totale populatie. In de tweede situatie zullen we in de praktijk vaak niet één aselecte steekproef uit een vaardigheids-verdeling ter beschikking hebben, maar, een bewust op die wijze getrokken gestratificeerde steekproef, bestaande uit aselecte steekproeven uit de vaardigheidsverdelingen voor elk onderscheiden niveau van de achtergrondvariabele.

Hetzelfde mogelijke onderscheid in subpopulaties speelt ook al een rol bij de MML-analyse in volledige designs. Bij een gestratificeerde steekproef zullen we daar, samen

met de itemparameters, de parameters van meer vaardigheidsverdelingen moeten schatten. Als we dat niet zouden doen, en de steekproef beschouwen als een aselecte uit één populatie, dan maken we een specificatiefout welke tot onjuiste schattingen leidt. Aangezien de situatie van volledige designs een bijzonder geval van is groepsgerichte designs, zullen we hieraan verder geen expliciet aandacht besteden.

Mislevy en Sheenan (1989) hebben aangetoond dat het voor de behandeling van de designvariabele in groepsgerichte designs in een MML-analyse niet uitmaakt of we nu een aselecte steekproef hebben uit één populatie of een gestratificeerde. Vandaar dat we er in deze paragraaf van uit zullen gaan dat we een aselecte steekproef hebben uit één vaardigheids-verdeling, die kan worden geschreven als een combinatie van B verdelingen, voor elke subpopulatie geassocieerd met een onderscheiden niveau van de achtergrondvariabele Y:

$$\begin{aligned} g_{\mu,\sigma^2}(\theta) &= \sum_{b=1}^{B} P(\theta, Y = y_b) = \sum_{b=1}^{B} P(\theta \mid Y = y_b).\ P(Y = y_b) \\ &= \sum_{b=1}^{B} g_{\mu_b,\sigma_b^2}(\theta).\pi_b \quad . \end{aligned} \tag{6.19}$$

In (6.19) zijn μ_b en σ_b^2 het gemiddelde en de variantie van de vaardigheidsverdeling verdeling in subpopulatie b en π_b de proportie personen in subpopulatie b in de totale populatie.

In groepsgerichte designs is de verdeling van de itemindicator gegeven in (6.5), waaruit met (6.19) volgt dat

$$P(\boldsymbol{R}_v = \boldsymbol{r}_b) = P(Y_v = y_b) = \pi_b.$$

Hetgeen uiteraard weer geldt voor alle personen $v = 1,...,n$ en alle boekjes of onderscheiden niveaus $b = 1,...,B$ van de achtergrondvariabele. Omdat de vaardigheid θ_v nooit geobserveerd wordt komt de vraag of we in deze designs de designvariabele kunnen negeren neer op de vraag of we in de analyse de achtergrondvariabele Y kunnen negeren ofwel moeten meenemen. Het antwoord op deze vraag kunnen we weer geven door de voorwaarden van Rubin te controleren.

In de MML-analyse zijn in dit geval de toevalsvariabelen die een rol zouden kunnen spelen $\boldsymbol{U} = (\boldsymbol{X}, \boldsymbol{Y}, \theta)$, met voor elke persoon de vector $\boldsymbol{X}_v$ met antwoorden op de k items, de waarde van de achtergrondvariabele Y_v en de vaardigheid θ_v. Als we de

achtergrond-informatie in de analyse meenemen dan wordt de opdeling van $\boldsymbol{U}$ door de missing data indicator $\boldsymbol{M}_v$ gegeven door

$$\left.\begin{aligned} \boldsymbol{U}_{obs,v} &= (\boldsymbol{X}_{obs,v}, Y_v) \\ \boldsymbol{U}_{mis,v} &= (\boldsymbol{X}_{mis,v}, \theta_v) \end{aligned}\right\}, \quad (v = 1,...,n). \tag{6.20}$$

En de verdeling van $\boldsymbol{M}_v$ door

$$P(\boldsymbol{M}_v = (\boldsymbol{r}_b, 1, 0)) = P(\boldsymbol{R}_v = \boldsymbol{r}_b) = P(Y_v = y_b),$$

ofwel

$$\left.\begin{aligned} P(\boldsymbol{M}_v = (\boldsymbol{r}_b, 1, 0) \mid Y_v = y_b) &= 1 \\ P(\boldsymbol{M}_v = (\boldsymbol{r}_b, 1, 0) \mid Y_v \neq y_b) &= 0 \end{aligned}\right\}, \quad b = 1,..., B; v = 1,...,n. \tag{6.21}$$

Waarbij de waarde 1 van het voorlaatste element van $\boldsymbol{M}_v$ aanduidt dat Y_v als waargenomen wordt beschouwd en het laatste element het niet waarnemen van θ_v indiceert. Uit (6.21) is eenvoudig te zien dat bij het meenemen van de achtergrondvariabele aan de MAR-voorwaarde (6.10) is voldaan: de verdeling van de missing data indicator hangt alleen af van geobserveerde waarden, en in de analyse kunnen we de designvariabele als geheel negeren en de marginale verdeling van alleen de geobserveerde waarden (6.9) hoeven we te beschouwen. Als we de kans beschouwen dat een aselect getrokken persoon uit de populatie een bepaald antwoordpatroon heeft in boekje b, dan kunnen we met de eerdere notatie (formule (6.17)) hiervoor schrijven:

$$\begin{aligned} &P_{\beta_{(b)}, \mu_b, \sigma_b^2, \pi_b}(\boldsymbol{x}_{obs,v}, Y_v = y_b) = \\ &\int_{\boldsymbol{x}_{mis,v}} \int_{\theta_v} P_{\beta_{(b)}, \mu_b, \sigma_b^2, \pi_b}(\boldsymbol{x}_{obs,v}, \boldsymbol{x}_{mis,v}, Y_v = y_b, \theta_v)\ d\theta_v\ d\boldsymbol{x}_{mis,v} = \\ &\int_{\theta_v} P_{\beta_{(b)}}(\boldsymbol{x}_{obs,v} \mid Y_v = y_b, \theta_v).\ P_{\mu_b, \sigma_b^2}(\theta_v \mid Y_v = y_b).P_{\pi_b}(Y_v = y_b)\ d\theta_v = \\ &\pi_b . \int_{\theta_v} P_{\beta_{(b)}}(\boldsymbol{x}_{obs,v} \mid \theta_v).g_{\mu_b, \sigma_b^2}(\sigma_v)\ d\theta_v . \end{aligned} \tag{6.22}$$

De tweede gelijkheid in (6.22) volgt uit de eigenschappen van voorwaardelijke kansen, terwijl in de derde gebruik gemaakt wordt van de lokale stochastische onafhankelijkheid in IRT-modellen. Bij n_b personen die boekje b maken wordt de marginale aannemelijkheidsfunctie gegeven door:

$$\prod_{b=1}^{B} \pi_b^{n_b} . \prod_{b=1}^{B} \prod_{v=1}^{n_b} \int_{\theta_v} P_{\beta_{(b)}}(\boldsymbol{x}_{obs,v} \mid \theta_v) . g_{\mu_b, \sigma_b^2}(\theta_v)\, d\,\theta_v . \tag{6.23}$$

We zien dat (6.23) uiteenvalt in een deel dat alleen afhangt van de trekkingskansen π_b, dat een persoon uit subpopulatie b komt en een deel dat het produkt is van in totaal B deels overlappende marginale aannemelijkheidsfuncties als (4.57). Voor de schatting van de parameters kunnen we deze functie maximaliseren naar β, μ_b, σ_b^2 en eventueel π_b, voor $b = 1,...,B$. De ML-schatter van π_b is gegeven door: $\hat{\pi}_b = n_b/n$.

Als we in groepsgerichte designs de achtergrondvariabele Y_i niet zouden meenemen dan wordt de opdeling van $\boldsymbol{U}$ gegeven door (vergelijk met (6.20))

$$\left.\begin{aligned} \boldsymbol{U}_{obs,v} &= \boldsymbol{X}_{obs,v} \\ \boldsymbol{U}_{mis,v} &= (\boldsymbol{X}_{mis,v}, Y_v, \theta_v) \end{aligned}\right\}, \quad (v = 1,...,n).$$

Immers Y_v beschouwen we dan als niet waargenomen gegevens. De verdeling van $\boldsymbol{M}_v$ is dan (vergelijk met (6.21)):

$$\left.\begin{aligned} P(\boldsymbol{M}_v = (\boldsymbol{r}_b, 0, 0) \mid Y_v = y_b) &= 1 \\ P(\boldsymbol{M}_v = (\boldsymbol{r}_b, 0, 0) \mid Y_v \neq y_b) &= 0 \end{aligned}\right\}, \quad b = 1,...,B \ , \ v = 1,...,n. \tag{6.24}$$

Het voorlaatste element is nu 0, omdat Y_v als niet waargenomen wordt beschouwd. Aan (6.24) is eenvoudig in te zien dat in dit geval niet voldaan is aan de MAR-voorwaarde (6.10) om de designvariabele te negeren, immers de verdeling van de missing data indicator hangt af van niet-waargenomen variabelen. In groepsgerichte designs zijn we dus verplicht de achtergrondvariabele mee te nemen in de analyse. Zouden we dat niet doen dan geeft een MML-analyse wel uitkomsten, deze zijn echter onjuist. Met een voorbeeld zullen wij dit illustreren.

We genereren onder het Raschmodel itemantwoorden voor twee groepen van 500 leerlingen. De eerste groep van 500 minder vaardige personen, met waarde y_1 van de achtergrond-variabele, is aselect getrokken uit een normale verdeling met gemiddelde -1 en variantie 1, $N(-1,1)$. De tweede vaardiger groep, met de waarde y_2, is aselect getrokken uit $N(1,1)$. Voor de eerste groep worden itemantwoorden op vijf items die

gemakkelijk zijn ($\beta_i = -2, i = 1,...,5$) en vijf middelmatig moeilijke items ($\beta_i = 0, i = 6,...,10$) gegenereerd. De tweede groep maakt naast de middelmatig moeilijke items 6 tot en met 10, vijf items moeilijke items met $\beta_i = 2, i = 11,\ldots,15$. Voor de aldus gegenereerde antwoorden voeren we twee MML-analyses uit: in de eerste negeren we de achtergrond-variabele, in de tweede nemen we de achtergrondvariabele mee in de analyse. Het resultaat, waarbij de normering zodanig is gekozen dat $\sum_{i=1}^{15} \hat{\beta}_i = 0$, staat in tabel 6.4. We zien in tabel 6.4 dat het niet meenemen van de achtergrondvariabele in groepsgerichte designs systematisch verkeerde schattingen van de itemparameters oplevert. De gemakkelijke items 1 tot en met 5 worden moeilijker geschat dan ze in werkelijkheid zijn. Van de moeilijke items 11 tot en 15 worden itemparameter onderschat. Ook de parameters van de vaardigheids-verdeling, zie onder in de tabel, worden als gevolg van de gemaakte specificatiefout verkeerd geschat. Zoals in tabel 6.4 te zien zijn de afwijkingen van de ingevoerde parameters doorgaans meer dan 2 standaardfouten. Als we de achtergrondinformatie wel meenemen zien we dat zowel de itemparameters als de parameters van de vaardigheidsverdelingen, rekening houdend met de standaardfouten naar verwachting worden teruggeschat.

Tabel 6.4
MML-analyse gesimuleerd groepsgericht design

		negeren y_b		meenemen y_b	
item	β_i	$\hat{\beta}_i$	$SE(\hat{\beta}_i)$	$\hat{\beta}_i$	$SE(\hat{\beta}_i)$
1	-2	-1.847	.127	-2.158	.113
2	-2	-1.786	.127	-2.099	.112
3	-2	-1.726	.126	-2.042	.111
4	-2	-1.761	.126	-2.076	.112
5	-2	-1.679	.125	-1.996	.110
6	0	0.018	.074	0.006	.076
7	0	-0.003	.074	-0.016	.076
8	0	-0.036	.074	-0.050	.076
9	0	0.018	.074	0.006	.076
10	0	0.018	.074	0.006	.076
11	2	1.706	.125	2.035	.111
12	2	1.753	.126	2.080	.112
13	2	1.813	.127	2.139	.113
14	2	1.637	.125	1.967	.110
15	2	1.874	.127	2.198	.114
		$\hat{\mu}$ = 0.018(.083)	$\hat{\sigma}$ = 1.326(.053)	$\hat{\mu}_1$ = -0.984(.061)	$\hat{\sigma}_1$ = 0.980(.049)
				$\hat{\mu}_2$ = 1.018(.065)	$\hat{\sigma}_2$ = 1.062(.050)

Bij groepsgerichte designs moeten we dus in een MML-analyse de achtergrondvariabele meenemen en tegelijk met de itemparameters de verdelingsparameters van alle groepen meeschatten. Omdat standaardprogrammatuur voor MML, zoals BILOG (Mislevy & Bock, 1986), deze optie niet kent en suggereert dat het geen rol speelt moet men in de praktijk hiervoor op zijn hoede zijn.

6.5.2 CML in stochastische designs

In paragraaf 6.5 werd reeds opgemerkt dat Rubins voorwaarden niet beslissend zijn voor het eventueel negeren van de designvariabele in de CML-analyse. Alvorens de

mogelijkheden voor CML-analyse in de drie stochastische designvormen te bespreken, zullen we de reden hiervoor uiteenzetten en de voor CML beslissende voorwaarden formuleren.

Stel dat we gebruik zouden willen maken van Rubins 'ignorability' principe in een CML-analyse. Dan analyseren we uiteindelijk de marginale verdeling van de geobserveerde itemantwoorden (zie (6.9)):

$$\prod_{v=1}^{n} \int_{\boldsymbol{u}_{mis,v}} f_\tau(\boldsymbol{u}_{obs,v}, \boldsymbol{u}_{mis,v})\, d\boldsymbol{u}_{mis,v} = \prod_{v=1}^{n} P_{\beta,\theta_v}(\boldsymbol{x}_{obs,v}).$$

De verdeling van het geobserveerde antwoordpatroon $\boldsymbol{X}_{obs,v}$ hangt hierin af van de moeilijkheidsparameters β en de individuele vaardigheidsparameter θ_v, die bij CML in tegenstelling tot bij MML niet als toevalsvariabele wordt beschouwd. Om de CML-methode te kunnen toepassen zou er voor elke persoon v een voldoende steekproefgrootheid of statistiek $S_{obs,v} = S_{obs,v}(\boldsymbol{X}_{obs,v})$ moeten bestaan voor θ_v waarop we dan zouden kunnen conditioneren, zodat de aannemelijkheidsfunctie onafhankelijk van θ_v wordt. In onvolledige designs bestaat zo'n voldoende statistiek echter niet in de verdeling van $\boldsymbol{X}_{obs,v}$, hetgeen we nu aan de hand van een voorbeeld zullen illustreren.

Stel we hebben drie items die het Raschmodel volgen en we hebben een gerandomiseerd design met twee boekjes, bestaande uit respectievelijk item 1 en 2, en item 1 en 3. De verdeling van de itemindicator wordt gegeven door

$$P(\boldsymbol{R} = \boldsymbol{r}_1 = (1,1,0) = \phi, \quad \text{en} \quad P(\boldsymbol{R} = \boldsymbol{r}_2 = (1,0,1)) = 1 - \phi.$$

In het Raschmodel verwachten we, zie hoofdstuk 4, dat de somscore op de geobserveerde items

$$S_{obs,v} = \sum_{j \in obs,v} X_{vj}, \tag{6.25}$$

voldoende zal zijn voor θ_v en dat dus door conditioneren hierop er per persoon een voorwaardelijke kans geldt die alleen afhangt van de itemparameters. De somscore (6.25) is echter niet voldoende in de verdeling van $\boldsymbol{X}_{obs,v}$.

Merk allereerst op dat in het voorbeeld dat we bespreken de verdeling van $\boldsymbol{X}_{obs,v}$ en de verdeling van alle toevalsvariabelen $(\boldsymbol{X}_{obs,v}, \boldsymbol{R}_v)$ exact gelijk zijn. Er geldt namelijk altijd dat

$$P(\boldsymbol{x}_{obs,v}) = P(\boldsymbol{x}_{obs,v} \mid \boldsymbol{R}_v = \boldsymbol{r}_1) \,.\, P(\boldsymbol{R}_v = \boldsymbol{r}_1) + P(\boldsymbol{x}_{obs,v} \mid \boldsymbol{R}_v = \boldsymbol{r}_2) \,.\, P(\boldsymbol{R}_v = \boldsymbol{r}_2). \tag{6.26}$$

En voor de verdeling van $(\boldsymbol{X}_{obs,v}, \boldsymbol{R}_v)$ geldt

$$P(\boldsymbol{x}_{obs,v}, \boldsymbol{R}_v = \boldsymbol{r}_b) = P(\boldsymbol{x}_{obs,v} \mid \boldsymbol{R}_v = \boldsymbol{r}_b) \,.\, P(\boldsymbol{R}_v = \boldsymbol{r}_b) \text{ voor } b = 1,2. \qquad (6.27)$$

Als we nu kijken naar de mogelijke waarden van $\boldsymbol{X}_{obs,v}$, dan is dat of de waarneming $\{X_1 = x_1, X_2 = x_2\}$ of $\{X_1 = x_1, X_3 = x_3\}$. In het eerste geval is het tweede deel van het rechterlid van (6.26) gelijk aan 0 omdat $P(X_1 = x_1, X_2 = x_2 \mid \boldsymbol{r}_2 = (1,0,1)) = 0$; de kans op een antwoord op item 1 en 2, gegeven dat item 1 en 3 zijn waargenomen is immers 0. Verder volgt dan direct dat formule (6.26) in dat geval gelijk is met (6.27). In het tweede geval is, volgens dezelfde redenering, het eerste deel van het rechterlid gelijk aan 0 en ook (6.26) weer gelijk aan (6.27).

In ons voorbeeld gaan we, om een kortere notatie te krijgen, de itemparameters en de persoonsparameters transformeren, respectievelijk $\varepsilon_i = \exp(-\beta_i)$, $i = 1,2,3$ en $\exp(\theta) = \xi$. Vervolgens beschouwen we alle mogelijke uitkomsten waarvoor de somscore (6.25) gelijk aan 1 is en geven in tabel 6.5 de relevante kansen.

Tabel 6.5

Kansen op alle uitkomsten met $S_{obs} = 1$ in Raschmodel met drie items

$\boldsymbol{x}_{obs}, \boldsymbol{r}$	$P(\boldsymbol{x}_{obs}) = P(\boldsymbol{x}_{obs}, \boldsymbol{r})$	$P(\boldsymbol{x}_{obs} \mid \boldsymbol{r}_1)$	$P(\boldsymbol{x}_{obs} \mid \boldsymbol{r}_2)$
(1)	(2)	(3)	(4)
$x_1=1, x_2=0, 110$	$\dfrac{\phi.\xi\varepsilon_1}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_2)}$	$\dfrac{\xi\varepsilon_1}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_2)}$	0
$x_1=0, x_2=1, 110$	$\dfrac{\xi\varepsilon_2}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_2)}$	$\dfrac{\xi\varepsilon_2}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_2)}$	0
$x_1=1, x_3=0, 101$	$\dfrac{(1-\phi).\xi\varepsilon_1}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_3)}$	0	$\dfrac{\xi\varepsilon_1}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_3)}$
$x_1=0, x_3=1, 101$	$\dfrac{(1-\phi).\xi\varepsilon_3}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_3)}$	0	$\dfrac{\xi\varepsilon_3}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_3)}$
1	$\dfrac{\phi.\xi(\varepsilon_1+\varepsilon_2)}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_2)} + \dfrac{(1-\phi).\xi(\varepsilon_1+\varepsilon_3)}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_3)}$	$\dfrac{\xi(\varepsilon_1+\varepsilon_2)}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_2)}$	$\dfrac{\xi(\varepsilon_1+\varepsilon_3)}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_3)}$
s_{obs}	$P(s_{obs})$	$P(s_{obs} \mid \boldsymbol{r}_1)$	$P(s_{obs} \mid \boldsymbol{r}_2)$

In kolom (1) van tabel 6.5 staan alle mogelijke uitkomsten. Beschouwen we eerst kolom (2). Hierin staan in het bovenste deel de kansen op deze uitkomsten en in het onderste deel de kans dat $S_{obs} = 1$. De voorwaardelijk kans op een willekeurige uitkomst,

gegeven $s_{obs} = 1$, verkrijgen we door het delen van de term uit het onderste deel van de tabel door een term uit het bovenste deel. Er geldt immers

$$P(\boldsymbol{x}_{obs}, \boldsymbol{r}) = \frac{P(\boldsymbol{x}_{obs}, \boldsymbol{r}, s_{obs})}{P(s_{obs})} = \frac{P(\boldsymbol{x}_{obs}, \boldsymbol{r})}{P(s_{obs})}.$$

Als we zo'n deling uitvoeren zien we dat het resultaat afhangt van individuele parameter ξ. Waaruit volgt dat S_{obs} niet voldoende is voor ξ en dus ook niet voor θ, en we kunnen CML dus niet toepassen in de verdeling van $\boldsymbol{X}_{obs}$ of van $(\boldsymbol{X}_{obs}, \boldsymbol{R})$.

Wat er echter wel mogelijk is zien we in de kolommen (3) en (4) van tabel 6.5. Hierin staan voor ons voorbeeld de conditionele kansen op de uitkomsten, $P(\boldsymbol{x}_{obs} \mid \boldsymbol{R}_v = \boldsymbol{r}_b)$, $b = 1,2$, een de conditionele kans dat de somscore 1 is, $P(S_{obs} = 1 \mid \boldsymbol{R}_v = \boldsymbol{r}_b)$, $b = 1,2$, beiden gegeven de waarde van itemindicator variabele. Eenvoudig is na te gaan dat in de conditionele verdeling van $\boldsymbol{X}_{obs}$ gegeven $\boldsymbol{R}$ de somscore wel voldoende is voor de individuele parameter ξ. De kans op een uitkomst gegeven de somscore bepalen we in deze conditionele verdelingen weer door in tabel 6.5 de kans uit het onderste deel te delen op een term uit het bovenste deel. Er geldt namelijk:

$$\frac{P(\boldsymbol{x}_{obs} \mid \boldsymbol{r})}{P(s_{obs} \mid \boldsymbol{r})} = \frac{P(\boldsymbol{x}_{obs}, s_{obs} \mid \boldsymbol{r})}{P(s_{obs} \mid \boldsymbol{r})} = P(\boldsymbol{x}_{obs} \mid s_{obs}, \boldsymbol{r}). \tag{6.28}$$

Voor alle gegeven uitkomsten en ook voor de andere uitkomsten is eenvoudig na te gaan dat het resultaat van deze deling onafhankelijk is van de individuele parameter ξ.

In de conditionele verdelingen, gegeven de itemindicator, zitten we dus in dezelfde positie als in het Raschmodel voor volledige data: we hebben een voldoende statistiek waarmee voor elke persoon de individuele parameter kunnen uitconditioneren uit de aannemelijkheidsfunctie. Daarmee is dan ook voldaan aan de eerste voorwaarde om de CML-schattingsmethode te kunnen toepassen. Merk op aan (6.28) dat we alternatief zouden kunnen zeggen dat alleen S_{obs} en $\boldsymbol{R}$ gezamenlijk voldoende zijn voor de individuele parameter ξ of θ. Ging het in de theorie van Rubin (1976) en ook in paragraaf 6.5.1, waar we MML in stochastische designs bespraken, steeds om de vraag of we in de analyse de designvariabele konden negeren, bij CML is deze vraag niet aan de orde. Willen we CML toepassen dan zullen we de designvariabele expliciet in de analyse moeten meenemen, omdat er anders geen voldoende statistiek voor de individuele vaardigheid bestaat. Dus Rubins voorwaarden kunnen niet beslissend zijn

voor de toepassing van CML in stochastische onvolledige designs. Welke dat wel zijn gaan we nu behandelen.

Als we CML gaan toepassen gaan we dus uit van de verdeling van alle waargenomen toevalsvariabelen. In het algemeen kan dit geschreven worden als:

$$P_{\theta,\beta,\phi}(\boldsymbol{x_{obs}}, \boldsymbol{r}) = \prod_{v=1}^{n} P_{\theta_v,\beta,\phi}(\boldsymbol{x}_{obs,v} \mid \boldsymbol{r}_v) \cdot P_{\phi}(\boldsymbol{r}_v). \tag{6.29}$$

We gebruiken dezelfde notatie als eerder. We onderscheiden B waarden van de design-variabele $\boldsymbol{r}_b$, $b = 1,\ldots,B$; n_b is het aantal personen dat boekje b maakt; $\beta_{(b)}$ is de k_b-vector met de parameters van de items in boekje b. Dan kunnen we (6.29) herschrijven als:

$$P_{\theta,\beta,\phi}(\boldsymbol{x_{obs}}, \boldsymbol{r}) = \prod_{b=1}^{B} \prod_{v=1}^{n_b} P_{\theta_v,\beta_{(b)},\phi}(\boldsymbol{x}_{obs,v} \mid \boldsymbol{r}_b) \cdot \prod_{b=1}^{B} \prod_{v=1}^{n_b} P_{\phi}(\boldsymbol{R}_v = \boldsymbol{r}_b). \tag{6.30}$$

We zien in (6.30) dat we de aannemelijkheidsfunctie van alle waarnemingen kunnen schrijven als het produkt van twee termen. Het is in te zien dat het eerste deel van het rechterlid van (6.30) niets anders is dan het produkt van B volledige data aannemelijkheidsfuncties, zoals in hoofdstuk 4 is besproken. In elk boekje is er, zoals bij de volledige data, zoals we in het voorgaande zagen (6.28), voor elke persoon een voldoende statistiek S_{obs}, zodat geldt

$$\prod_{v=1}^{n_b} P_{\theta_v,\beta_{(b)},\phi}(\boldsymbol{x}_{obs,v} \mid \boldsymbol{r}_b) = \prod_{v=1}^{n_b} P_{\beta_{(b)}}(\boldsymbol{x}_{obs,v} \mid s_{obs,v}, \boldsymbol{r}_b) \cdot P_{\theta_v,\beta,\phi}(s_{obs,v} \mid \boldsymbol{r}_b). \tag{6.31}$$

Het eerste deel van het rechterlid van (6.31) hangt alleen nog maar af van de itemparameters $\beta_{(b)}$ en dit deel wordt in de CML-methode gemaximaliseerd naar de parameters β in plaats van het linkerlid. De maxima geven de itemparameterschattingen. De rechtvaardiging van de CML-methode hangt mede af van het feit of we het tweede deel van het rechterlid van (6.31) mogen weglaten uit de analyse. Zou het tweede deel van het rechterlid onafhankelijk zijn van β dan is het duidelijk dat het niet uitmaakt of we het linkerlid, de volledige aannemelijkheids-functie, dan wel alleen het eerste deel van het rechterlid, de conditionele aannemelijkheids-functie gebruiken. We zien echter dat ook het tweede deel van het rechterlid van (6.31), de verdeling van S_{obs}, afhangt van β. Het zo maar weglaten van dit deel zal in zijn algemeenheid natuurlijk niet dezelfde resultaten voor de itemparameterschattingen opleveren. Het is echter aangetoond (Andersen, 1973b) dat voor IRT-modellen die behoren tot de exponentiële familie, zie hoofdstuk 4, zoals het

Raschmodel en het OPLM model, die afhankelijkheid van het tweede lid van β een zeer speciale structuur heeft, waardoor het in dat geval gerechtvaardigd is het in de analyse buiten beschouwing te laten, en dat de resulterende schattingen de in hoofdstuk 4 gememoreerde goede statistische eigenschappen hebben. De speciale structuur komt er op neer dat de verdeling van S_{obs} niet rechtstreeks afhankelijk is van β; de afhankelijkheid is altijd gekoppeld aan de afhankelijkheid van de persoonsparameter. We zullen hier niet verder op ingaan en verwijzen voor details naar Andersen (1973b).

De voorgaande beschouwing geldt voor elk volledig boekje in onvolledige designs en natuurlijk ook voor aannemelijkheidsfunctie voor alle boekjes. Dus het is in onze modellen gerechtvaardigd om ook in onvolledige designs in plaats van het produkt over B boekjes van het linkerlid van (6.31) uit te gaan van het produkt over B boekjes van het eerste deel van het rechterlid: de conditionele aannemelijkheidsfunctie:

$$L_c = \prod_{b=1}^{B} \prod_{v=1}^{n_b} P_{\beta_{(b)}}(\boldsymbol{x}_{obs,v} \mid s_{obs,v}, \boldsymbol{r}_b) \tag{6.32}$$

Of het in stochastische designs gerechtvaardigd is om alleen (6.32) te beschouwen, hangt dan alleen nog maar af van de vraag of we ook het rechterdeel van de aannemelijkheidsfunctie (6.30):

$$\prod_{b=1}^{B} \prod_{v=1}^{n_b} P_{\phi}(\boldsymbol{R}_v = \boldsymbol{r}_b), \tag{6.33}$$

in de analyse weg kunnen laten. Het antwoord hierop is analoog aan de redenering hiervoor. Zolang (6.33) onafhankelijk is van de itemparameters β, dan is dat gerechtvaardigd. Als er afhankelijkheid is dan moet voor de rechtvaardiging van CML in stochastische designs de eerder omschreven speciale structuur aanwezig zijn. Is er rechtstreekse afhankelijkheid van (sommige) itemparameters in (6.33) dan is CML niet toegestaan. We bespreken nu de mogelijkheid van CML voor de drie stochastische designvormen.

CML in gerandomiseerde onvolledige designs

De designverdeling in gerandomiseerde designs wordt gegeven door (6.2):

$$\prod_{b=1}^{B} \prod_{v=1}^{n_b} P_{\phi}(\boldsymbol{R}_v = \boldsymbol{r}_b) = \prod_{b=1}^{B} \prod_{v=1}^{n_b} \phi_b. \tag{6.34}$$

En we zien dat (6.34) geheel onafhankelijk is van de itemparameters β, en dus dat toepassen van CML in gerandomiseerde onvolledige designs evenals bij MML geen problemen oplevert.

CML in meerfasen onvolledige designs

In meerfasen onvolledige designs kunnen we (6.33), met behulp van de itemindicator verdeling (6.4), schrijven als:

$$\prod_{b=1}^{B}\prod_{v=1}^{n_b} P_{\phi}(\boldsymbol{R}_v = \boldsymbol{r}_b) = \prod_{b=1}^{B}\prod_{v=1}^{n_b} P_{\phi}(\boldsymbol{R}_v = \boldsymbol{r}_b \mid \boldsymbol{x}_{obs,v}).P_{\beta_{(obs)},\theta_v}(\boldsymbol{x}_{obs,v}). \qquad (6.35)$$

In (6.35) zien we dat het tweede deel van het rechterlid rechtstreeks afhangt van de itemparameters van de items, waarvan de waargenomen waarden bepalen wie welk boekjes gaat maken. De speciale afhankelijkheidsstructuur, waarvan bij de rechtvaardiging van CML in het algemeen sprake is, is hier niet aanwezig. CML in meerfasen designs is dus niet mogelijk. Dit in tegenstelling tot MML waarbij, zoals we eerder zagen in paragraaf 6.5.1, de designvariabele in de analyse kon worden genegeerd om tot correcte resultaten te komen. Wij zullen dit met een voorbeeld met gesimuleerde data illustreren. Daarvoor beschouwen opnieuw het voorbeeld uit paragraaf 6.3.2. De tien middelmatig moeilijke items 1 tot 10 uit de sorteertoets hebben een moeilijkheid in het Raschmodel van 0. Voor de gemakkelijke items is $\beta_i = -1, i = 11,...,15$ en voor de moeilijke $\beta_i = 1\ , i = 16,...,20$. Als we 1000 itemantwoorden genereren voor vaardigheden getrokken uit een standaard normale verdeling en in de analyse alleen de antwoorden op de moeilijke items beschouwen voor de personen met een score van 6 of meer op de sorteertoets en de antwoorden op de gemakkelijke items alleen voor de personen met een score van 5 of minder op de sorteertoets, dan leveren analyses van deze gegevens de resultaten op uit tabel 6.6.

We zien in tabel 6.6 dat in de MML-analyse de itemmoeilijkheden bij het negeren van de designvariabele in dit tweefasen design goed worden geschat: er zijn geen geschatte moeilijkheden $\hat{\beta}_i$ die meer dan twee geschatte standaardfouten van de ingevoerde moeilijk-heden afliggen. Hetzelfde geldt voor de verdelingsparameters die onder in de tabel staan vermeld. Voor de CML-schattingen van de moeilijkheid geldt dit alleen maar voor de items van de sorteertoets (1 tot 10). Ze verschillen nauwelijks van de MML-schattingen. De overige itemmoeilijkheden worden systematisch onjuist geschat. De gemakkelijke items (11 tot 15) worden gemakkelijker geschat dan ze in werkelijkheid zijn en de moeilijke items (16 tot 20) moeilijker. Steeds is het verschil

tussen de geschatte moeilijkheid $\hat{\beta}_j$ en de echte moeilijkheid β_j meer dan twee geschatte standaardfouten. Tenslotte zij opgemerkt dat in de realisatie van deze simulatie van de 1000 personen die de sorteertoets maakten er vervolgens 556 met de gemakkelijke items verder gingen en 444 met de moeilijke. Dit verklaart de verschillen tussen de items in de geschatte standaardfouten in tabel 6.6.

Tabel 6.6

CML- en MML-analyse gesimuleerd meerfasen design

		CML		MML	
Item	β_i	$\hat{\beta}_i$	$SE(\hat{\beta}_i)$	$\hat{\beta}_i$	$SE(\hat{\beta}_i)$
1	0	0.043	.069	0.064	.068
2	0	-0.101	.069	-0.075	.069
3	0	-0.007	.069	0.016	.068
4	0	-0.081	.069	-0.056	.069
5	0	-0.036	.069	-0.013	.068
6	0	-0.076	.069	-0.051	.069
7	0	0.038	.069	0.059	.068
8	0	0.023	.069	0.044	.068
9	0	-0.026	.069	-0.003	.068
10	0	-0.071	.069	-0.046	.069
11	-1	-1.391	.090	-1.144	.097
12	-1	-1.286	.089	-1.033	.095
13	-1	-1.192	.090	-0.933	.095
14	-1	-1.310	.090	-1.058	.096
15	-1	-1.318	.090	-1.067	.096
16	1	1.314	.098	1.012	.105
17	1	1.410	.099	1.114	.106
18	1	1.420	.099	1.124	.106
19	1	1.381	.098	1.083	.106
20	1	1.266	.098	0.962	.105
	$\mu = 0$			$\hat{\mu} = 0.026(.038)$	
	$\sigma = 1$			$\hat{\sigma} = 0.944(.031)$	

Uit dit voorbeeld moge duidelijk zijn dat CML in een meerfasen design geen correcte resultaten oplevert en dus niet toegestaan is. Aangezien standaardprogrammatuur voor CML-analyse, bijvoorbeeld OPLM, geen rekening houdt met hoe de onvolledige gegevens zijn ontstaan, dient men hiervoor op de hoede te zijn.

CML in groepsgerichte designs

In groepsgerichte designs is (6.33) af te leiden uit de verdeling van de itemindicator variabele (6.5):

$$\prod_{b=1}^{B}\prod_{v=1}^{n_b} P_{\phi}(\boldsymbol{R}_v = \boldsymbol{r}_b) = \prod_{b=1}^{B}\prod_{v=1}^{n_b} P_{\pi_b}(Y_v = y_b). \qquad (6.36)$$

Het zal duidelijk zijn dat uitdrukking (6.36) niet van de itemparameters β afhangt. De kans dat een persoon tot een bepaalde groep b behoort wordt natuurlijk niet bepaald door de items die deze persoon maakt. Hieruit volgt dat CML met de conditionele aannemelijkheidsfunctie (6.32) in groepsgerichte stochastische designs zonder problemen kan plaatsvinden.

Ter illustratie volgt tenslotte het resultaat van de CML-analyse van de gesimuleerde gegevens in een groepsgericht design, waarvoor in tabel 6.4 de resultaten van de MML-analyses werden gegeven.

Tabel 6.7
CML-analyse in een gesimuleerd groepsgericht design

item	β_i	$\hat{\beta}_i$	$SE(\hat{\beta}_i)$
1	-2	-2.158	.113
2	-2	-2.099	.112
3	-2	-2.042	.111
4	-2	-2.076	.112
5	-2	-1.996	.110
6	0	0.006	.076
7	0	-0.016	.076
8	0	-0.050	.076
9	0	0.006	.076
10	0	0.006	.076
11	2	2.035	.111
12	2	2.080	.112
13	2	2.139	.113
14	2	1.967	.110
15	2	2.198	.114

In tabel 6.7 zien we dat alle CML-schattingen van de moeilijkheid $\hat{\beta}_i$ in dit groepsgerichte design minder dan twee standaardfouten van de ingevoerde waarden β_i afliggen. Als we resultaten vergelijken met de MML-analyse, waarbij we de achtergrondvariabele Y expliciet in de analyse meenemen, zie tabel 6.4, dan zien dat resultaten bijna perfect overeenstemmen.

De omstandigheid dat CML-analyses zelfs in stochastische groepsgerichte designs zonder problemen kunnen worden uitgevoerd is nog eens bevestiging van het feit, dat bij CML, ook bij volledige designs, geen rekening gehouden hoeft te worden met de wijze waarop de steekproef personen uit een populatie is getrokken. Dit in tegenstelling tot MML, waarbij altijd expliciet rekening moet worden gehouden met de wijze van steekproeftrekking en met het in dit geval relevante lidmaatschap van subpopulaties van personen.

6.6 Schatten van persoonsparameters in stochastische designs

Voor de persoonsparameterschattingen zijn in de IRT verschillende methoden beschikbaar. In paragraaf 4.5 werden behandeld de ML-schatter (grootste aannemelijkheid), de WML- schatter (gewogen-grootste-aannemelijkheid) en de EAP-schatter (de verwachting van de a posteriori verdeling van de vaardigheid). Bij het schatten van de persoonsparameter θ_v gaan we ervan uit dat de itemparameters uit het IRT-model waar we mee werken voldoende nauwkeurig zijn geschat om ze bekend te veronderstellen. We gaan dus uit van gecalibreerde itemverzamelingen. Reeds in paragraaf 6.1 werd gesteld dat een van de positieve eigenschappen van het werken met IRT-modellen is dat de vaardigheid van de personen met verschillende opgaven, deelverzamelingen uit een gecalibreerde itemverzameling, op dezelfde schaal worden geschat. Deze eigenschap impliceert dat voor de schatting van de vaardigheid de designvariabele geen rol speelt in de analyse. In deze paragraaf zullen nagaan of dit in het algemeen bij de drie besproken stochastische designtypen ook het geval is. We moeten daarbij in de bespreking onderscheid maken naar enerzijds de ML- en de WML-schatter en anderzijds de EAP-schatter van θ_v.

6.6.1 ML- en WML-vaardigheidsschatting in stochastische designs

In stochastische designs is steeds de vraag aan de orde of we in de analyse rekening moeten houden met het toevalsproces dat de designs genereert, dan wel dat we het stochastisch karakter van de designvariabele kunnen negeren. Omdat in de ML-schatting en de WML-schatting van de persoonsparameter dezelfde toevalsvariabele wordt beschouwd, namelijk het antwoordpatroon van persoon v op de items $\boldsymbol{X}_v = (X_{v1},...,X_{vk})$, heeft deze vraag bij beide methoden hetzelfde antwoord. We zullen daarom alleen de ML schatting nader beschouwen. De theorie van Rubin, behandeld in paragraaf 6.5. is ook hier weer direct toepasbaar.

In de eerdere notatie is de toevalsvariabele die ons interesseert $\boldsymbol{U}_v = \boldsymbol{X}_v$ waarvan de verdeling $f_\tau(\boldsymbol{u}_v)$ alleen afhangt van de onbekende parameter $\tau = \theta_v$. In gerandomiseerde en in meerfasen designs deelt de missing data indicator $\boldsymbol{M}_v$, die hier hetzelfde is als de itemindicator $\boldsymbol{R}_v$, de variabelen $\boldsymbol{U}_v$ op in:

$$\boldsymbol{U}_{obs,v} = \boldsymbol{X}_{obs,v} \text{ en } \boldsymbol{U}_{mis,v} = \boldsymbol{X}_{mis,v}.$$

In deze gevallen is eenvoudig na te gaan dat de verdeling van de itemindicator, respectievelijk (6.2) voor gerandomiseerde design en (6.4) voor meerfasen designs, op zijn minst voldoet aan de MAR-voorwaarde (6.10) voor het negeren van het design in

de analyse. Dus in deze designs kan de schatting gebaseerd worden op de marginale verdeling van de observaties:
$P_{\theta_v}(\boldsymbol{x}_{obs,v})$. Opgemerkt kan worden dat het negeren van de designvariabele bij het schatten van de persoonsparameter eveneens gerechtvaardigd is bij het adaptief toetsen, hetgeen immers een limietgeval is van meerfasen toetsen (zie paragraaf 6.3.2).

Bij groepsgerichte designs moet bij het schatten van de persoonsparameter analoog bij de MML-calibratie (paragraaf 6.3.3) onderscheid gemaakt worden tussen het wel en niet meenemen van de achtergrondvariabele Y in de analyse. Bij wel meenemen geldt

$$\boldsymbol{U}_{obs,v} = (\boldsymbol{X}_{obs,v}, Y_v) \text{ en } \boldsymbol{U}_{mis,v} = \boldsymbol{X}_{mis,v}. \tag{6.37}$$

De verdeling van de missing data indicator is (vergelijk met (6.21)):

$$\left.\begin{array}{l} P(\boldsymbol{M}_v = (\boldsymbol{r}_b, 1) \mid Y_v = y_b) = 1 \\ P(\boldsymbol{M}_v = (\boldsymbol{r}_b, 1) \mid Y_v \neq y_b) = 0 \end{array}\right\}, \quad b = 1,...,B; v = 1,...,n. \tag{6.38}$$

In (6.38) is $\boldsymbol{r}_b$ weer de $k-$vector met k_b maal een 1 op plaatsen die de geobserveerde items in boekje b indiceren, en $k-k_b$ maal een 0. De laatste 1 in de waarde van $\boldsymbol{M}_v$ indiceert het waarnemen van Y_v. Duidelijk is dan dat aan de MAR-voorwaarde (6.10) is voldaan en we in de analyse de designvariabele kunnen negeren en ons kunnen baseren op de marginale verdeling van de observaties $P_{\theta_v,\pi_b}(\boldsymbol{x}_{obs,v}, y_v)$. Merk op dat we deze verdeling kunnen schrijven als:

$$P_{\theta_v,\pi_b}(\boldsymbol{x}_{obs,v}, y_v) = P_{\theta_v}(\boldsymbol{x}_{obs,v} \mid y_v) \, . \, P_{\pi_b}(Y_v = y_b). \tag{6.39}$$

In (6.39) zien we dat voor het maximaliseren ervan naar θ_v we kunnen volstaan met het maximaliseren van het eerste deel van het rechterlid. In de IRT-modellen die wij beschouwen geldt hiervoor, vanwege de lokale stochastische onafhankelijkheid:

$$P_{\theta_v}(\boldsymbol{x}_{obs,v} \mid y_v) = \prod_{j \in obs,v} P_{\theta_v}(x_{vj} \mid y_v) = \prod_{j \in obs,v} P_{\theta_v}(x_{vj}). \tag{6.40}$$

Hierin staat $P_{\theta_v}(x_{vj})$ voor het IRT-model dat we beschouwen. We zien dus dat de aannemelijkheidsfunctie (6.40) die we, eventueel vermenigvuldigd met een functie van θ bij WML, die we maximaliseren voor het verkrijgen van de persoonsparameterschatting onafhankelijk is van de achtergrondvariabele Y. Dus ook hier geldt dat de persoons-

parameterschatting onafhankelijk is van de toevallige items, hier bepaald door de waarde van de achtergrondvariabele, die uit de gecalibreerde itemverzameling zijn afgenomen.

Als we in groepsgerichte designs de achtergrondvariabele niet zouden meenemen dan krijgen we voor de opdeling door de designvariabele van alle variabelen in plaats van (6.37):

$$\boldsymbol{U}_{obs,v} = \boldsymbol{X}_{obs,v} \text{ en } \boldsymbol{U}_{mis,v} = (\boldsymbol{X}_{mis,v}, Y_v). \tag{6.41}$$

En de verdeling van de designvariabele is als in (6.38), met dien verstande dat het laatste element altijd de waarde 0 heeft in plaats van 1, welke niet voldoet aan de MAR-voorwaarde (6.10), hetgeen betekent dat het design niet genegeerd kan worden. In dit geval echter zou het negeren geen consequenties hebben: het alleen beschouwen van de marginale verdeling van de observaties $P_{\theta_v}(\boldsymbol{x}_{obs,v})$ levert, vanwege eigenschap (6.40), dezelfde uitdrukking op voor de aannemelijkheidsfunctie als bij het wel meenemen van de achtergrondvariabele.

6.6.2 EAP vaardigheidsschatting in stochastische onvolledige designs

De EAP-schatter voor de vaardigheid is in tegenstelling tot alle voorgaande schattingsmethoden een bayesiaanse schatter en geen grootste-aannemelijkheidsschatter. Dat betekent dat de algemene theorie voor het negeren van de designvariabele in de analyse, zoals behandeld in paragraaf 6.5, hier niet direct van toepassing is. Rubin (1976) heeft echter ook voor bayesiaanse schattingsmethoden aangegeven onder welke voorwaarden het design in de analyse genegeerd kan worden. Het zou in het kader van dit boek te ver voeren om ook dit onderwerp uitgebreid te behandelen. We volstaan met op te merken dat voor het negeren van het design in een bayesiaanse analyse naast de voorwaarden die al gelden voor de ML-schattingen nog een extra voorwaarde moet gelden. Of aan deze voorwaarde voldaan is zullen we hierna voor de drie besproken stochastische designtypen kort bespreken.

De extra voorwaarden heeft betrekking op de eigenschappen van de a priori verdelingen die in de bayesiaanse analyse worden gebruikt. In het algemeen is aan de voorwaarden voor het negeren van de designvariabele in een bayesiaanse analyse voldaan, als de a priori verdelingen van de betrokken parameters onafhankelijk zijn. Bij het schatten van de persoonsparameters in stochastische designs hebben we te maken met twee parameters: de persoonsparameter θ en de parameter ϕ van de

verdeling van de designvariabele. Bij de mogelijkheid de designvariabele te negeren bij de EAP-schatting van θ zullen we de a priori relatie tussen deze parameters moeten beschouwen.

In gerandomiseerde designs zal er geen enkele a priori relatie zijn tussen θ en ϕ. Voor de gezamenlijke a priori verdeling van deze parameters zal dan ook voldaan zijn aan de onafhankelijkheidsvoorwaarde:

$$P(\theta,\phi) = P(\theta) \,.\, P(\phi). \tag{6.42}$$

Omdat ook aan de MAR-voorwaarde is voldaan levert het negeren van het design ook voor de EAP-schatting van θ geen probleem op.

Hetzelfde geldt voor meerfasen designs: de parameter ϕ wordt volledig bepaald door uitkomsten van waargenomen variabele, die op zichzelf natuurlijk wel van de vaardigheid θ afhangen, maar voor de waarnemingen zijn gedaan is er geen enkele aanname over het verband tussen θ en ϕ. Dus ook hier is de aanname (6.42) reëel. Met het voldoen aan de MAR-voorwaarde is dit samen voldoende om ook in meerfasen designs bij het bepalen van de EAP-schatting de designvariabele in de analyse te negeren. Zowel bij gerandomiseerde als meerfasen designs kunnen we dus, na specificatie van een a priori verdeling, met behulp van (4.119) en (4.120) een EAP-schatting bepalen.

Anders is de situatie bij groepsgerichte designs daar hebben we al in paragraaf 6.6.1 al gezien dat om te voldoen aan de MAR-voorwaarde de achtergrondvariabele in de analyse moeten meenemen. Echter ook geredeneerd vanuit de a priori verdelingen is het in te zien dat het a priori aannemen van onafhankelijkheid van θ en ϕ hier niet reëel is. De parameter van de designverdeling ϕ wordt immers volledig bepaald door de achtergrondvariabele. Zouden we (6.42) aannemen dat zou dat betekenen dat we a priori geen relatie zien tussen de vaardigheid θ en de waarde van achtergrondvariabele Y, echter de relatie tussen deze twee variabelen is evenwel juist de reden om met groepsgerichte designs te werken. Dus (6.42) geldt zeker niet. Om toch EAP-schatters te kunnen verkrijgen in groepsgerichte designs zullen we dus Y expliciet in de analyse moeten meenemen. Om te voldoen aan Rubins voorwaarden hebben we de geldigheid van (6.42) niet meer nodig echter alleen dat er gegeven de achtergrondvariabele, onafhankelijkheid is tussen de a priori verdelingen:

$$P(\theta,\phi \mid Y_v = y_b) = P(\theta \mid Y_v = y_b) \,.\, P(\phi \mid Y_v = y_b).$$

Deze aanname omtrent de a priori verdeling van parameters zal in de praktijk geen problemen opleveren. Voor een persoon v in groepsgerichte designs, met waarde y_b van achtergrond-variabele, kan de EAP-schatting dan met a priori verdeling $g(\theta) = P(\theta \mid \boldsymbol{Y}_v = y_b)$ bepaald worden.

7

Toepassingen van itemresponstheorie

In dit hoofdstuk komen een drietal toepassingen van itemresponstheorie (IRT) aan de orde. Ze zijn enerzijds bedoeld als illustratie van de theoretische uiteenzettingen in de vorige drie hoofdstukken, anderzijds dienen ze om enkele theoretische problemen die niet besproken werden, toe te lichten en een mogelijke oplossing voor te stellen.

De eerste toepassing gaat over een grootschalig Cito-project, de periodieke peiling van het onderwijsniveau (PPON). Het doel van deze peiling is het uitvoeren van metingen en daarover verslag doen. Een van de problemen waarmee het project werd geconfronteerd was het ontbreken van meetinstrumenten. De constructie van de meetinstrumenten en de eigenlijke peiling dienden in één fase te gebeuren. In paragraaf 7.1 worden de psychometrische aspecten van deze dubbele opdracht besproken.

De tweede toepassing behoort tot een domein dat in de psychologie bekend staat als leesbaarheidsonderzoek, een traditie die haar oorsprong vindt in het onderzoek van Vogel en Washburne (1928). De praktische vraagstelling bij dit soort onderzoek betreft de relatie tussen de leesvaardigheid van een jonge lezer en de moeilijkheid of leesbaarheid van een tekst. Met andere woorden, de vraag is of er een maat ontwikkeld kan worden die aangeeft of een bepaalde persoon met goed gevolg een gegeven tekst kan lezen. Hoewel iedereen wel bekend zal zijn met leeftijdscodes op boeken in jeugdbibliotheken, is een dergelijke aanduiding veel te ruw: de spreiding van de leesvaardigheid bij kinderen van dezelfde leeftijd is dermate groot dat deze leeftijdsaanduidingen te enen male onvoldoende zijn. In paragraaf 7.2 worden enkele aspecten van het leesbaarheidsonderzoek van Staphorsius (1992b) besproken.

De derde toepassing heeft betrekking op een beroemde test uit de psychologie, de 'verborgen-figurentest' van Witkin (1950). Met behulp van IRT is door Pennings (1991) een gemodificeerde versie van deze test gemaakt, zodat hij beter geschikt wordt voor diagnostische doeleinden dan de oorspronkelijke test, waarbij alleen aantal juiste antwoorden en gemiddelde antwoordtijd worden geregistreerd. Het is meteen een illustratie van een creatief gebruik van een IRT-model voor polytome items. Deze toepassing wordt in paragraaf 7.3 besproken.

7.1 De PPON-rekenpeiling

In 1987 begon in opdracht van het Ministerie van Onderwijs het project 'Periodieke Peiling van het Onderwijsniveau' (PPON) in het basisonderwijs. Het eerste vakgebied dat werd gepeild was rekenen aan het einde en in het midden van het basisonderwijs, dat wil zeggen bij leerlingen van ongeveer twaalf respectievelijk negen jaar. Het algemene doel van peilingsonderzoek in Nederland kan omschreven worden als: systematisch bijdragen aan het verkrijgen van een beeld van het leeraanbod en de effecten van onderwijs. PPON moet een empirische basis verschaffen voor de algemene maatschappelijke discussie over de inhoud en het niveau van het onderwijs. Concreet betekent dit bijvoorbeeld dat verschillen in leer-prestaties tussen belangrijke subpopulaties in kaart gebracht dienen te worden. De reken-peiling van 1987 is een eerste peiling in een reeks van periodiek herhaalde peilingen, en de resultaten moeten dienen als algemeen referentiepunt om ontwikkelingen in de tijd te kunnen evalueren. Dit aspect van de opdracht, samen met de verplichting om na elke peiling een gedeelte van de items te publiceren, vormt de eerste grote complicatie van de opdracht. De toetsen die gebruikt worden in opeenvolgende peilingen kunnen niet identiek zijn. Dit schept het probleem dat er maatregelen getroffen moeten worden, zodat verschillen in de tijd op gemiddelde prestatie niet ten onrechte kunnen worden toegeschreven aan verschillen in moeilijkheidsgraad.

Een tweede complicerende factor betrof de steekproeftrekking. Omdat het tot de opdracht behoorde betrouwbare en vrij nauwkeurige uitspraken te doen over relatief kleine subpopulaties, bijvoorbeeld etnische minderheden, kon niet worden volstaan met een eenvoudige aselecte steekproef uit de leerlingpopulatie. In dat geval zouden deze minder-heden in onvoldoende aantal in de steekproef vertegenwoordigd zijn. Daarom werd besloten een gestratificeerde steekproef te trekken op zo'n wijze dat scholen met veel leerlingen uit etnische minderheden proportioneel oververtegenwoordigd waren. Bovendien is het om praktische redenen onuitvoerbaar om binnen elk stratum een aselecte steekproef te trekken. Daarom werd gebruikt gemaakt van getrapte steekproeftrekking. Eerst werd uit de populatie van basisscholen een aselecte steekproef getrokken, en dan werd er binnen elke school uit de relevante leeftijdsgroep weer een aselecte steekproef getrokken.

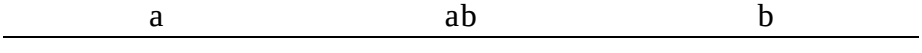
De derde complicatie had te maken met het feit dat de meetinstrumenten nog ontwikkeld moesten worden. Normaliter zou men in een dergelijk grootscheeps onderzoek een constructiefase verwachten waarin de meetinstrumenten ontwikkeld worden, en waarbij een afzonderlijke calibratiesteekproef getrokken wordt om de eigenschappen van het meet-instrument vast te stellen. Door de tijdsdruk bleek dit

echter niet mogelijk te zijn, zodat dezelfde steekproef moest fungeren als calibratiesteekproef en peilingssteekproef, met het theoretische risico dat bepaalde instrumenten van zo'n slechte kwaliteit konden blijken te zijn, dat er van peiling geen sprake meer zou zijn. Bovendien speelden nog andere aspecten van tijdsdruk mee: men kan leerlingen niet een willekeurig lange tijd items laten beantwoorden, en men kan de steekproef niet willekeurig groot maken, wil men de dataverzameling in een realistische periode afronden.

Om een gedetailleerde verslaglegging toe te laten, werd besloten het hele vakgebied rekenen op te delen in inhoudelijk zeer homogene deelgebieden, en voor elk deelgebied een afzonderlijke schaal te construeren. Zo werd bijvoorbeeld het onderwerp 'breuken' opgedeeld in de schalen 'optellen en aftrekken' en 'vermenigvuldigen en delen'. In totaal werden 27 deelgebieden onderscheiden voor de 12-jarigen en 13 deelgebieden voor de 9-jarigen. Voor een gedetailleerde onderwijskundige verantwoording van deze opdeling, zie Wijnstra (1988). Deze opdeling is natuurlijk een gelukkige omstandigheid om het werken met unidimensionale IRT-modellen aanvaardbaar te maken.

De verdere uiteenzetting heeft betrekking op de constructie van één schaal voor één deelgebied. Aan het einde van deze paragraaf komen we nog even terug op de vraagstukken die te maken hebben met het tegelijkertijd hanteren van meer schalen.

In hoofdstuk 4 is het begrip informatiefunctie uiteengezet, waarbij beklemtoond werd dat itemantwoorden niet altijd evenveel informatie geven over de onderliggende vaardigheid. Voor een praktische toepassing als PPON betekent dit dat het nutteloos is hele moeilijke items door hele zwakke leerlingen en zeer gemakkelijke items door hele vaardige leerlingen te laten beantwoorden, omdat die antwoorden nauwelijks informatie opleveren voor het schatten van de itemparameters of de individuele vaardigheid. Om het verzamelen van nutteloze gegevens zoveel mogelijk te vermijden, werd tot de volgende proefopzet besloten. Op grond van het oordeel van de leerkracht, en enkele objectieve criteria zoals het niveau van het geplande vervolgonderwijs, werden alle leerlingen die aan de peiling deelnamen toegewezen aan één van twee niveaugroepen, verder aan te duiden als A en B, waarbij B als vaardiger werd beoordeeld dan A. Merk op dat de groepsindeling slechts één keer plaats vond, en gebruikt werd voor elk van de schalen die de leerlingen beantwoordden. Door de itemconstructeurs werden de items die voor de schaal werden ontwikkeld, ingedeeld in drie niveaus: a voor de gemakkelijke, b voor de moeilijke en ab voor de middelmatig moeilijke items. Het afnamedesign dat werd gebruikt is weergegeven in figuur 7.1. Het betreft dus een onvolledig, groepsgericht design (zie hoofdstuk 6).



Figuur 7.1
Design in het PPON-onderzoek

De designvariabele, het al dan niet aanbieden van een item, is afhankelijk van de schatting van het niveau door de leerkracht, waarbij het aannemelijk is dat deze schatting enige validiteit heeft voor de latente variabele die door de items wordt gemeten, maar anderzijds weer niet volledig samenvalt met de antwoorden op de items die wel zijn aangeboden. Het al dan niet aanbieden van bepaalde items is dus niet volledig bepaald door de geobserveerde itemantwoorden, maar is ook afhankelijk van een variabele die correleert met de niet geobserveerde antwoorden. Dit wil zeggen dat de procedure waardoor het design tot stand gekomen is, niet verwaarloosd mag worden bij ML-schattingen van de modelparameters, op straffe van onzuivere en inconsistente schattingen. Zie hoofdstuk 6 voor een theoretische uiteenzetting hierover. Deze vaststelling heeft een paar heel belangrijke implicaties.

Ze betekent in de eerste plaats dat we een model moeten maken waarin niet alleen de kansen beschreven worden op een goed antwoord, gegeven dat het item aangeboden wordt, zoals bijvoorbeeld het Raschmodel, maar dat we tevens de kansen moeten beschrijven dat een bepaalde leerling, met een bepaalde vaardigheid θ, in de A- of B-groep terecht komt. Stel dat we aannemen dat in de totale populatie θ normaal verdeeld is, dan is het niet realistisch aan te nemen dat alle leerlingen met een θ-waarde boven een bepaalde grenswaarde θ_0 aan de B-groep worden toegewezen, en alle andere leerlingen aan de A-groep. Dit zou immers impliceren dat de toewijzingsprocedure foutloos is, dit wil zeggen dat het leerkrachtoordeel perfect betrouwbaar is en perfect valide met betrekking tot θ. Dit betekent dat in het model de grenswaarde θ_0, de betrouwbaarheid en de validiteit van de leerkrachtoordelen moeten worden opgenomen. Bovendien is dit nog maar een grove benadering van de werkelijkheid, want niet alle leerkrachten beoordelen even betrouwbaar en valide. Dus de verschillen tussen leerkrachten zouden eigenlijk ook gemodelleerd moeten worden.

De tweede implicatie heeft te maken met de wijze van steekproeftrekken. Zelfs al is de veronderstelling waar dat de vaardigheid in de populatie normaal verdeeld is, dan kunnen we dit niet zonder meer gaan invoeren als een modelveronderstelling, omdat de steekproef niet aselect uit de populatie is getrokken. Er moet minstens een model gehanteerd worden voor elk stratum dat voor de steekproeftrekking is gedefinieerd.

Willen we standaard ML-schattingen gaan toepassen, dan zijn we dus verplicht een zeer complex model te gaan ontwikkelen. Nu zou men kunnen redeneren dat al die

argumenten betrekking hebben op de marginale verdeling van θ, en aangezien itemparameterschattingen met MML robuust zijn tegen schendingen van de normaliteitsassumptie (zie het voorbeeld in paragraaf 4.3.6), het niet veel zal uitmaken als we MML-schattingen maken met de modelaanname van één enkele normale verdeling. Jammer genoeg is in dit geval het model niet robuust genoeg, en treden er belangrijke vervormingen op in de schattingen van de itemparameters: de moeilijkheid van de moeilijke b-items wordt systematisch onderschat en die van de gemakkelijke a-items wordt systematisch overschat (Eggen, 1990).

Iets algemener geformuleerd komt het hele probleem erop neer dat we voor de constructie van een meetinstrument opgezadeld worden met een aantal netelige bijkomende problemen die in feite niets met de validiteit van het meetinstrument te maken hebben, maar wel met de verdeling in de populatie van de latente variabele die we met het meetinstrument willen gaan meten. Men zou kunnen opperen dat de onderzoekers, door zo'n ingewikkelde proefopzet te kiezen, dit probleem grotendeels aan zichzelf te wijten hebben. Echter, met een eenvoudige proefopzet is het probleem niet opgelost. Stel dat er een enkelvoudige aselecte steekproef uit de populatie was getrokken, en dat alleen de eenvoudige vraag moest worden beantwoord of jongens gemiddeld meer, minder of evenveel presteren als meisjes, waarbij echter ook in de toekomst moet kunnen worden nagegaan of een eventueel verschil met de tijd toeneemt of afneemt. Door gebruik te maken van een MML-schattingsprocedure om de itemparameters te schatten zijn we verplicht vooraf, per hypothese, een standpunt in te nemen over de structuur van de latente variabele in de populatie. Indien we geloven dat er geen verschil is, kunnen we volstaan met de assumptie van één normale verdeling. Denken we echter dat er verschil zal zijn dan dienen we een verschillende verdeling aan te nemen voor jongens en voor meisjes. Door het invoeren van een hypothese over de verdeling van de latente vaardigheid worden meetprobleem (de eigenschappen van het meetinstrument) en het structurele probleem (de verdeling van de vaardigheid in de populatie) in één samengesteld model met elkaar vermengd. En de grote problemen duiken op indien het model, als geheel, verworpen dient te worden, omdat het statistische toetsingsarsenaal waarover we beschikken niet garandeert dat er in alle gevallen een scherp onderscheid gemaakt wordt tussen schendingen in de meetcomponent en de structurele component van het model.

Het is natuurlijk een veel comfortabeler positie indien het meetmodel gevalideerd kan worden zonder dat aannamen over het structurele model hoeven te worden gemaakt. Dit is mogelijk indien de parameters die betrekking hebben op het meetmodel met de CML-schattingsmethode kunnen worden geschat. Toen het onderzoek uitgevoerd werd, was echter alleen het Raschmodel beschikbaar als IRT-model waar

CML mogelijk was. Het Raschmodel is echter nogal restrictief door de eis van gelijke discriminatie voor alle items, waardoor bij de constructie van een schaal in veel gevallen tamelijk veel items moeten worden verwijderd. Daarom is OPLM ontwikkeld als een soort compromis. Dit model heeft de flexibiliteit van het tweeparameter-logistische model maar het laat CML-schatting van zijn moeilijkheidsparameters toe. De theorie van OPLM is besproken in hoofdstuk 5. Van de ongeveer 500 items in de 40 schalen van de peiling rekenen moest minder dan vijf procent verwijderd worden op grond van de statistische toetsen die in het OPLM-programma zijn geïmplementeerd.

Wanneer het meetmodel eenmaal geaccepteerd is, kan het meetinstrument gebruikt worden om onderzoek te doen naar structurele vraagstukken. Dit kan op verschillende manieren gebeuren. Om een duidelijk idee te hebben van de werkwijze beperken we ons hier tot het geval van twee achtergrondvariabelen, geslacht (jongen-meisje) en herkomst (Nederlands - buitenlands). Als algemene hypothese nemen we aan dat beide variabelen een deel van de variabiliteit in de leerprestatie verklaren. Stellen we de afhankelijke variabele voor als Y_{vjk}, waarbij de index *v* verwijst naar een individu, de index j naar de subpopulatie van de jongens $(j=1)$ respectievelijk meisjes $(j=2)$ en de index *k* naar de subpopulatie van Nederlanders $(k=1)$ respectievelijk buitenlanders $(k=2)$. Een simpel lineair model is gegeven door

$$Y_{vjk} = \mu + \alpha_j + \beta_k + \varepsilon_{vjk}, \tag{7.1}$$

waarin μ een algemene constante is, α_j het effect van de j-de waarde van de geslachts-variabele, en β_k het effect van de k-de waarde van de herkomstvariabele. De grootheid ε_{vjk} is het zogenaamde residu, en wordt beschouwd als een toevalsvariabele waarvoor een bepaalde verdeling wordt aangenomen. We zullen, in overeenstemming met de gewone veronderstellingen uit de variantie-analyse, aannemen dat alle residuen normaal verdeeld zijn met gemiddelde 0 en variantie σ^2:

$$\varepsilon_{vjk} \sim N\left(0,\sigma^2\right). \tag{7.2}$$

Het model, gegeven door (7.1), is niet geïdentificeerd, omdat voor elke gegeven oplossing een andere gemaakt kan worden door α_j met een willekeurige constante c_1 en β_k met een willekeurige constante c_2 te vermeerderen, en ter zelfder tijd $c_1 + c_2$ van μ af te trekken. Er zijn dus oneindig veel mogelijke oplossingen en willen we zinvol over het model praten dan dienen we een oplossing te kiezen. Dat doen we door wat vaak 'technische restricties' genoemd worden, op te leggen aan de parameters. Wij zullen de restricties zo kiezen dat alle effectparameters die '1' hebben als index gelijk worden gesteld aan 0. Dus

$$\alpha_1 = \beta_1 = 0. \tag{7.3}$$

Merk op dat het gemiddelde van nul voor de residuen ook zo'n technische restrictie is en dat we ook een willekeurige andere waarde voor dit gemiddelde hadden kunnen kiezen. De restricties die we hier gekozen hebben, geven echter een elegante interpretatie aan de parameter μ. Beschouw daartoe de verwachte waarde van Y_{v11}:

$$\begin{aligned} \mathcal{E}(Y_{v11}) &= \mu + \alpha_1 + \beta_1 + \mathcal{E}(\varepsilon_{vjk}) \\ &= \mu + 0 + 0 + 0 = \mu . \end{aligned} \tag{7.4}$$

De parameter μ is dus de verwachte waarde van de afhankelijke variabele voor de subpopulatie waar alle categorieën hun 'eerste' of beter gezegd hun referentiewaarde aannemen. In het voorbeeld is 'jongen' de referentiecategorie voor de variabele 'geslacht' en 'Nederlander' de referentiecategorie voor de variabele 'herkomst'. De parameter μ is dus de gemiddelde θ-waarde van de jongens van Nederlandse herkomst.

Om de modelparameters $(\alpha_2, \beta_2, \sigma^2)$ consistent te schatten is het niet nodig dat de steekproef een aselecte steekproef is uit de totale populatie. De twee achtergrondvariabelen samen delen de totale populatie op in vier subpopulaties, en het is voldoende dat de steekproef uit elke subpopulatie beschouwd kan worden als een aselecte steekproef. De schattings-methode die gebruikt wordt is ML, waarbij de schattingen van de itemparameters uit de calibratiefase als de 'echte' itemparameters, dus als bekende constanten worden behandeld.

Een belangrijke vraag is natuurlijk wat we moeten nemen als de afhankelijke variabele *Y* in (7.1). Als we (7.1) werkelijk als een lineair model voor de vaardigheid θ beschouwen, lijkt het voor de hand te liggen *Y* in (7.1) door θ te vervangen, maar dan hebben we het probleem dat θ latent, dus niet geobserveerd, is. Een mogelijke oplossing is θ te vervangen door een zogenaamde 'proxy', bijvoorbeeld een schatting van θ. De Warm-schatter is een goede kandidaat omdat die schatter voor alle scores bestaat, en bijna zuiver is. Een andere goede kandidaat is de gewogen toetsscore, omdat deze voor niet al te extreme scores een bijna lineaire functie van de Warm-schatter is. Toch kleven aan beide benaderingen een paar nadelen, die men niet moet verwaarlozen.

Het eerste nadeel betreft het verlies aan nauwkeurigheid: de schattingen van θ zijn behept met een schattingsfout. Vullen we in het linkerlid van (7.1) zo'n schatting in, dan moet het residu ε_{vjk} geïnterpreteerd worden als de som van een 'waar' residu, dit wil zeggen, de fout bij het voorspellen van θ uit de predictoren, en de schattingsfout. Daardoor zal de residuele variantie toenemen, maar tevens de standaardfout van de schatters van de regressieparameters μ, α_2, β_2.

Het tweede nadeel heeft te maken met de overblijvende onzuiverheid, en de ongelijke verdeling van die onzuiverheid over de vier subpopulaties. Stel dat in één van

de vier subpopulaties relatief veel perfecte en relatief weinig nulscores voorkomen, dan is de gemiddelde Warm-schatting van de steekproef uit deze subpopulatie een onderschatting van het populatiegemiddelde, en deze onzuiverheid zal ook de schatting van de regressie-parameters beïnvloeden.

Deze twee overwegingen hebben er toe geleid dat in (7.1) toch θ werd ingevuld als afhankelijke variabele. Hoewel θ zelf niet geobserveerd is, hebben we toch informatie over θ via de itemantwoorden. Hierna volgt een korte schets van de schattingsprocedure.

Stellen we het antwoordpatroon van persoon v uit de (j,k)-de subpopulatie voor door $\boldsymbol{x}_{vjk}$ en de bijbehorende score door s_{vjk}, en definiëren we $\lambda = (\mu, \alpha_2, \beta_2, \sigma^2)$, dan kan de aannemelijkheidsfunctie gegeven dit antwoordpatroon, geschreven worden als:

$$\begin{aligned} L(\lambda ; \boldsymbol{x}_{vjk}) &= P(\boldsymbol{x}_{vjk} \mid s_{vjk})\, P(s_{vjk}) \\ &= P(\boldsymbol{x}_{vjk} \mid s_{vjk}) \int_{-\infty}^{+\infty} P(s_{vjk} \mid \theta)\, g_{jk}(\theta;\lambda)\, d\theta\,, \end{aligned} \tag{7.5}$$

waarin $g_{jk}(\theta;\lambda)$ de dichtheidsfunctie is van de verdeling van θ in de (j,k)-de subpopulatie. Het residu ε_{vjk} in het rechterlid van (7.1) is de enige toevalsvariabele, en uit (7.1) en (7.2) volgt dus dat θ_{jk}, dat is de toevalsvariabele θ in de (j,k)-de subpopulatie, normaal verdeeld is met gemiddelde $\mu + \alpha_j + \beta_k$ en variantie σ^2. De eerste factor in het rechterlid van (7.5) is geen functie van de parameters λ, en kan dus behandeld worden als een constante. De aannemelijkheidsfunctie gegeven de itemantwoorden van alle personen samen is het produkt van uitdrukkingen zoals het rechterlid van (7.5), en de ML-schattingen zijn die waarden van de parameters die de aannemelijkheidsfunctie maximaliseren. Een gedetailleerde uiteenzetting van de schattingsprocedure is gegeven in Verhelst en Eggen (1989).

In tabel 7.1 is een voorbeeld gegeven van de effectschattingen van zeven achtergrond-variabelen voor de schaal 'meten en maateenheden' voor de 9-jarigen. De variabele 'stratum' is de stratificatievariabele die gebruikt werd bij het steekproeftrekken, de variabele 'herkomst' geeft aan of de leerling Nederlander (N), dan wel buitenlander (B) was. De variabele 'leertijd' maakt onderscheid tussen kinderen die op het moment van de dataverzameling een kalenderleeftijd hadden van niet meer dan 109 maanden (L), en leerlingen die ouder waren (H). Omdat de data afkomstig zijn van leerlingen die in groep 5, voorheen derde klas, zaten, betreft deze laatste categorie dus leerlingen die één of meer keren gedoubleerd hebben. De variabele 'methode' verwijst naar de gebruikte rekenmethode. Voor de effectschattingen is gebruik gemaakt van de tweedeling Modern-Traditioneel. Categorie '1' van de variabele 'aanbod' verwijst naar leerlingen die, op het moment van de dataverzameling reeds onderwijs hadden

gekregen in de basisprincipes waarop de items een beroep doen. Naast deze variabelen is ook de variabele 'design' opgenomen. Categorie A verwijst naar de kinderen die de 'a' en 'ab' items hebben beantwoord, en categorie B naar de kinderen die de items 'ab' en 'b' voorgelegd kregen. Bij het schatten van de parameters worden de effecten uitgedrukt in de schaal die door de itemparameters is gedefinieerd. In tabel 7.1 is echter een lineaire transformatie toegepast op de schaal, waardoor het geschatte gemiddelde van de totale populatie gelijk is aan 250 en de standaarddeviatie 50. Voor elke variabele is de eerst gerapporteerde categorie gekozen als referentiecategorie. De verhouding z tussen parameter-waarden en standaardfout is bij benadering standaardnormaal verdeeld en kan gebruikt worden als toetsingsgrootheid om voor een parameter α_j de nulhypothese $\alpha_j = 0$ te toetsen. Het is interessant op te merken dat men aan de hand van deze tabel ook enig inzicht kan krijgen in de validiteit van het leerkrachtoordeel: de leerlingen die de moeilijkste items hebben gekregen liggen gemiddeld ongeveer tweederde standaardafwijking boven de kinderen die de gemakkelijke items voorgelegd kregen. Een gedetailleerder onderzoek naar de informatiewinst bij groepsgerichte designs kan men vinden in Verhelst (1989).

Tabel 7.1

Effectschattingen van zeven achtergrondvariabelen
op de schaal 'meten en maateenheden'

Variabele	Cat.	*n*	Eff.	*SE*(eff)	*z*=eff/*SE*
Stratum	1	333	0	---	---
	2	350	-11.49	4.02	-2.86
	3	403	-19.16	4.22	-4.55
Gewicht	N	927	0	---	---
	B	159	-36.72	4.96	-7.40
Geslacht	M	557	0	---	---
	V	529	-7.16	3.19	-2.24
Leertijd	L	902	0	---	---
	H	184	-17.51	4.27	-4.10
Methode	M	654	0	---	---
	T	432	-15.70	3.27	-4.80
Aanbod	1	834	0	---	---
	0	252	-7.59	3.78	-2.01
Design	A	514	0	---	---
	B	572	36.60	3.22	11.36

De effecten in de kolom 'Eff' geven het contrast aan met de referentiecategorie. Het effect van de categorie V van de variabele 'geslacht' bedraagt -7.16 eenheden, dit is ongeveer een zevende deel van de standaardafwijking in de populatie. De geassocieerde *z*-waarde van -2.24 is significant op het 5%-niveau, waarmee wordt aangegeven dat het geslacht, naast de andere variabelen die in de analyse zijn opgenomen, een niet te verwaarlozen effect op de prestatie heeft. Bij de interpretatie van de gerapporteerde contrasten dient men, net als bij de gewone regressie-analyse, zeer voorzichtig te zijn. Uit de tabel volgt niet dat meisjes gemiddeld 7.16 punten lager scoren dan jongens. Het is zelfs mogelijk dat meisjes gemiddeld hoger scoren, zoals uit het volgende fictieve voorbeeld blijkt. Veronderstel dat er slechts twee achtergrond-variabelen van belang zijn, 'geslacht' en 'leertijd', en dat de populatiewaarden van de effecten gelijk zijn aan de geschatte waarden uit tabel 7.1, namelijk -7.16 voor de categorie V van de variabele 'geslacht' en -17.51 voor de categorie H van de variabele 'leertijd'. Veronderstel verder dat de gezamenlijke verdeling van de variabelen 'geslacht' en 'leertijd' overeenkomt met tabel 7.2. Dan is het niet moeilijk na te rekenen dat de gemiddelde θ-waarde van de jongens μ_M gegeven is door

$$\mu_M = [.1(\mu + \alpha_1 + \beta_1) + .4(\mu + \alpha_1 + \beta_2)]/.5$$
$$= [.1(\mu + 0 + 0) + .4(\mu + 0 - 17.51)]/.5 = \mu - 14.008 ,$$

terwijl het populatiegemiddelde van de meisjes,

$$\mu_V = [.4(\mu + \alpha_2 + \beta_1) + .1(\mu + \alpha_2 + \beta_2)]/.5$$
$$= [.4(\mu - 7.16 + 0) + .1(\mu - 7.16 - 17.51)]/.5 = \mu - 10.662 \text{ bedraagt}.$$

Tabel 7.2
Niet-orthogonale verdeling van achtergrondvariabelen, leidend tot Simpsons paradox.

	geslacht	
leertijd	M: $\alpha_1 = 0$	V: $\alpha_2 = -7.16$
L: $\beta_1 = 0$	0.1	0.4
H: $\beta_2 = -17.51$	0.4	0.1

Dus, zowel in de subpopulatie 'leertijd = L' als in de subpopulatie 'leertijd = H' doen de meisjes het minder goed dan de jongens, doch gemiddeld over de hele populatie doen de meisjes het beter. De verklaring van dit paradoxale fenomeen is gelegen in het feit dat beide variabelen, 'geslacht' en 'leertijd' in de populatie niet onafhankelijk zijn,

of zoals men meestal zegt, niet orthogonaal zijn. Dit fenomeen is voor het eerst in de literatuur beschreven door Simpson (1951), en staat bekend als Simpsons paradox. De interpretatie van het geslachtseffect dient dan ook conditioneel te gebeuren: de meisjes scoren gemiddeld 7.16 punten lager dan de jongens indien de andere achtergrondvariabelen constant worden gehouden. Merk op dat de gemiddelde θ-waarde van de jongens of van de meisjes niet uit tabel 7.1 kan worden berekend, omdat de gezamenlijke verdeling van de zeven achtergrond-variabelen niet gegeven is.

Met betrekking tot de standaardfouten dient opgemerkt te worden dat de gerapporteerde getallen een beetje te optimistisch zijn om drie redenen. Ten eerste, de standaardfouten, berekend uit de informatiematrix gelden alleen asymptotisch. In eindige steekproeven zijn de standaardfouten groter. In de tweede plaats is er geen rekening gehouden met het feit dat de itemparameters niet bekend zijn, en dat we ons beholpen hebben met schattingen. Deze schattingen bevatten echter een schattingsfout waarmee geen rekening is gehouden bij het berekenen van de standaardfouten van de regressieparameters. Ten derde is het zo dat de variabelen in tabel 7.1 niet allemaal dezelfde status hebben. De variabelen 'stratum' en 'methode' zijn geen leerlinggebonden variabelen, maar schoolvariabelen. Alle leerlingen in de steekproef die uit dezelfde school komen hebben dezelfde rekenmethode gevolgd. Dit betekent dat, indien 'methode' een effect heeft, de residuen voor leerlingen uit dezelfde school niet onafhankelijk van elkaar zijn. Deze afhankelijkheid is in de analyse veronachtzaamd; er is gedaan alsof alle variabelen leerlinggebonden zijn. Het resultaat is dat de gerapporteerde standaardfouten systematisch te klein zijn. Vergelijk met hoofdstuk 2, de discussie over intraklassecorrelatie. Een correcte analyse zou vereisen dat elke variabele op zijn juiste niveau geanalyseerd wordt. Dergelijke analysemethoden worden aangeduid als multi-niveau- of multi-level-analyses. Er is echter geen programmatuur voorhanden om een multiniveau-analyse uit te voeren waarbij de afhankelijke variabele niet geobserveerd is. Het effect van de fout is, hoewel niet precies bekend, in het geval van de PPON-analyses waarschijnlijk erg klein, omdat de proefopzet zo werd ingericht dat van eenzelfde school niet meer dan vier leerlingen de items van eenzelfde schaal beantwoordden.

Tenslotte zij er nog op gewezen dat de data verzameld zijn in een onvolledige proefopzet, zie figuur 7.1. Voor de schatting van de effectparameters vormt dit geen enkel probleem, omdat in formule (7.5) rekening gehouden wordt met het design, hoewel dat niet expliciet is aangegeven. De factor $P(s_{vjk}|\theta)$ is een functie van de parameters van de items die persoon *v* heeft beantwoord.

7.2 De Cito leesbaarheidsindex voor het basisonderwijs

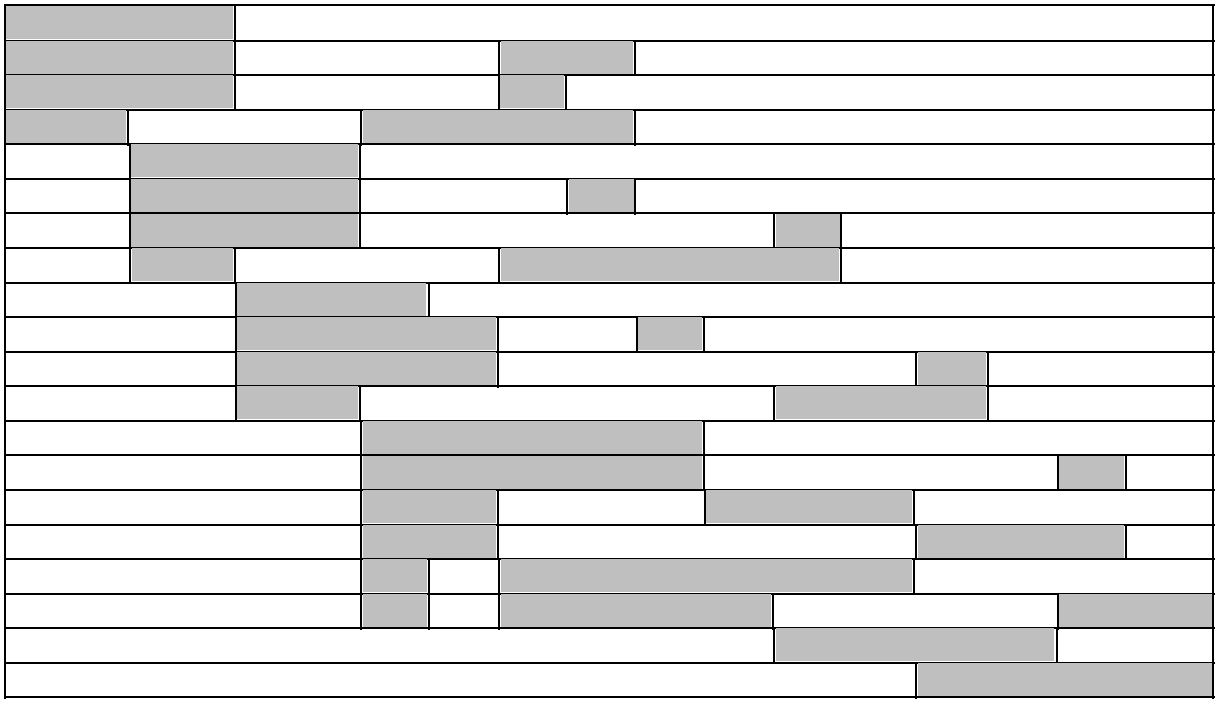
Leesbaarheid

Leesbaarheidsonderzoek heeft in verreweg de meeste gevallen als praktische bedoeling het construeren van een leesbaarheidsindex. Een bruikbare methode hiertoe is de zogenaamde cloze-procedure. Deze procedure bestaat uit het weglaten van woorden uit een tekst volgens een vast patroon. Leerlingen wordt gevraagd de ontbrekende woorden in te vullen. In het te bespreken onderzoek werd elk zevende woord weggelaten, elke tekst heeft zo zeven varianten. Middelen van het aantal correcte antwoorden in een representatieve steekproef over de varianten van de tekst, is nu een maat voor de moeilijkheid van de tekst. Teksten kunnen op deze manier worden geordend naar moeilijkheid. Het is natuurlijk niet praktisch om voor elke nieuwe tekst waarvan men de leesbaarheid wil bepalen deze cloze-procedure toe te passen. Daarom wordt gezocht naar formele tekstkenmerken die in combinatie de gemiddelde score van de tekst goed konden voorspellen. Goede voorspellers zijn onder meer de gemiddelde woordlengte, de gemiddelde zinslengte en het percentage frequente woorden in de tekst. Deze predictoren, die gemakkelijk en betrouwbaar kunnen worden gemeten, worden dan gebruikt als onafhankelijke variabelen in een regressievergelijking. Staphorsius (1992a; maar zie ook Staphorsius & Krom, 1985a en 1985b) vond een multipele correlatie van .85 bij het voorspellen van de gemiddelde cloze-score. De regressie-coëfficiënten die in dit onderzoek zijn gevonden, kunnen dan toegepast worden op willekeurige teksten waarvan de formele kenmerken zijn bepaald. De uitkomst van deze regressieformule, dat wil zeggen de voorspelde gemiddelde cloze-score, wordt de CLIB-waarde van de tekst genoemd. CLIB is de afkorting van Cito leesbaarheidsindex voor het basisonderwijs.

De leesbaarheidsindex van een tekst laat wel toe teksten in moeilijkheid te ordenen, doch hij is niet voldoende om aan te geven of een bepaalde persoon geschikt is voor een gegeven tekst, dat wil zeggen of die persoon de tekst kan lezen en begrijpen. Wat daartoe nodig is, is een maat voor de leesvaardigheid van de persoon en de relatie tussen die lees-vaardigheid en de CLIB-waarde van de tekst. Met andere woorden, we moeten antwoord kunnen geven op de vraag of een leerling met leesvaardigheid x in staat is een tekst met CLIB-waarde y te begrijpen.

Leesvaardigheid

Staphorsius (1992b) heeft een teksttoets ontwikkeld waarbij gebruik werd gemaakt van IRT. De items van de toets bestaan uit tekstfragmenten waaruit een of meer woorden zijn weggelaten. De leerlingen moeten het fragment completeren door uit vijf gegeven antwoordalternatieven het juiste te kiezen. De items zijn zo geconstrueerd dat het juiste antwoord alleen gevonden kan worden indien de tekst die voorafgaat aan en volgt op het ontbrekende stuk, is begrepen. In totaal werden 42 teksten gebruikt die werden opgedeeld in zes fragmenten van ongeveer 180 woorden, zodat er in totaal meer dan 250 items waren. Het spreekt vanzelf dat niet alle items aan eenzelfde persoon ter beantwoording konden worden aangeboden. Het hele onderzoek had betrekking op leerlingen van groep 4 tot en met groep 8 en de variatie in de moeilijkheid van de teksten was voldoende groot om bij het toewijzen van de teksten rekening te kunnen houden met verschillen in leesvaardigheid tussen de leerlingen. Aldus ontstond een onvolledig design dat in principe dezelfde structuur had als het design in figuur 7.1. Het was iets gecompliceerder, omdat de dataverzameling zich over verschillende jaren uitstrekte, zodat een aantal leerlingen gedurende hun hele schoolloopbaan gevolgd kon worden. Een gedeelte van het uiteindelijk gerealiseerde design is afgebeeld in figuur 7.2. De rijen in de figuur komen overeen met groepen leerlingen, geordend volgens geschat leesniveau; de kolommen komen overeen met items geordend volgens geschat moeilijkheidsniveau. In totaal werden meer dan 20.000 antwoordpatronen verzameld, waarbij elk antwoordpatroon de antwoorden bevatte op tussen de 30 en 60 items. Het aantal leerlingen dat aan het onderzoek deelnam was beduidend minder omdat een behoorlijk aantal leerlingen verschillende keren aan de testafname, met gedeeltelijk andere items, deelnam. Elk item werd minimaal 850 keer beantwoord.

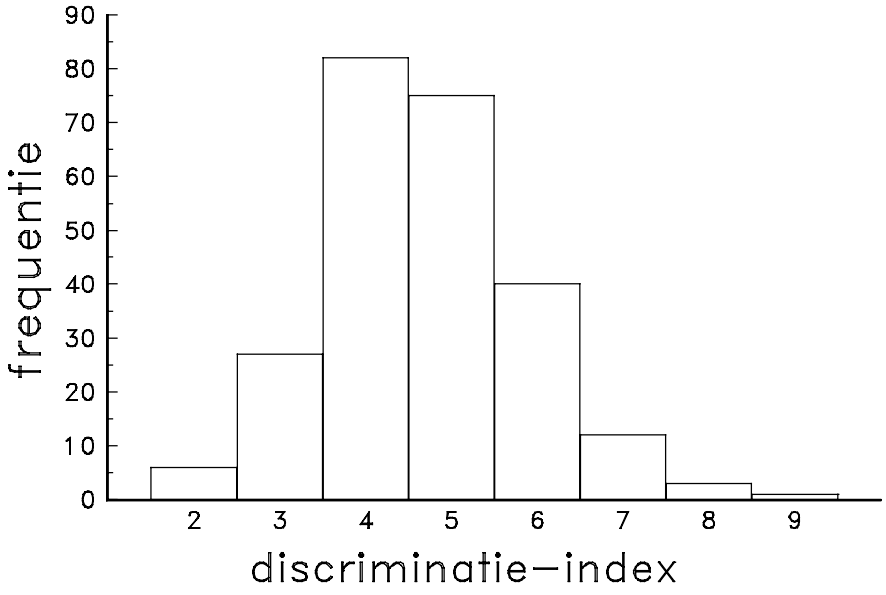


Figuur 7.2
Design van het leesvaardigheidsonderzoek

Net zoals in het PPON-onderzoek vereist een schattingsmethode met MML een vrij gecompliceerd model waarin de designvariabelen, het al dan niet aanbieden van items, gemodelleerd worden. Bovendien treedt hier een extra complicatie op, omdat de steekproeven, overeenkomend met de rijen van figuur 7.2 niet onafhankelijk zijn van elkaar. Verschillende leerlingen namen meer keren aan het onderzoek deel, en deze afhankelijkheid dient gemodelleerd te worden wil men een correcte MML-procedure toepassen. Wordt daarentegen met een CML-procedure gewerkt, dan spelen deze overwegingen geen rol, en ook niet het feit dat leerlingen meermaals aan de test deelnamen. Immers, het is aannemelijk dat na een tussenperiode van een jaar de leesvaardigheid θ veranderd is, en voor het model maakt het niets uit of die twee verschillende θ-waarden afkomstig zijn van één dan wel van twee personen. Voor de verdeling van θ maakt het wel uit: de θ-waarden van twee aselect uit de populatie getrokken personen zijn per definitie onafhankelijk van elkaar, terwijl de θ-waarde van dezelfde persoon op twee verschillende tijdstippen dat niet zijn; dat kunnen we althans niet veronderstellen, anders zou het hele onderzoek zinloos worden.

Het schatten van de itemparameters werd uitgevoerd met het programma OPLM, waarbij de discriminatie-indices een aantal keren werden aangepast. In de uiteindelijke oplossing werden 246 items opgenomen. De verdeling van de discriminatie-indices is afgebeeld in figuur 7.3. Bedenk dat de absolute waarden van deze indices onbelangrijk

zijn, alleen hun onderlinge verhoudingen zeggen iets over het relatieve discriminerende vermogen. Uit de figuur blijkt heel duidelijk dat voor het merendeel van de items de paarsgewijze verhoudingen tamelijk dicht bij 1 liggen, maar toch weer verschillend genoeg zijn om het Raschmodel niet als nulhypothese te kunnen handhaven.

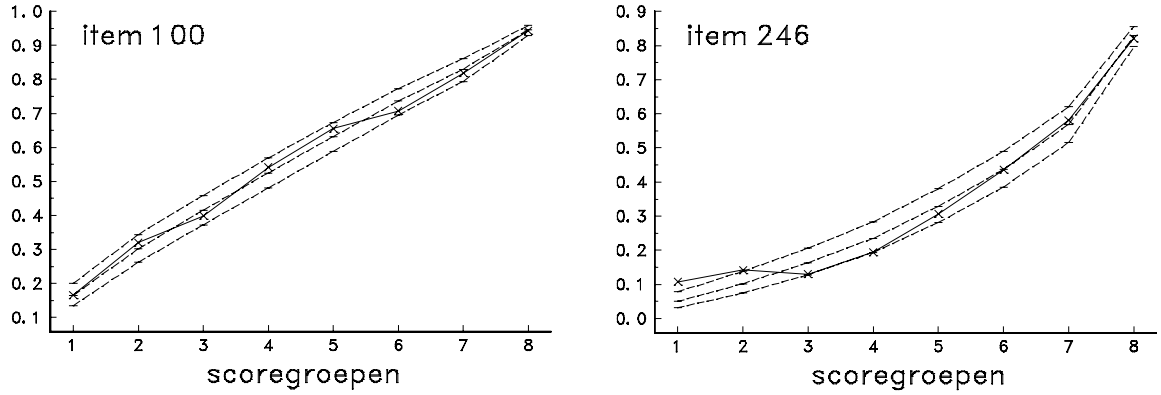


Figuur 7.3
Discriminatie-indices van de 246 items in het leesonderzoek

Om een indruk te geven van de passing van het model, zijn de gegevens waarop de S_i-toetsen gebaseerd zijn, afgebeeld in figuur 7.4 voor twee items. De volle lijnen verbonden door x-symbolen geven de geobserveerde proporties juiste antwoorden weer voor het item, de middelste stippellijn verbindt de voorspelde proporties, en de twee buitenste lijnen geven bij benadering het 95%-betrouwbaarheidsinterval aan. Het item dat links is afgebeeld is een typisch voorbeeld van de meeste items die in de schaal werden opgenomen. Het is bovendien een item dat niet al te moeilijk is: in de hoogst scorende groep is de proportie correcte antwoorden ongeveer 0.9. Het item dat rechts is afgebeeld is het slechtst passende item, en de afbeelding laat meteen ook zien wat de reden van deze slechte passing is. Het is een moeilijk item, en de twee laagst scorende groepen scoren duidelijk hoger dan door het model wordt voorspeld. Dit zou een effect kunnen zijn van het raden bij meerkeuzevragen.

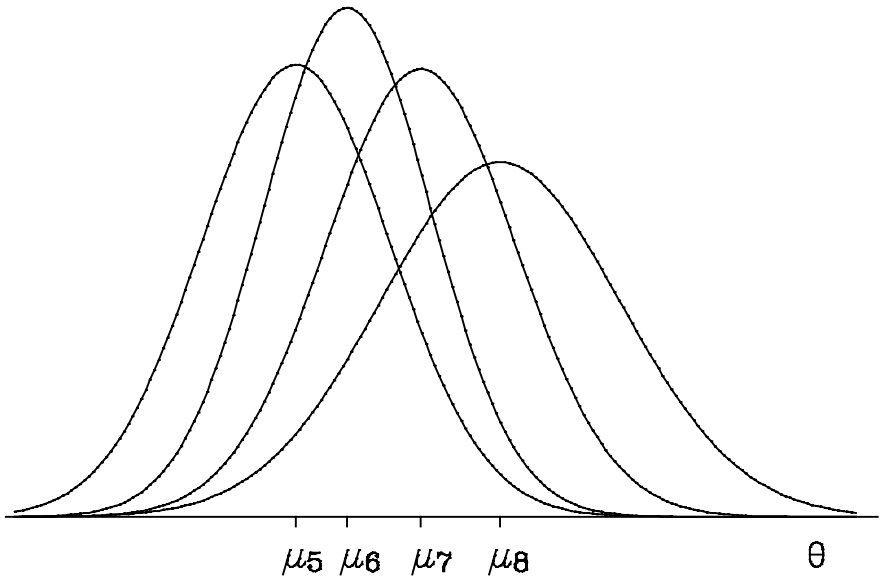
De beoordeling van de algemene modelpassing is een lastig probleem in dit onderzoek: door het zeer groot aantal observaties krijgen de statistische toetsen zeer veel onderscheidend vermogen. Effecten als weergegeven in het rechter gedeelte van figuur 7.4, zelfs als ze veel minder uitgesproken zijn, kunnen bij dergelijke steekproefgrootte gemakkelijk tot significantie aanleiding geven. De procedure van

Hommel die in hoofdstuk 4 is besproken, leidde tot verwerping van het model op het 1%-niveau. Door verwijdering van het slechtst passende item was Hommels toets echter niet significant op het 5%-niveau.



Figuur 7.4
Modelpassing van twee items uit het leesbaarheidsonderzoek

Om een idee te krijgen van de verdeling van de leesvaardigheid in de verschillende jaargroepen werden uit de totale steekproef vier deelsteekproeven gebruikt die representatief konden worden geacht voor de vier onderscheiden populaties, de groepen 5 tot 8. Elke steekproef bevatte ongeveer 1200 leerlingen. In totaal waren er 219 items door de vier deelsteekproeven gemaakt. Op analoge wijze als in paragraaf 7.1 werd beschreven, werden van elke populatie het gemiddelde en de standaardafwijking geschat. Een grafische weergave van de resultaten is gegeven in figuur 7.5.



Figuur 7.5

Verdeling van de leesvaardigheid voor de jaargroepen 5 tot 8

Uit de figuur blijkt zeer duidelijk dat de variabiliteit van de leesvaardigheid groot is in vergelijking met de spreiding tussen de gemiddelden μ_i van de respectievelijke jaargroepen. Dit geeft achteraf gezien een bevestiging van de zinvolheid van het onderzoek: alleen een jaargroep aangeven als indicatie voor de geschiktheid van lectuur negeert de variabiliteit binnen de jaargroepen. De variantie tussen de jaargroepen bedraagt 38% van de totale variantie. Dit betekent dat, indien de jaargroep beschouwd wordt als een maat van lees-vaardigheid, dat wil zeggen een één-item toets, deze een betrouwbaarheid heeft van .38 met betrekking tot de totale populatie van 5- tot 8-jarigen. De uiteindelijk geconstrueerde toetsen (Staphorsius, 1992b) die nu in het onderwijs worden gebruikt, hebben een betrouwbaarheid van boven de .95 met betrekking tot dezelfde populatie, en verklaren dus meer dan 95% van de variabiliteit.

Validiteit

Bij het gebruik van een IRT-model, gaat men uit van bepaalde axioma's, en de statistische toetsen worden gebruikt om de aanvaardbaarheid van deze axioma's te toetsen. Deze toetsen maken dus deel uit van het valideringsonderzoek. Doch daarmee is het valideringsonderzoek natuurlijk niet afgelopen, enerzijds omdat er modelschendingen kunnen zijn die de statistische toetsen niet ontdekken, anderzijds omdat er aspecten zijn aan valideringsonderzoek waarvoor de gebruikelijke statistische modeltoetsen niet geschikt zijn. Er is bijvoorbeeld geen enkele mogelijkheid om uit alleen de leesvaardigheidsdata het besluit te trekken dat de items leesvaardigheid en niet iets anders meten. Voor dit aspect van de validiteit hebben we een extern criterium nodig. We bespreken eerst een bijkomende manier om de geldigheid van het model te controleren, en vervolgens gaan we in op een aspect van de criteriumvaliditeit.

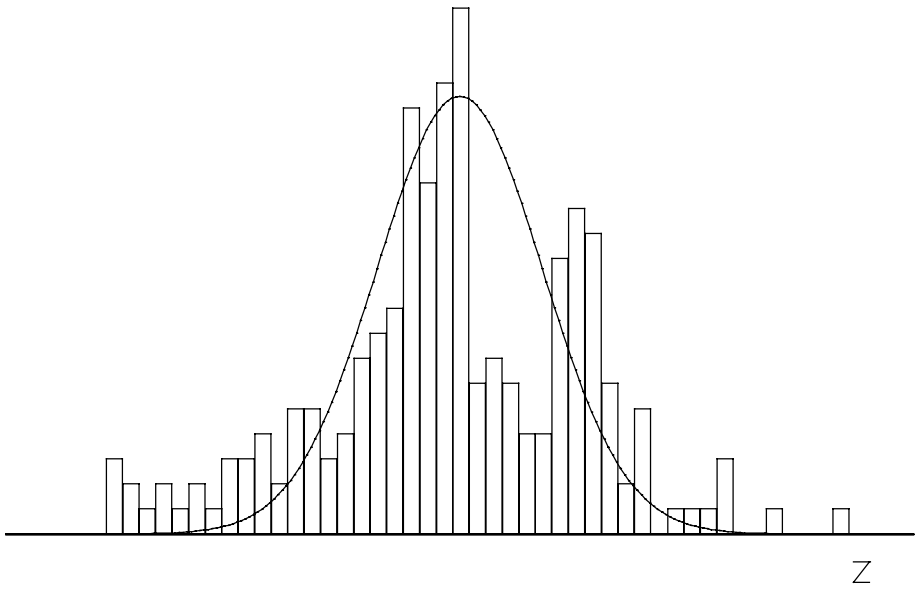
In de klassieke testtheorie wordt de moeilijkheid van een item doorgaans aangegeven met zijn theoretische *p*-waarde, de kans dat het item door een aselect getrokken persoon uit de populatie juist wordt beantwoord. De proportie juiste antwoorden in de steekproef is een schatting van de theoretische *p*-waarde, die we zullen aanduiden als π_i voor item *i*. Indien een IRT-model geldig is, met itemresponsfuncties $f_i(\theta)$, en de verdeling van θ in een bepaalde populatie is gegeven door de dichtheidsfunctie $g(\theta)$, dan geldt dat

$$\pi_i = \int_{-\infty}^{+\infty} f_i(\theta) g(\theta) d(\theta). \qquad (7.6)$$

Zowel $f_i(\theta)$ als $g(\theta)$ is een functie van de modelparameters. Vullen we in die functies nu schattingen van de parameters in, dan is het rechterlid van (7.6) een schatter van π_i, die niet noodzakelijkerwijze precies moet gelijk zijn aan de proportie juiste antwoorden, omdat de data die hier gebruikt worden een deelverzameling zijn van de data waaruit de itemparameters zijn geschat. Maar het verschil tussen beide schatters: $\hat{\pi}_i$, berekend door in het rechterlid van (7.6) de schattingen van de parameters in te vullen, en de geobserveerde proportie p_i, mag niet al te groot zijn, want beide zijn consistente schatters van dezelfde grootheid π_i. Voor alle items die gebruikt werden bij het schatten van de verdelingen in de jaargroepen 5 tot 8 zijn beide grootheden uitgerekend. In figuur 7.6 is het histogram van de gestandaardiseerde afwijkingen

$$z_{(p_i - \hat{\pi}_i)} = \frac{\sqrt{n_i}\,(p_i - \hat{\pi}_i)}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}, \tag{7.7}$$

gegeven, waarbij n_i het aantal personen is dat item i heeft gemaakt. De gestandaardiseerde afwijkingen, gegeven door (7.7) zijn bij benadering normaal verdeeld met gemiddelde 0. De standaardafwijking is echter niet gelijk aan 1, omdat geen rekening is gehouden met het feit dat $\hat{\pi}_i$ niet de werkelijke parameter is, doch een schatting. Omdat de calibratiesteekproef zo groot is, zal het effect van deze fout waarschijnlijk niet al te groot zijn. Het effect van deze verwaarlozing van de schattingsfout maakt dat de gestandaardiseerde afwijkingen gegeven in (7.7) een standaardafwijking hebben die groter is dan 1. Om toch enige indruk te krijgen van de passing van het model is een standaardnormale verdeling bij het histogram getekend.



Figuur 7.6
Gestandaardiseerde afwijkingen tussen geobserveerde en voorspelde proporties

Zelfs al is de standaardafwijking van de theoretische verdeling onderschat, dan blijkt uit de figuur nog heel duidelijk een relatief te groot aantal negatieve z-waarden met grote absolute waarde, terwijl afwijkingen met kleine positieve waarden niet vaak genoeg voorkomen. Een negatieve *z*-waarde betekent dat de voorspelde waarde $\hat{\pi}_i$ groter is dan de geobserveerde proportie p_i. Een verklaring voor dit effect ligt wellicht wederom in raadgedrag als gevolg van het gebruik van meerkeuzevragen. Het item dat in figuur 7.4 rechts is afgebeeld, leverde de kleinste *z*-waarde op ($z = -4.23$). Uit de figuur blijkt het raadgedrag duidelijk bij de twee laagste scoregroepen, doch dit betekent natuurlijk niet dat raadgedrag tot die twee groepen beperkt is gebleven. Men kan geredelijk aannemen dat er ook geraden is, hoewel in mindere mate, in de andere scoregroepen. Bij de schatting van de itemparameters wordt de geobserveerde proportie juist gelijkgesteld aan de voorspelde proportie, dat wil zeggen, het item wordt gemakkelijker geschat dan het werkelijk is, omdat een gedeelte van de juiste antwoorden is toe te schrijven aan raden en niet aan voldoende leesvaardigheid. Dit heeft dan als gevolg dat er een systematische fout in de itemparameterschattingen wordt geïntroduceerd, die op haar beurt doorwerkt in de schatting van de populatieparameters. Of hierin inderdaad een voldoende verklaring ligt voor de afwijkingen is echter niet helemaal duidelijk, en dient onderwerp te zijn van verder onderzoek.

Wij volstaan hier met een algemene beschouwing, die aansluit op wat in hoofdstuk 4 werd gesteld. Het gebruik van het Raschmodel of van een ander model dat CML-schattingen toelaat, heeft het grote voordeel van de zogenaamde steekproefonafhankelijkheid, waarbij het er niet toe doet hoe de steekproef uit de populatie is getrokken. In het onderzoek van Staphorsius is van dit voordeel op grote schaal gebruik gemaakt: de totale steekproef waarop de calibratie is uitgevoerd, getuigt op het eerste gezicht van een soort wildgroei, die elke poging om tot een min of meer realistische beschrijving van de verdeling van θ bij voorbaat tot een hopeloze onderneming maakt. De ingewikkeldheid van het design heeft echter zijn redenen, omdat veel data werden verzameld met andere doeleinden dan alleen het toepassen van een meetmodel. Het verzamelen van herhaalde metingen bij dezelfde personen bijvoorbeeld heeft geleid tot het inpassen van dit onderzoek in het leerlingvolgsysteem dat op het Cito is ontwikkeld. Het grote voordeel van de steekproefonafhankelijkheid kan echter alleen geclaimd worden indien het meetmodel geldig is. Indien meerkeuzevragen gebruikt worden, en er wordt in meer of mindere mate geraden, dan verdwijnt dit voordeel. Zelfs bij redelijk goed uitvallende modeltoetsen, zoals bij de data van Staphorsius, treden er systematische fouten op zodra het model wordt toegepast op populaties die systematisch verschillen van de populatie die bij de

calibratie werd gebruikt, zoals uit figuur 7.6 blijkt. Dit betekent natuurlijk niet dat de onderzoeksgegevens van Staphorsius onbruikbaar zijn. Bij 90% van de items is het absolute verschil tussen geobserveerde en voorspelde *p*-waarde kleiner dan 0.035, en bij 80% is het kleiner dan 0.02. De praktische consequenties zijn tweevoudig: ten eerste kan het toepassen van de geconstrueerde schaal leiden tot een verkeerde schatting van verschillen tussen populaties waar het raadgedrag systematisch gaat verschillen; ten tweede levert het gebruik van meerkeuze-items in modellen die niet voorzien in raadgedrag, dus andere modellen dan bijvoorbeeld het drieparametermodel, bijna automatisch de hierboven beschreven problemen op. Hoewel op het eerste gezicht het gebruik van dit soort ingewikkelder modellen voor de hand schijnt te liggen, is de CML-schattingsmethode hierbij uitgesloten, en is men bij ingewikkelde designs aangewezen op een zeer ingewikkelde modellering van de verdeling van θ, waarbij men zich vaak tevreden zal moeten stellen met benaderingen waarvan het allerminst zeker is of ze een even goede predictie opleveren als in figuur 7.6 is afgebeeld. Een suggestie die vanuit psychometrisch oogpunt voor de hand lijkt te liggen, namelijk afzien van meerkeuze-items, lijkt de oplossing van het probleem te zijn. Voor de praktische haalbaarheid van deze oplossing zal het oordeel van de veldonderzoeker wellicht zwaarder moeten wegen dan een suggestie uit de psychometrie.

Voor het tweede onderdeel van de validiteitsstudie, namelijk de relatie met externe variabelen, beperken we ons tot één gedeelte uit het onderzoek van Staphorsius. Indien de teksttoets dezelfde vaardigheid meet als een cloze-toets, dan bestaat de voor de hand liggende controle erin, de teksten van de teksttoetsen te 'be-clozen' en het verband na te gaan tussen individuele cloze-scores en de geschatte vaardigheid θ die door de teksttoets wordt gemeten. De dataverzameling voor dit doel is begonnen, doch bij het schrijven van dit hoofdstuk waren de resultaten nog niet beschikbaar. Toch kunnen we indirecte evidentie voor dit verband krijgen door de $\hat{\pi}_i$-waarden die met (7.6) te berekenen zijn, te beschouwen als 'proxies' voor de cloze-scores. Van alle 246 items werd de gemiddelde $\hat{\pi}_i$-waarde berekend over de jaar-groepen 5 tot 8. Om de overeenkomst met de cloze-procedure te bevorderen, werden de $\hat{\pi}_i$- waarden van items die tot dezelfde tekst behoren, gemiddeld en beschouwd als 'proxy' voor de cloze-scores. Indien de teksttoets dezelfde vaardigheid meet als de cloze-score, dan moet de voorspelling van de gemiddelde $\hat{\pi}_i$-waarden uit formele tekstkenmerken goed overeenkomen met de CLIB-waarde van die teksten. De multipele correlatie tussen de gemiddelde $\hat{\pi}_i$-waarden en formele tekstkenmerken bedroeg .967. Het feit dat deze correlatie hoger is dan de correlatie tussen deze formele tekstkenmerken en de gemiddelde cloze-scores, is voor een deel te verklaren uit het feit dat de gemiddelde $\hat{\pi}_i$-waarden een grotere spreiding vertonen dan de gemiddelde cloze-scores. Bovendien

waren de teksten waarop de cloze-scores zijn bepaald, een steekproef uit bestaande teksten, waarvan sommige zeer specifieke kennis vereisten en zodoende de cloze-score drukten. Bij het formuleren van de teksttoetsen daar-entegen was veel zorg besteed om de antwoorden zoveel mogelijk onafhankelijk te maken van specifieke kennis of informatie die niet in de tekst gegeven was. De hoge correlaties tussen enerzijds cloze-score en formele tekstkenmerken, en anderzijds tussen gemiddelde $\hat{\pi}_i$-waarden en formele tekstkenmerken, impliceren een hoge correlatie tussen gemiddelde cloze-score en gemiddelde $\hat{\pi}_i$-waarden. De correlatie tussen de voorspelde waarde van de gemiddelde $\hat{\pi}_i$-waarden en de CLIB bedroeg 0.987.

De correlatie tussen individuele cloze-scores en de geschatte θ-waarde zal ongetwijfeld lager uitvallen; maar niettemin zijn deze resultaten duidelijke evidentie dat teksttoetsen en cloze-toetsen dezelfde vaardigheid aanspreken.

Het verband tussen leesvaardigheid en leesbaarheid

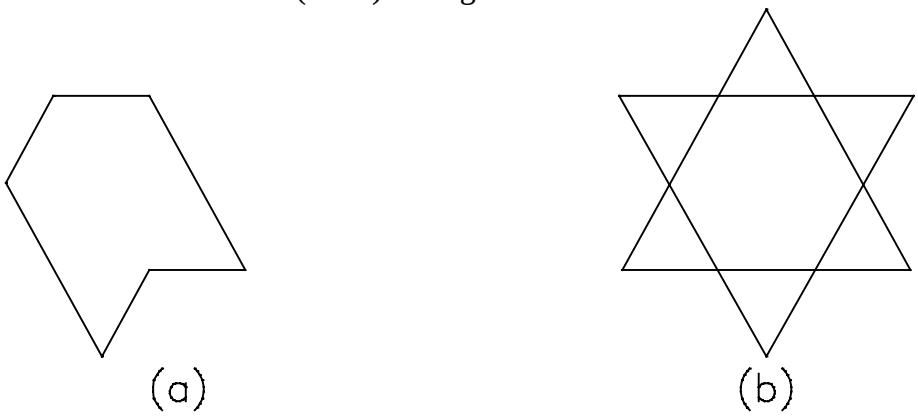
Het hierboven beschreven valideringsonderzoek levert ook de sleutel om leesbaarheid en leesvaardigheid op eenzelfde schaal te brengen. Voor een tekst T uit de teksttoets die bestaat uit zes items kunnen we voor een willekeurige waarde van θ de verwachte gestandaardiseerde score berekenen met de formule

$$\mathscr{E}(X_T) = \frac{\sum_{i \in T} a_i f_i(\theta)}{\sum_{i \in T} a_i}. \qquad (7.8)$$

Stellen we nu dat beheersing van de tekst gelijk staat met een gestandaardiseerde verwachte score van minstens *c* (bijvoorbeeld 0.7), dan kan in het rechterlid van (7.8) θ zo bepaald worden dat de verwachte score gelijk is aan *c*. We duiden deze waarde aan als θ_c. Uit de zeer hoge correlatie tussen de gemiddelde $\hat{\pi}_i$-waarden en de CLIB volgt dat de CLIB-waarde voor deze tekst in de populatie van personen met $\theta = \theta_c$ ongeveer gelijk zal zijn aan *c*. Omgekeerd -en in de mate dat het verband tussen CLIB en leesvaardigheidstoets te veralgemenen is- volgt dat een tekst met CLIB-waarde gelijk aan *c*, begrepen wordt door personen met een θ-waarde groter θ_c. Kennen we de θ-waarde van een persoon en de CLIB-waarde van een tekst, dan hebben we een rationele grond om te beslissen of de tekst al dan niet voor die persoon geschikt is. Omdat θ geschat moet worden, wordt de schatting natuurlijk niet gebaseerd op één tekst met zes items, maar op een teksttoets van redelijke lengte, zodat de meetfout (dit is de schattingsfout van $\hat{\theta}$) voldoende klein wordt gehouden.

7.3 De diagnostische verborgen-figurentest

Binnen de cognitieve psychologie worden trainingsprogramma's opgesteld om het cognitieve functioneren te beïnvloeden en om eventuele achterstanden weg te werken. Het 'Instrumental Enrichment'-programma van Feuerstein (1980) neemt hier een leidende positie in. Het programma bestaat uit 14 instrumenten die voornamelijk oefeningen in de vorm van testfiguren bevatten. Het is de bedoeling om via deze training de cognitieve capaciteiten en het algemene leervermogen van adolescenten te verhogen. Een van de instrumenten die Feuerstein gebruikte om zijn programma te evalueren is de verborgen-figurentest (Embedded Figures Test, verder afgekort als EFT), ontwikkeld door Witkin (1950). In figuur 7.7 is een item uit deze test afgebeeld.



Figuur 7.7
Voorbeeld van een verborgen-figuren opgave

De eenvoudige figuur (*a*) zit verborgen in het complexe patroon (*b*). Bij toepassing van Witkins test wordt aan de persoon eerst gevraagd het complexe patroon te beschrijven; daarna moet de eenvoudige figuur gememoriseerd worden, en tenslotte moet aangewezen worden waar de eenvoudige figuur in het complexe patroon verborgen zit. De antwoordtijd en de correctheid van het antwoord worden genoteerd.

Uit de evaluatiestudie bleek dat de personen die het 'Instrumental Enrichment' programma hadden gevolgd, gemiddeld sneller antwoordden en meer juiste antwoorden gaven dan een controlegroep die een minder specifiek trainingsprogramma had gevolgd. Bradley (1983) betoogde echter dat uit dit resultaat niet volgt dat door het trainingsprogramma cognitieve strategieën gewijzigd kunnen worden. Immers, uit de

verschillen in antwoordtijd en aantal items juist volgt niet automatisch dat er andere cognitieve strategieën gebruikt worden in de twee condities. Het probleem met de interpretatie van de EFT wordt bijvoorbeeld duidelijk geïllustreerd door de vele theoretische interpretaties die Witkin zelf en anderen aan de test hebben gegeven (Witkin & Goodenough, 1981; Pennings, 1991). In meer algemene termen gesteld, betekent dit dus dat er problemen zijn met de constructvaliditeit van de EFT. Het is niet zonder meer duidelijk wat de EFT eigenlijk meet. Op basis van een theoretische studie over de gebruikte strategieën in de EFT, kwam Pennings (1988) tot de volgende conclusies:

(1) Zeer korte antwoordtijden komen tot stand door het gebruiken van een simultane (ook genoemd holistische, synthetische of figuratieve) strategie, waarbij vorm, grootte en positie van de eenvoudige figuur als geheel in gedachten worden gehouden bij het bekijken van het complexe patroon. Het antwoord komt tot stand door een 'matching' van deze voorstelling met een gedeelte van het complexe patroon;

(2) middellange antwoordtijden resulteren bij gebruik van een successieve (analytische) strategie, waarbij onderdelen van de eenvoudige figuur (bijvoorbeeld een lijnstuk) successievelijk opgezocht worden in het complexe patroon;

(3) als de antwoordtijden, bij volwassenen en adolescenten, heel lang worden, kan toch een oplossing gevonden worden door het externaliseren van oplossingsoperaties, zoals het volgen van bepaalde lijnstukken met een aanwijsstokje op het complexe patroon;

(4) wanneer kinderen de items erg moeilijk vinden, vinden ze toch vaak de oplossing als ze een doorzichtig figuurtje in de vorm van de eenvoudige figuur mogen manipuleren over het complexe patroon. Dit wordt aangeduid als een globaal-manipulatorische strategie.

Deze vier genoemde strategieën komen bovendien overeen met een ontwikkelingslijn in de cognitieve ontwikkeling van kinderen: van een globaal-manipulatorische strategie, die helemaal extern is, naar een geïnternaliseerde strategie die verloopt van successieve en gecontroleerde operaties naar simultaan en geautomatiseerd. De vier beschreven strategieën in de volgorde (4) tot (1) weerspiegelen dus ook de chronologische ontwikkeling in het normale functioneren van een kind.

Om deze strategieën meer zichtbaar te maken dan door de pure tijdopname in de EFT, ontwikkelde Pennings een variant, het Verborgen-Figuren Diagnosticum genaamd. Daarbij wordt eenzelfde soort items gebruikt als in de EFT, doch de wijze van afname en de scoring is verschillend. De algemene procedure is een 'antwoord-totdat-juist' procedure:

(1) een juist antwoord binnen vijf seconden wordt geïnterpreteerd als evidentie voor een (succesvolle) simultane strategie, en levert een score op van vier punten;

(2) bij geen of een fout antwoord onder conditie (1), krijgt de proefpersoon speciale instructie om een successieve strategie te gebruiken. Een juist antwoord binnen de 55 seconden levert drie punten op;

(3) indien (2) niet succesvol is, krijgt de proefpersoon staafjes die in lengte overeenkomen met de lijnstukken van de eenvoudige figuur, die op het complexe patroon kunnen worden neergelegd om de eenvoudige figuur te vormen (maximale tijd 75 seconden). Succes levert een score van twee punten op;

(4) indien nog steeds geen oplossing is gevonden, kan de proefpersoon manipuleren met een doorzichtig perspex model van de eenvoudige figuur (maximale tijd 45 seconden). Een goed antwoord levert één punt op. Lukt het niet binnen de maximaal toegestane tijd dan is de itemscore nul punten.

De belangrijkste vraag met betrekking tot de constructvaliditeit van het aldus geconstrueerde meetinstrument is of deze scoringsregel zinvol is: bestaat er een abstract unidimensionaal begrip θ, zodat een grotere waarde van θ een hogere verwachte score betekent op elk item in de test. Een geschikt model om deze vraag te beantwoorden is OPLM voor polytome data (zie hoofdstuk 5).

De data waren afkomstig van 480 kinderen, 30 jongens en 30 meisjes in de leeftijd van 5, 6, 7, 8, 9, 10, 11 en 12 jaar. De test bevat zes items en de resultaten van de CML-schattings- en toetsingsprocedure zijn weergegeven in tabel 7.3. Hoewel de passing van het model niet overweldigend is, is er ook geen duidelijke evidentie om het model te verwerpen. De conclusie dat de scoringsregel zinvol is, wordt door deze analyse dus goeddeels gesteund.

Het tweede aspect van de hypothese, namelijk dat θ de individuele ontwikkeling weerspiegelt, kan gevalideerd worden door het verband tussen de leeftijd van de proefpersonen en θ te onderzoeken. Op dezelfde wijze als in paragraaf 7.1 wordt een lineair model gespecificeerd voor de latente variabele θ:

$$\theta_{vjk} = \mu + \alpha_j + \beta_k + \varepsilon_{vjk} \qquad (7.9)$$

waarin het residu ε_{vjk} normaal verdeeld is met gemiddelde nul en gemeenschappelijke variantie σ^2. Hoewel leeftijd een continue variabele is, werd de totale groep opgesplitst in vier leeftijdscategorieën: 1 = 5-6 jaar; 2 = 7-8 jaar; 3 = 9-10 jaar en 4 = 11-12 jaar.

Tabel 7.3

Parameterschattingen en toetsen voor de diagnostische EFT

Item	Cat.	*a*	β	*SE*(β)	*S*	*vg*	*p*	*M*	*M*2	*M*3

1	1	4	-.931	.085	---	-	---	3.17	-.09	-.30
	2		-.275	.046	1.41	3	.702	1.44	-.02	.26
	3		-.104	.035	5.70	4	.222	-2.12	-1.27	-1.89
	4		.582	.040	2.54	3	.467	-1.37	-1.91	-.86
2	1	3	-.815	.093	---	-	---	-1.49	-.30	-.68
	2		-.459	.060	7.38	3	.061	-1.49	.03	.03
	3		-.035	.045	1.65	5	.895	-.87	-.36	-.95
	4		.317	.044	13.06	5	.023	.01	1.41	-.68
3	1	2	-.398	.100	.42	3	.937	.72	.30	.22
	2		-.336	.082	4.74	5	.448	.61	2.02	1.74
	3		.149	.072	8.41	6	.209	1.28	.49	1.80
	4		.271	.074	3.39	5	.640	-1.51	-.97	-1.56
4	1	3	-.697	.073	.12	1	.730	.98	-.66	-.62
	2		-.126	.054	9.01	4	.061	2.70	2.44	2.84
	3		-.130	.045	3.70	5	.594	.14	-.05	.56
	4		.797	.057	1.28	3	.734	.37	.00	.86
5	1	3	-.507	.053	4.32	3	.229	-2.22	-.12	-.90
	2		.147	.043	2.91	5	.714	.72	.63	.85
	3		.407	.050	9.46	4	.051	.40	2.56	1.65
	4		1.082	.108	---	-	---	-.86	4.52	2.63
6	1	4	-.288	.043	1.25	3	.742	-.07	-.63	.89
	2		-.009	.037	4.35	4	.361	-2.46	-1.51	-2.43
	3		.344	.037	3.79	4	.435	-.21	-.58	.00
	4		1.016	.088	---	-	---	.01	-.21	-.57
R_{1c} = 85.80 (*vg* = 67; *p* =.061)										

De effecten van de leeftijdscategorieën worden weergegeven door de parameters β_k. Omdat Witkin ook verschillen tussen jongens en meisjes rapporteert voor de EFT, werd geslacht als tweede achtergrondvariabele meegenomen. De effectparameters zijn α_j (1 = 'jongen', 2 = 'meisje'). De resultaten zijn weergegeven in tabel 7.4. De schaal waarop de resultaten zijn gerapporteerd is zo geconstrueerd dat de som van de categorieparameters gelijk is aan 0 en het produkt van de discriminatie-indices gelijk is aan 1. De analysemethode is identiek aan de methode beschreven in paragraaf 7.1.

Tabel 7.4

Effectschattingen van het onderzoek met de diagnostische EFT

Parameter	Schatting	Stand. fout (*SE*)	z=schatting/*SE*
σ^2	0.54		
μ	-1.50	0.14	-10.77
α_1	0	---	---
α_2	0.12	0.09	1.26
β_1	0	---	---
β_2	1.56	0.14	10.96
β_3	2.00	0.14	14.00
β_4	2.62	0.14	18.12

De binnengroeps-standaardafwijking, σ, is gelijk aan $\sqrt{.54} = 0.735$. Het verschil tussen de tweede leeftijdsgroep en de referentiegroep (de jongste kinderen), $\beta_2 - \beta_1$, bedraagt dus meer dan twee maal de binnengroeps-standaardafwijking, terwijl de verandering van de tweede naar de volgende leeftijdsgroepen veel minder sterk uitgesproken is. De resultaten van deze analyse bevestigen dus zeer duidelijk de hypothese dat θ de individuele ontwikkeling weerspiegelt.

8

Equivaleren

Een leerling van het VWO doet een herexamen (tweede tijdvak) voor het vak natuurkunde en behaalt een hogere score dan tijdens het reguliere examen (eerste tijdvak). Waarom? We zouden kunnen concluderen dat deze hogere score een grotere vaardigheid weerspiegelt: de leerling heeft tussen de beide examens flink wat bijgeleerd. Aan de andere kant is het mogelijk dat het examen uit het tweede tijdvak gemakkelijker was dan dat uit het eerste. Zelfs bij een gelijk gebleven vaardigheid zou de leerling dan een hogere score behalen. Gezien het grote belang dat examens hebben, is het duidelijk dat de leerling een score moet krijgen die een zo goed mogelijke afspiegeling van haar of zijn vaardigheid is, ongeacht welk examen gemaakt is. Dit betekent in ieder geval dat voor iedere score op het tweede tijdvak een score op het eerste tijdvak gevonden moet worden die dezelfde vaardigheid representeert. Het zoeken van vergelijkbare scores is een voorbeeld van wat men equivaleren noemt.

De psychometrische theorie over equivaleren is zeer omvangrijk. Voor overzichten verwijzen we naar Angoff (1971), Holland en Rubin (1982) en Petersen, Kolen en Hoover (1989). In dit hoofdstuk zullen wij ons zoveel mogelijk beperken tot het behandelen van equivaleermethoden die in de praktijk veelvuldig gebruikt worden. De belangrijkste factor die bepalend is voor de wijze waarop de equivalering plaatsvindt is het gebruikte meetmodel. Zoals we gezien hebben in de hoofdstukken 3, 4 en 5 heeft elk model zijn eigen manier om met een toets de vaardigheid te bepalen. Voor de bepaling van de vaardigheid gebruiken we in de klassieke testtheorie (KTT) doorgaans geobserveerde scores op een toets, terwijl in de itemresponstheorie (IRT) de vaardigheid als parameter, die in het model is opgenomen, geschat wordt. Alvorens echter het equivaleren per meetmodel te bespreken, zullen we in paragraaf 8.1 eerst een globaal overzicht geven van het equivaleren. Aspecten die daarbij aan de orde zullen komen spelen zowel een rol bij equivaleren in de KTT als in de IRT. In paragraaf 8.2 gaan we vervolgens de equivalering in de KTT behandelen. In paragraaf 8.3 volgt equivaleren in de IRT. In de laatste paragraaf 8.4 worden de conclusies en aanbevelingen uit dit hoofdstuk kort samengevat.

8.1 Overzicht equivaleren

Zoals uit de inleiding blijkt, ontstaat de behoefte aan equivaleren als we de vaardigheid van twee personen met een verschillend meetinstrument meten en de resultaten met elkaar willen vergelijken. De eerste vraag die we hierbij zouden moeten beantwoorden is of equivaleren in de praktijk niet vermeden kan worden. Men zou kunnen denken dat in het voorbeeld uit de inleiding geen problemen waren ontstaan als het examen van het tweede tijdvak hetzelfde geweest was als dat van het eerste tijdvak. Omdat de examens identiek zijn, zullen ook de scores op beide examens gelijk dezelfde vaardigheid weerspiegelen. Het is maar al te duidelijk dat we niet op deze manier te werk kunnen gaan. Leerlingen die tijdens het tweede tijdvak examen doen, zijn dan bevoordeeld daar zij de inhoud van het af te nemen examen reeds kennen. Daarom, op grond van eerlijkheid, is het noodzakelijk om het herexamen verschillend van het eerste te laten zijn. Om de scores van een leerling, of meer algemeen voor verschillende leerlingen, op twee verschillende examens op een zinvolle manier met elkaar te kunnen vergelijken, zal men dus rekening moeten houden met de, mogelijk verschillende, moeilijkheid van beide examens. Het is immers onterecht als een tweede tijdvak kandidaat een hoger cijfer haalt dan een eerste tijdvak kandidaat, alleen maar omdat zij of hij een eenvoudiger examen gemaakt heeft.

Het ideaal van het vermijden van equivaleren wordt in zekere zin bereikt, zoals we later zullen zien, als we toetsen samenstellen uit een itembank die gecalibreerd is onder een IRT-model. In de praktijk is evenwel meestal het equivaleerprobleem aan de orde als we de scores op twee bestaande, vaste, toetsen vergelijkbaar willen maken. Overigens is in de KTT een andere werkwijze ook niet mogelijk, omdat we daar altijd uitgaan van de score op een toets. We zullen in dit hoofdstuk het equivaleerprobleem dan ook via deze weg benaderen.

Meer algemeen gesteld zouden we het probleem van het equivaleren als volgt kunnen omschrijven. Twee of meer groepen personen maken verschillende versies van een toets. Hoe kunnen de scores op de ene toets vertaald of naar een zelfde schaal getransformeerd worden als de scores op de andere toets, opdat ze vergelijkbaar worden? Het zal blijken dat het equivaleren van twee toetsen in feite neerkomt op het vinden van een functie die de scores op een toets Y transformeert naar de schaal van de scores op een toets X. Deze functie, die we de equivaleerfunctie noemen, noteren we met $e_X(Y)$. Het zal duidelijk zijn dat als we twee toetsen kunnen equivaleren, we ook meer toetsen kunnen equivaleren. In dit hoofdstuk zullen we dan ook steeds spreken over het equivaleren van twee toetsen.

We kunnen stellen dat de vergelijking van de scores op twee toetsen niet mag afhangen van wie welke toets heeft gemaakt. De score van een persoon op een toets zal echter afhangen van de moeilijkheid van de voorgelegde toets. Ook de twee situaties waarin de toetsen werden afgenomen mag op de vergelijking niet van invloed zijn. De score op een toets kan immers ook afhangen van externe factoren, zoals lawaai of extreme hitte tijdens de afname. Helaas zijn deze laatste effecten in de toetspraktijk vaak aanwezig. Alhoewel het soms mogelijk is om voor een lagere score tengevolge van externe factoren te corrigeren, zullen we ons hier in dit hoofdstuk niet mee bezig houden. Als we spreken over equivaleren dan willen we alleen corrigeren voor verschil in moeilijkheid.

In de praktijk kunnen we twee situaties onderscheiden waarin we willen equivaleren. In de eerste plaats is dat de situatie waarin we willen corrigeren voor niet geplande verschillen tussen de toetsen. Bij deze zogenaamde horizontale equivalering gaan we ervan uit dat we twee toetsen hebben die in principe hetzelfde meten, van dezelfde moeilijkheidsgraad zijn en bedoeld zijn voor één populatie. In deze situatie willen we dus onbedoelde ruis in metingen wegwerken. Deze ruis kan ontstaan doordat het bijvoorbeeld niet geheel gelukt is twee even moeilijke toetsen te maken. Het kan ook voorkomen dat de groepen leerlingen die de toetsen maken toch op de een of andere manier een weinig in vaardigheid verschillen. Een voorbeeld waar horizontale equivalering wordt toegepast is de Eindtoets Basisonderwijs van het Cito (in het vervolg Eindtoets). De Eindtoets, welke bestaat uit de drie onderdelen taal, rekenen en informatieverwerking, is een schoolvorderingentoets die jaarlijks wordt afgenomen in groep 8 van de basisschool. Deze toets heeft twee functies. Enerzijds levert de Eindtoets informatie over individuele leerlingen in verband met de overgang naar het voortgezet onderwijs, anderzijds levert de toets informatie ten behoeve van de evaluatie van het gegeven onderwijs (Uiterwijk & Engelen, 1993). Bij de constructie van een nieuwe versie van deze toets wordt er, onder andere, expliciet naar gestreefd om deze dezelfde moeilijkheidsgraad te geven als de oudere versie. Bovendien valt het te verwachten dat de groepen leerlingen die de Eindtoets maken, steeds leerlingen uit groep 8 van het basisonderwijs, van jaar tot jaar niet al te veel in vaardigheid zullen verschillen. Een ander voorbeeld, waarbij we horizontaal willen equivaleren, zijn de eindexamens van het eerste en het tweede tijdvak.

De tweede situatie waarin we zouden willen equivaleren is die waarbij we de prestaties op twee toetsen willen vergelijken die een verschillende moeilijkheidsgraad hebben en dan ook bedoeld zijn voor groepen met verschillende vaardigheidsniveaus. Bij deze zogenaamde verticale equivalering willen we dus corrigeren voor reeds vooraf geplande verschillen in moeilijkheidsgraad tussen de toetsen. Als we bijvoorbeeld

Mavo-C en Mavo-D examens willen equivaleren, dan hebben we te maken met verticale equivalering. Immers, het Mavo-D examen is getracht moeilijker te maken dan het Mavo-C examen terwijl ook de populaties leerlingen in vaardigheid zullen verschillen.

Gezien de extra complicaties (ongelijke moeilijkheid en vaardigheden) zal het duidelijk zijn dat verticaal equivaleren in het algemeen problematischer zal verlopen dan horizontaal equivaleren. Historisch gezien is de theorie van het equivaleren dan ook ontwikkeld voor de situatie waarin we horizontaal willen equivaleren; verticaal equivaleren is pas later ontstaan. Alhoewel er ook binnen het kader van de KTT al enige aandacht aan wordt besteed, is toepassing van verticaal equivaleren eigenlijk pas goed mogelijk als we met IRT werken. We komen hier later nog op terug. In paragraaf 8.1.1 geven we een beknopt overzicht van de psychometrische voorwaarden die in de loop der tijd aan equivalering zijn gesteld. We willen hier reeds opmerken dat in de praktijk niet strikt aan deze voorwaarden wordt vastgehouden. Voor de volledigheid en voor een beter begrip van het equivaleerprobleem worden ze hier toch besproken. Vervolgens bespreken we in paragraaf 8.1.2 de eerste stap van elk equivaleerprobleem: volgens welk design moeten de gegevens die nodig zijn voor het equivaleren, verzameld worden?

8.1.1 Psychometrische voorwaarden voor equivaleren

We kunnen equivaleren als een psychometrisch, maar ook als een statistisch probleem opvatten. We zullen uitleggen wat we hiermee bedoelen. Laten we eerst maar eens aannemen dat we aan een statisticus zonder kennis van de psychometrie vragen om twee toetsen te equivaleren. Daar deze statisticus geen notie van het begrip ware score heeft, is voor hem alleen maar de geobserveerde score van belang. Equivaleren betekent voor hem het zoeken van een relatie tussen de geobserveerde scores van de twee toetsen. Om deze relatie te vinden zal hij bepaalde statistische aannames moeten maken, zoals bijvoorbeeld de aanname dat de geobserveerde scores normaal verdeeld zijn. Vervolgens gebruikt hij een of andere statistische methode om de functionele dan wel structurele relatie tussen de geobserveerde scores vast te leggen. Hoe dit alles precies in zijn werk gaat, is hier niet van belang. De gevolgde werkwijze van de statisticus zullen we statistisch equivaleren noemen. Het moge duidelijk zijn dat equivaleren op deze manier een relatief eenvoudige empirische procedure geworden is: alleen de data en de statistiek zijn hier van belang. De psychometrie wordt in het geheel niet gebruikt. Statistisch equivaleren zoals hierboven beschreven, legt geen

enkele psychometrische restrictie aan de toetsen op. De twee toetsen zouden bijvoorbeeld verschillende betrouwbaarheden kunnen hebben of zelfs verschillende vaardigheden kunnen meten. Als we spreken over (psychometrisch) equivaleren, zullen we dus altijd de psychometrie op de een of andere manier in het verhaal moeten betrekken. Het zal dan ook blijken dat het noodzakelijk is om psychometrische voorwaarden op te leggen aan de te equivaleren toetsen. Bovendien zal blijken dat ook de equivaleerfunctie aan bepaalde voorwaarden moet voldoen. De rest van deze paragraaf zal een beschrijving van deze voorwaarden geven.

Voordat we echter een beschrijving van deze eisen geven, willen we eerst een opmerking maken. Bij het equivaleren van twee toetsen is het, zoals later zal blijken, van groot belang om de betrokken populatie(s) goed te definiëren. De belangrijkste reden hiervoor is dat ook de gebruikte meetmodellen, de KTT en de IRT, altijd met een (of meer) populaties werken. Zo is bijvoorbeeld de betrouwbaarheid van een toets in de KTT populatie-afhankelijk. We komen hier later nog op terug.

Theoretische overwegingen (Angoff, 1971) leiden tot de volgende, vrij algemeen aanvaarde, vier voorwaarden of eisen (Petersen e.a., 1989) met betrekking tot het equivaleren van twee toetsen:

(1) De toetsen moeten dezelfde vaardigheid meten.
(2) De geëquivaleerde scores op de twee toetsen moeten uitwisselbaar zijn.
(3) De equivaleerfunctie moet invariant over groepen personen zijn.
(4) De equivalering moet symmetrisch zijn.

We zullen aangeven wat deze theoretische eisen voor de praktijk van het equivaleren betekenen.

De eerste voorwaarde kan gezien worden als een gezond verstand voorwaarde. Hierbij kunnen we opmerken dat het geen enkele zin heeft om een toets engels met een toets natuurkunde te equivaleren. Dit zou namelijk kunnen leiden tot uitspraken zoals Piets vaardigheid in engels is even groot als Jans natuurkunde vaardigheid. Bij equivaleren met behulp van de KTT zijn er verschillende mogelijkheden om aan de eerste voorwaarde te voldoen. De zwakst mogelijke is die waarbij we eisen dat de twee toetsen congeneriek zijn; de sterkste is die van parallelliteit. Voor meer informatie omtrent de begrippen congeneriek en parallelliteit verwijzen we naar paragraaf 3.6.1 (zie ook tabel 3.1). Op dit moment volstaat de opmerking dat naarmate de voorwaarden die we stellen aan de te equivaleren toetsen strenger worden, de equivalering van de toetsen eenvoudiger en beter wordt. Immers, als de eisen die we stellen om over dezelfde vaardigheid te kunnen spreken sterker worden, gaan de toetsen meer op elkaar lijken: de toetsen zelf worden dan al meer 'equivalent'. Bij equivaleren met behulp van de IRT dient de eerste eis, strikt genomen, vervangen te worden door de

sterkere eis van unidimensionaliteit. We verwijzen voor de betekenis hiervan naar paragraaf 4.3.1. De laatste jaren zijn er echter ook voor meerdimensionale IRT-modellen equivaleermethoden ontwikkeld. Daar deze methoden nooit aan de unidimensionaliteitseis kunnen voldoen, zullen we deze 'quasi-equivalering' noemen. Een voorbeeld hiervan zullen we bespreken in paragraaf 8.3.4.

De tweede voorwaarde, de uitwisselbaarheid van de scores, ook wel de rechtvaardigheidseis genoemd, is oorspronkelijk geformuleerd door Angoff (1971), die er de volgende inhoud aan gaf. Het mag voor personen niet uitmaken welke van de twee geëquivaleerde scores gebruikt worden, bijvoorbeeld om een zak/slaag beslissing te nemen. Angoff werkte in het kader van de KTT en stelde vast dat deze voorwaarde noodzakelijkerwijs parallelliteit van de toetsen veronderstelt. Angoff neemt dus daarmee ook de sterkst mogelijke versie van de eerste eis aan. Maar dat zou betekenen dat we alleen maar parallelle toetsen kunnen equivaleren. Daarom is deze strikte voorwaarde door hem afgezwakt tot het even betrouwbaar zijn van de toetsen.

Lord (1980) heeft de rechtvaardigheidseis voor equivalering met behulp van de IRT gepreciseerd als: twee toetsen X en Y zijn uitwisselbaar of sterk equivalent als geen enkele persoon, met een gegeven vaardigheid, een reden heeft om de ene boven de andere toets te prefereren. Het moge duidelijk zijn dat sterk equivalente toetsen het ideaal is. Dat de constructie van sterk equivalente toetsen echter veelal onmogelijk zal zijn kunnen we eenvoudig aantonen. Beschouw daartoe twee toetsen die elk slechts één item bevatten. Willen deze toetsen sterk equivalent zijn, dan moet voor elke willekeurig gekozen persoon de kans op een goed antwoord voor beide items precies gelijk zijn. Maar dit betekent dat de beide items dezelfde itemparameters moeten hebben, ze moeten dus even moeilijk zijn. In het algemeen zal dus gelden dat twee willekeurige toetsen dan en slechts dan sterk equivalent zijn, als er voor elk item uit de ene toets een item uit de andere toets te vinden is dat gelijke itemparameters heeft en omgekeerd. We zien dan gelijk dat een noodzakelijke voorwaarde hiervoor is dat de toetsen ook precies even lang moeten zijn. De eigenschap dat er voor elk item uit de ene toets een 'vergelijkbaar' item uit de andere toets gevonden kan worden, is wat Samejima (1977) het sterk parallel zijn van twee toetsen noemt. Uit de praktijk blijkt dat het vrijwel onmogelijk is om sterk parallelle (equivalente) toetsen te construeren. Deze observatie heeft dan ook geleid (Divgi, 1981 en Yen, 1983) tot een afzwakking van Lords rechtvaardigheidseis tot: twee toetsen zijn zwak geëquivaleerd als elke persoon in de populatie dezelfde verwachte score op beide toetsen heeft. Merk op dat de gebruikte begrippen sterk en zwak logische benamingen zijn. Uit de definities volgt immers eenvoudig dat sterk geëquivaleerde (sterk parallelle) toetsen ook zwak geëquivaleerd zijn. De bovenstaande overwegingen zijn strikt genomen alleen voor het

equivaleren met behulp van de IRT geldig. Omdat de KTT gezien kan worden als een speciaal geval van de IRT (Lord, 1980), wordt er vaak beweerd dat parallelliteit, maar dan in de KTT betekenis, ook voor equivaleren met behulp van de KTT moet gelden. Maar de KTT houdt zich niet bezig met items, doch met toetsscores zodat het voorgaande zeer de vraag is. Bovendien is het zo dat als we de rechtvaardigheidseis zo strikt zouden nemen als Lord, we voor wat de KTT betreft weer terug zijn bij de aanvankelijke voorwaarde van parallelliteit van Angoff. In de praktijk van het equivaleren zal zowel in de KTT als in de IRT zelden voldaan zijn aan de sterkst mogelijke variant van de uitwisselbaarheidsvoorwaarde; in het algemeen zal slechts aan de besproken zwakke varianten zijn voldaan.

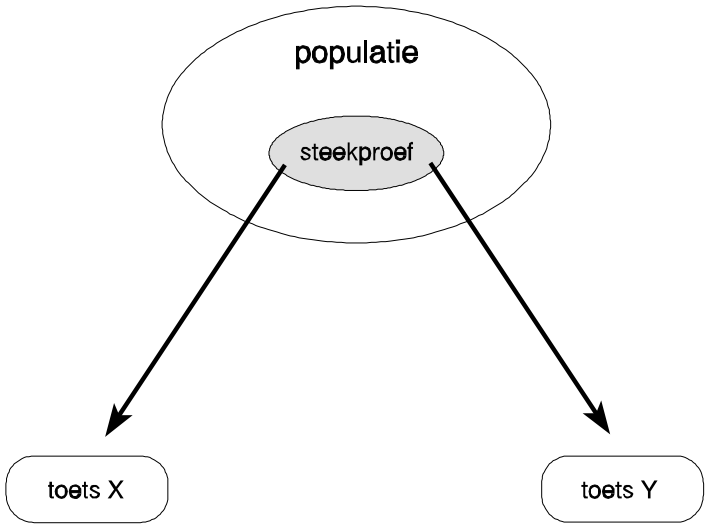
De laatste twee eisen, de invariantie- en symmetrie-eis, zijn het logisch gevolg van het eigenlijke doel van het equivaleren, namelijk het vinden van gelijkwaardige scores. Als scores op twee toetsen gelijkwaardig zijn, dan moet er een één-één relatie bestaan tussen die scores. Maar een één-één relatie is zowel uniek als inverteerbaar. De uniciteit vindt zijn weer-spiegeling in de derde eis, de invariantie over groepen. Als voorbeeld van twee groepen nemen we de opsplitsing van de populatie op basis van sexe. De invariantie eis stelt dan dat de equivaleerfunctie voor de jongens gelijk moet zijn aan die van de meisjes. Als dit niet zo zou zijn, dan is er een score op de ene toets die voor een jongen een andere equivalente score heeft op de tweede toets dan voor een meisje. De twee verschillende equivaleerfuncties hebben één score omgezet in twee verschillende scores. De vierde eis, de symmetrie-eis, kan gezien worden als de inverteerbaarheidsconditie. Stel dat voor een willekeurige score x_0 op toets X een equivalente score y_0 op toets Y gevonden is. De symmetrie eis zegt nu dat als we voor y_0 een equivalente score op toets X zoeken, dat deze score x_0 moet zijn. De derde eis, de invariantie-eis, maakt wederom duidelijk dat we de populatie precies moeten definiëren. Als we namelijk de populatie in het voorgaande definiëren als 'de meisjes', dan is er wat betreft de derde eis wellicht geen probleem meer. We schrijven hier wellicht omdat ook deze populatie weer opgedeeld kan worden, bijvoorbeeld naar leeftijd. Voor de praktijk van het equivaleren betekent dit, dat men er in ieder geval zeker van moet zijn dat de toetsen, in de eventueel te onderscheiden subpopulaties, geen verschillende vaardigheden moeten meten. Dit onderwerp, onzuiverheid, wordt in hoofdstuk 9 besproken. Aan de vierde eis, de symmetrie-eis, kan in de praktijk bijna altijd voldaan worden.

8.1.2 Designs voor equivaleren

De eerste stap die bij equivalering genomen moet worden is het vaststellen van het design voor de verzameling van de data. Voor elk design geldt dat we bij equivaleren altijd uitgaan van een of meer populaties, waaruit een steekproef (of steekproeven) van leerlingen de te equivaleren toetsen maken. Alhoewel we in sommige equivaleerproblemen vrij zijn in de keuze van een design, zij vooraf opgemerkt dat de keuze in de praktijk vaak voor een groot deel wordt bepaald door praktische randvoorwaarden. Bij equivalering wordt veelal gebruik gemaakt van een van de volgende drie basisdesigns, welke in de figuren 8.1, 8.2 en 8.3 schematisch worden weergegeven, het single group design, het random group design en het ankertoetsdesign.

Single group design

Bij dit design maakt één groep leerlingen alle te equivaleren toetsen. Als we twee toetsen willen equivaleren, zeg toets X en toets Y, dan maakt deze groep leerlingen eerst toets X en daarna toets Y. Als vermoeidheidsaspecten een rol spelen, dan is het mogelijk dat toets Y



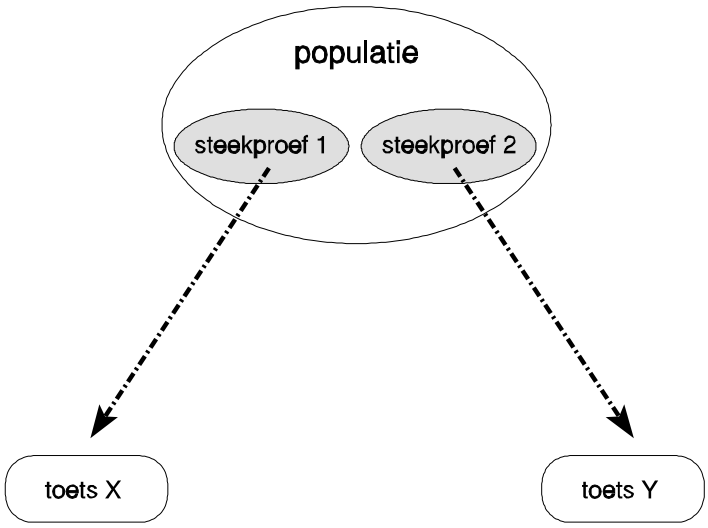
Figuur 8.1
Single group design

relatief moeilijker lijkt dan hij in werkelijkheid is. Anderzijds is het ook mogelijk dat er een zeker oefeneffect optreedt, toets Y lijkt dan gemakkelijker. Om deze effecten te vermijden, wordt bij dit design vaak gebruikt gemaakt van verwisseling: een helft van de kandidaten maakt eerst toets X en daarna Y, terwijl de andere helft eerst toets Y

en daarna X maakt. De idee is uiteraard dat oefen- en vermoeidheidseffecten elkaar dan opheffen. Helaas is het niet goed mogelijk om te onderzoeken of dit inderdaad ook gebeurt. Een ander probleem dat hiermee niet opgelost kan worden is het tijdsduureffect. Als beide toetsen een afnametijd van, zeg drie uur vragen, zal voor de afname van beide toetsen praktisch een hele dag nodig zijn. Bovendien is het vaak zo dat men een nieuwe versie van een toets wil equivaleren met een oudere, zoals bij examens en de Eindtoets. Bij dit design zal de steekproef dus zowel de oude als de nieuwe toets moeten maken. Dit zijn geen gewenste zaken, daarom wordt dit design niet vaak toegepast.

Random group design

Bij dit design, zie figuur 8.2, maken twee aselect getrokken groepen leerlingen uit één populatie elk één toets. De nadelen die we bij het single group design hebben aangegeven, zijn bij het random group design niet aanwezig. Bij nieuwe en oude versies van een toets of examen kan de geheimhouding van de oude echter wel een rol spelen. Bij dit design hebben we de extra aanname gemaakt dat we beschikken over twee vergelijkbare steekproeven, dat



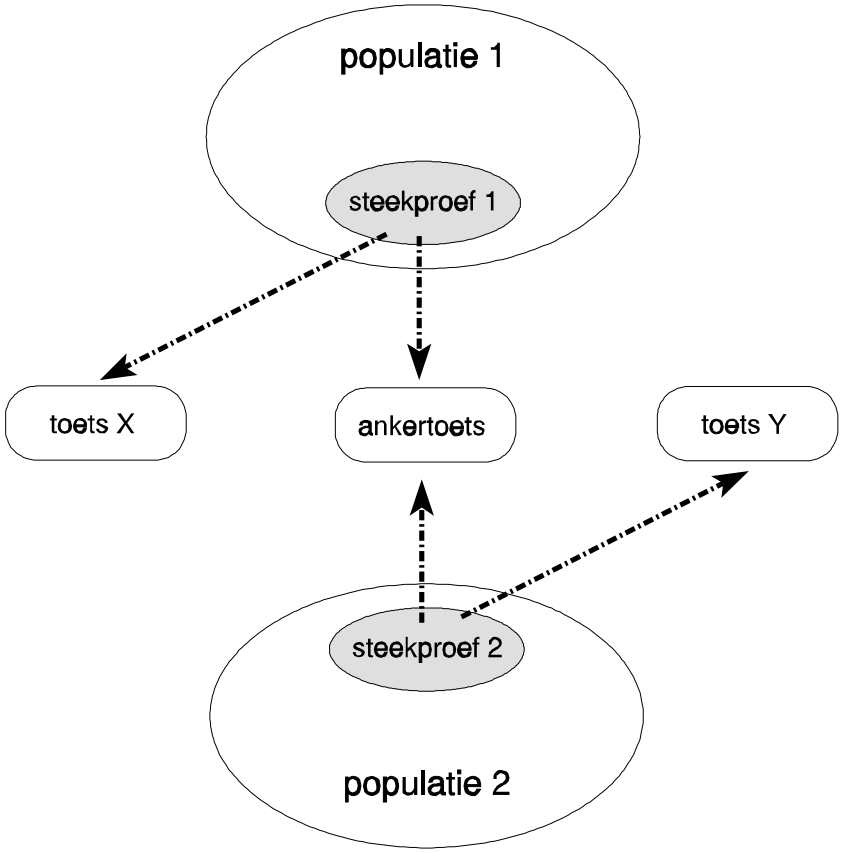
Figuur 8.2
Random group design

wil zeggen met dezelfde vaardigheidsverdeling. Deze vergelijkbaarheid wordt in de praktijk verkregen door slechts één steekproef van leerlingen te trekken en de toetsen daarna aselect toe te wijzen aan de leerlingen. De leerlingen die toets X maken vormen

dan steekproef 1, terwijl steekproef 2 bestaat uit die leerlingen die toets Y maken. Alhoewel op deze wijze de vergelijkbaarheid van de twee steekproeven zeer aanneembaar geworden is, kunnen we deze vergelijkbaarheid niet toetsen.

Ankertoetsdesign

Bij het laatste basisdesign, het ankertoetsdesign, maken twee aselect getrokken groepen leerlingen twee toetsen die een aantal items gemeen hebben. Deze groepen kunnen getrokken zijn uit één populatie, maar ook uit twee verschillende populaties. De variant met twee populaties staat in figuur 8.3. De gemeenschappelijke deeltoets wordt het anker genoemd. De bij de eerste twee basisdesigns genoemde bezwaren zijn bij dit design opgeheven. Immers,
alle leerlingen maken slechts een toets, inclusief de ankertoets. Bovendien biedt de ankertoets de mogelijkheid om voor eventuele verschillen tussen de beide groepen leerlingen te corrigeren. Stel bijvoorbeeld dat de tweede groep gemiddeld hoger scoort op de ankertoets dan de eerste: de tweede groep is dan gemiddeld vaardiger dan de eerste. Deze informatie kunnen we gebruiken om voor een eventueel verschil in moeilijkheidsgraad tussen de toetsen te corrigeren. Hoe dit precies in zijn werk gaat staat beschreven in de volgende paragrafen.



Figuur 8.3
Ankertoetsdesign

Tenslotte willen we een opmerking maken over de status van de ankertoets. Als we gebruik maken van de KTT, zullen we in dit hoofdstuk steeds aannemen dat de ankertoets extern is, dat wil zeggen dat de score op toets X alleen bepaald wordt door de antwoorden op toets X (Y). Het is namelijk ook mogelijk dat de ankertoets opgevat wordt als een deel van de te equivaleren toetsen, hetgeen in de literatuur als intern wordt omschreven. De score op toets X (en ook op Y) bestaat dan dus voor een gedeelte uit het aantal goed gemaakte opgaven uit de ankertoets. Bij het equivaleren in de IRT zullen we steeds veronderstellen dat de ankertoets intern is.

Gezien de bovengenoemde nadelen bij de eerste twee basisdesigns, is het derde basisdesign, het ankertoetsdesign, verreweg het meest gebruikte en bestudeerde equivaleerdesign (Petersen e.a., 1989; Harris & Crouse, 1992). Dit gegeven over de gebruikersfrequentie laat onverlet dat in bepaalde situaties de eerste twee basisdesigns, en dan met name het tweede, best geschikt kunnen zijn. Merk bovendien op dat het tweede en het derde basisdesign voorbeelden zijn van designs die datamatrices geven die onvolledig zijn: elke leerling heeft slechts een gedeelte van de items gemaakt. Zoals reeds in hoofdstuk 6 beschreven is, dienen dit soort designs aan bepaalde voorwaarden

te voldoen om naderhand zinvolle conclusies te kunnen trekken. We komen hier later nog op terug.

Op de drie basisdesigns zijn zeer veel varianten en combinaties ontwikkeld. Zonder volledigheid na te streven noemen we er hier een paar. Het design waarin twee aselect getrokken groepen beide toetsen maken en het design waarbij twee groepen ieder een toets maken terwijl een derde groep beide toetsen maakt, het ankergroepdesign, zijn beide voorbeelden van een combinatie van de basisdesigns. Als variant op het ankertoetsdesign kan ook het, eventueel geblokte, kettingdesign (zie hoofdstuk 6) worden genoemd. Voor alle genoemde designs geldt, zoals we later zullen zien, dat ze voor sommige equivaleermethoden wel en voor andere niet bruikbaar zijn.

Equivaleerdesign van de Eindtoets

We eindigen deze paragraaf met een voorbeeld van een design uit de praktijk. Dit betreft het design van de Eindtoets voor de jaren 1990-1993, welke in figuur 8.4 schematisch is weergegeven. Horizontaal in de figuur staan verschillende ankertoetsen en eindtoetsen (EB met jaar), verticaal de jaargroepen leerlingen. In de figuur is met grijs aangegeven wie welke toetsen maakt.

	anker K	anker L	anker M	anker N	EB90	EB91	EB92	EB93
1990	■				■			
		■			■			
					■			
1991	■					■		
		■				■		
			■			■		
						■		
1992		■					■	
			■				■	
				■			■	
							■	
1993		■						■
			■					■
				■				■
								■

Figuur 8.4
Afnamedesign Eindtoets Basisonderwijs 1990-1993

De Eindtoets wordt ieder jaar bij ongeveer 60% van de leerlingen uit groep 8 van het basisonderwijs afgenomen. Bovendien maakt elk jaar een steekproef van ongeveer 3000 leerlingen, behalve de Eindtoets van hun eigen jaar, een ankertoets. Zo'n ankertoets

is een verkleinde versies (45 items) van de Eindtoets (180 items): zowel qua inhoud alsook qua psychometrische eigenschappen zijn beide toetsen vergelijkbaar. Deze ankertoetsen houden, in tegenstelling tot de Eindtoets, dezelfde samenstelling en dienen louter voor de equivalering. Aangezien de inhoud van de Eindtoets in de loop der tijd aangepast wordt aan het veranderende onderwijs, moeten de ankertoetsen, om nog vergelijkbaar met de Eindtoets te blijven, na verloop van tijd vervangen worden. Voor twee verschillende jaren in welke dezelfde ankertoets afgenomen is, hebben we te maken met een ankertoetsdesign. Het totale design is een voorbeeld van een combinatie van de basisdesigns.

8.2 Equivaleren in de klassieke testtheorie

Voorafgaande aan de bespreking van het equivaleren in de KTT, willen we eerst een algemene opmerking maken omtrent de KTT die in dit verband van belang is. Zoals in hoofdstuk 4 is beschreven, is een van de grootste bezwaren van de KTT de onmogelijkheid om de moeilijkheid van een toets en de vaardigheid van de populatie te scheiden. Of, met andere woorden, alle uit de KTT bekende begrippen zoals *p*-waarden, r_{it} en betrouwbaarheid, hebben steeds betrekking op één populatie en (vaak) één toets. Bij het equivaleren, waar we te maken hebben met verschillende toetsen, eventueel met verschillende moeilijkheden, en (eventueel) met verschillende populaties, kan dit bezwaar ons natuurlijk nog extra parten spelen. Toch wordt equivaleren met behulp van de KTT nog steeds vrij regelmatig gebruikt. Een eerste reden hiervoor is de grote hoeveelheid van beschikbare methoden, die in de praktijk naar tevredenheid van de gebruiker worden toegepast. Een tweede reden is dat in die gevallen waar equivaleren met behulp van de IRT onmogelijk is, men wel met behulp van de KTT moet equivaleren. De eisen die de KTT stelt zijn immers zwakker als die van de IRT.

Binnen de KTT kunnen we grofweg twee klassen van equivaleermethoden onderscheiden. De eerste klasse maakt alleen gebruik van geobserveerde scores (’observed score equating’) terwijl de tweede klasse werkt met ware scores (’true score equating’). In de praktijk worden meestal alleen equivaleermethoden gebruikt die werken met de geobserveerde scores. Een eerste reden hiervoor is de eenvoud. Een tweede, en minstens zo’n belangrijke, reden is dat als men toch wil werken met ware scores, IRT vaak te prefereren is. In het kader van de KTT zullen we ons dan ook zo veel mogelijk beperken tot equivaleermethoden die gebruik maken van geobserveerde scores, soms zullen we echter ook de ware scores in het verhaal betrekken. Hierbij

houden we uiteraard rekening met de psychometrische voorwaarden zoals die in paragraaf 8.1.1 zijn behandeld.

In paragraaf 8.2.1 zullen we de basisequivaleermethoden binnen de KTT beschrijven. Aangaande de voorwaarden uit paragraaf 8.1.1, zullen we altijd aannemen dat we toetsen willen equivaleren die dezelfde vaardigheid meten. Op z'n minst moeten de toetsen dus congeneriek zijn. Een extra psychometrische aanname die vaak gemaakt wordt is dat de toetsen even betrouwbaar zijn. Het belang van gelijke betrouwbaarheid van de toetsen is evident. Zouden de toetsen namelijk niet even betrouwbaar zijn, dan zou een zwakke leerling de voorkeur geven aan een minder betrouwbare toets, terwijl de goede leerling meer baat zou hebben bij de meer betrouwbare toets. Immers, de zwakke leerling heeft bij een slechter meetinstrument een grotere kans om bijvoorbeeld boven de cesuur te scoren. Zelfs aan de zwakke versie van de rechtvaardigheidseis kan dus nooit voldaan worden voor toetsen die niet even betrouwbaar zijn. Bovendien blijkt uit de praktijk dat in de meeste situaties de (geschatte) betrouwbaarheid van de te equivaleren toetsen (ongeveer) gelijk is. Wij zullen de eis van gelijke betrouwbaarheden hier dan ook maken. In paragraaf 8.2.2 zullen we de equivaleermethoden in de KTT voor het ankertoetsdesign bespreken.

8.2.1 Basismethoden voor equivaleren

In de KTT zijn er drie basismethoden in gebruik om een equivaleerfunctie te bepalen tussen twee toetsen: de equipercentiel methode, de lineaire methode en de regressie methode, die we nu achtereenvolgens beschrijven. Om een beter inzicht te krijgen in de problematiek, zullen we in eerste instantie steeds aannemen dat we over de data beschikken van één gehele populatie $\wp$. Daarna zullen we de parameters van de vaardigheidsverdeling in die populatie schatten uit de getrokken steekproef. Een opmerking omtrent de notatie. De te equivaleren toetsen worden aangegeven met hoofdletters (bijvoorbeeld X), terwijl de (geobserveerde) scores op die toetsen cursief genoteerd worden (bijvoorbeeld X).

Equipercentielmethode

De equipercentielmethode werkt als volgt: Kies de equivaleerfunctie zodanig dat de scores op de toetsen X en Y geëquivaleerd zijn als ze corresponderen met dezelfde percentiele rang in de populatie, waaronder we verstaan het percentage scores in de

populatie dat gelijk of kleiner is. Deze equivaleermethode is historisch gezien de belangrijkste en werd vroeger zelfs als definitie gehanteerd: "Two scores, one on form X and the other on form Y (where X and Y measure the same function with the same degree of reliability), may be considered equivalent if their corresponding percentile ranks in any group are equal" (Lord, 1950; Flanagan, 1951).

Laat dus nu $\mathcal{P}$ een populatie van leerlingen zijn waarvoor de te equivaleren toetsen X en Y geschikt zijn. Elke leerling uit $\mathcal{P}$ kan dus getoetst worden met X en/of Y. Stel dat $F(x)$ en $G(y)$ de verdelingsfuncties van de geobserveerde scores van de toetsen X en Y in de populatie $\mathcal{P}$ zijn, dat wil zeggen

$$\begin{cases} F(x) = \text{proportie leerlingen in } \mathcal{P} \text{ met } X \leq x, \\ \\ G(y) = \text{proportie leerlingen in } \mathcal{P} \text{ met } Y \leq y. \end{cases} \tag{8.1}$$

Bij de equipercentielmethode worden alle percentiele rangen gelijkgesteld, hetgeen natuurlijk alleen mogelijk is als voor een willekeurige waarde van een percentiele rang p^* met $p^* = 100p$ geldt dat

$$F(x) = p \quad \text{en} \quad G(y) = p. \tag{8.2}$$

Het is eenvoudig na te gaan dat voor strikt monotone F en G er dan geldt dat

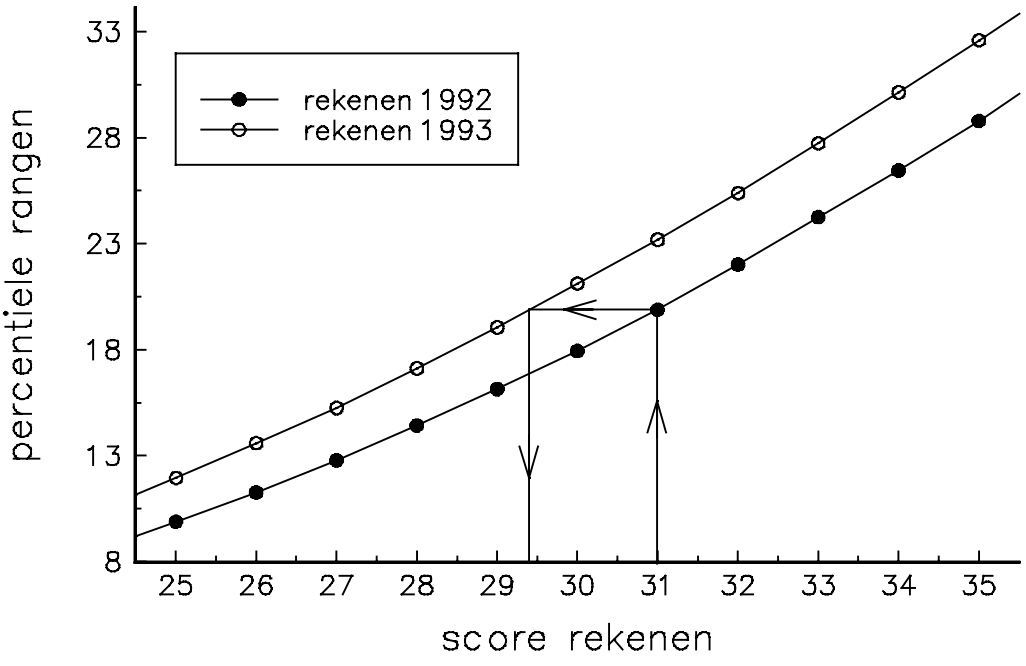
$$x = F^{-1}(G(y)). \tag{8.3}$$

De inverse functie van F, F^{-1}, wordt gegeven door het voorschrift dat $F^{-1}(p)$ die waarde van x is waarvoor geldt dat $F(x) = p$. Merk op dat x nu een functie van y geworden is. We geven dit aan met $e_X(y)$. Dus $e_X(y) = F^{-1}(G(y))$ equivaleert X en Y op $\mathcal{P}$, waarbij $e_X(y)$ de equivaleerfunctie is. Nu kan de percentiele rang p^* elke mogelijke waarde tussen 0 en 100 aannemen. De scores hebben echter slechts een eindig bereik daar alleen scores tussen 0 (alle items fout) en de maximale score (alle items goed) mogelijk zijn. De verdelingsfuncties F en G zijn dan niet meer strikt monotoon. Maar dit betekent ook dat de inverse functie nooit helemaal exact bekend is en dat de waarde van de inverse functie op de onbekende plaatsen op de een of andere manier moet worden ingevuld. Dit proces van invullen staat bekend onder de naam 'smoothen'. We zullen dit later aan de hand van een voorbeeld demonstreren.

Een ander moeilijk probleem blijft natuurlijk het bepalen van $F(x)$ en $G(y)$ omdat we in de praktijk nooit over de gehele populatie $\mathcal{P}$, maar slechts over steekproeven uit $\mathcal{P}$ beschikken. We zullen ons dus altijd moeten behelpen met schattingen van de functies F en G. Bovendien moeten complete verdelingsfuncties met in principe oneindig veel parameters geschat worden. Als we over een aselecte steekproef beschikken, dan zou

als schatting van de verdelingsfunctie natuurlijk de geobserveerde kunnen dienen. De geobserveerde verdelingsfunctie is eenvoudig uit de geobserveerde frequentieverdeling te construeren en kan bovendien met veel statistische pakketten uitgerekend worden. Hoe goed deze schatting is, hangt uiteraard af van de populatie, de steekproef en de toetsen. Het moge duidelijk zijn dat bij grotere steekproeven de geschatte verdelingsfunctie de ware beter zal benaderen.

Als voorbeeld zullen we nu laten zien hoe twee versies van de Eindtoets met behulp van de equipercentielmethode geëquivaleerd kunnen worden. We zullen ons hier beperken tot het onderdeel rekenen (60 items) voor de jaren 1992 en 1993. Als eerste stap moeten we dan de beschikking hebben over één populatie $\wp$. We kunnen dit doen als we de populatie $\wp$ definiëren als 'alle kinderen die in een willekeurig jaar in groep 8 zitten'. In werkelijkheid beschikken we natuurlijk niet over $\wp$, maar slechts over twee steekproeven van leerlingen, een die aan de Eindtoets van 1992 en een die aan de Eindtoets van 1993 deelnam; beide steekproeven bevatten ongeveer 100.000 leerlingen. De verdelingsfunctie van de scores van 1992 noemen we G en die van 1993 noemen we F. Merk op dat bij de Eindtoets de scores gegeven worden door middel van het aantal goed beantwoorde opgaven. Daar we over een zeer grote steekproef beschikken, mogen we aannemen dat de geobserveerde verdelingsfunctie $\hat{G}$ een goede schatting is van G. Hetzelfde verhaal gaat natuurlijk op voor F en $\hat{F}$. De geobserveerde verdelingsfuncties $\hat{F}$ en $\hat{G}$ zijn voor scores tussen 25 en 35 weergegeven in figuur 8.5. We hebben bovendien beide verdelingsfuncties een vloeiend verloop gegeven, dat wil zeggen een nette benaderende lijn door de discrete verdelingsfunctie getrokken. Dit is



Figuur 8.5
Equipercentiel equivaleren Eindtoets

wat we hiervoor smoothen genoemd hebben. Merk op dat de verdelingsfunctie van 1993, voor de gegeven scores, overal boven die van 1992 ligt. Ook voor de niet gepresenteerde scores bleek dit zo te zijn. Uit de aanname dat beide steekproeven getrokken zijn uit een en dezelfde populatie volgt dus dat de Eindtoets van 1993 moeilijker is dan de Eindtoets van 1992, uiteraard voor het onderdeel rekenen. Nu zijn alle gegevens voor de equipercentiel equivalering beschikbaar. Beschouw nu bijvoorbeeld een score van 31 op het onderdeel rekenen van de Eindtoets van 1992; bij deze score hoort een percentiel van (ongeveer) 20. Bij datzelfde percentiel zou een reken score op de Eindtoets van 1993 van (ongeveer) 29.4 horen. Maar deze score kan niet voorkomen, zodat we de dichtstbijzijnde score kiezen, of, met andere woorden, we ronden 29.4 af tot 29.

Lineaire methode

De lineaire methode kan omschreven worden met de volgende regel: 'Kies de equivaleerfunctie zodanig dat twee scores op X en Y equivalent zijn als ze hetzelfde aantal standaarddeviaties afwijken van de gemiddelden, ofwel dezelfde standaardscore hebben'. Voor toets X (Y) duiden we het gemiddelde van de geobserveerde scores in de populatie $\wp$ aan met μ_X (μ_Y) en de standaarddeviatie van de scores met σ_X (σ_Y). Het gelijk stellen van de standaardscores is dan equivalent met

$$\frac{X - \mu_X}{\sigma_X} = \frac{Y - \mu_Y}{\sigma_Y}. \tag{8.4}$$

Herschikken van termen in (8.4) geeft dan direct de formule voor het lineair equivaleren:

$$e_X(Y) = \mu_X + \frac{\sigma_X}{\sigma_Y}(Y - \mu_Y). \tag{8.5}$$

We merken hier op, dat als we de sterke variant van de rechtvaardigheidseis in de KTT, de toetsen zijn parallel, zouden hebben aangenomen, dat (8.4) dan reduceert tot $X = Y$. De scores op de toetsen zijn dan dus per definitie lineair geëquivaleerd.

Lineair equivaleren kan ook gezien worden als een bijzonder geval van equipercentiel equivaleren in de zin dat slechts de eerste twee momenten van de scoreverdelingen gelijkgesteld worden (Braun & Holland, 1982). Er kan namelijk eenvoudig aangetoond

worden dat bij equipercentiel equivaleren alle momenten aan elkaar gelijk gesteld worden. Een extra aanname bij lineaire equivalering is dus dat de hogere momenten van de scoreverdelingen van beide toetsen identiek zijn. Deze benadering start dan ook met de aanname dat F en G schaalinvariante functies zijn. Schaalinvariante functies zijn functies waarvan de ene functie op een lineaire transformatie na, gelijk is aan de andere. Met andere woorden, schaalinvariante functies hebben dezelfde vorm. Bij equipercentiel equivaleren moeten complete verdelingsfuncties geschat worden, hetgeen een groot nadeel van die methode is. Omdat het in het algemeen beter is om minder dan meer parameters te schatten, verdient lineair equivaleren, daar waar toepasbaar, de voorkeur.

Net zoals bij het equipercentiel equivaleren, zijn ook bij het lineair equivaleren de populatie gegevens, in dit geval de gemiddelden en de standaarddeviaties, niet bekend. Deze moeten dus altijd uit de data geschat worden en vervolgens ingevuld worden in (8.5). Als schatters voor μ_X en σ_X komen bijvoorbeeld de steekproefmomenten $\bar{X}$ en s_X in aanmerking.

Als de toetsen X en Y niet even betrouwbaar zijn, kunnen we ook lineair equivaleren. Het is duidelijk dat we nu niet meer alleen met geobserveerde scores uit de voeten kunnen. De betrouwbaarheid is immers een functie van zowel de ware als van de geobserveerde scores. De ware scores dienen nu dus op de een of andere manier expliciet gebruikt te worden. De simpelste manier is nu om (8.4) te herschrijven tot een vergelijking tussen de ware scores. Hiertoe dienen we dan zowel de geobserveerde scores als ook de parameters van de geobserveerde variabelen te vervangen door de ware score equivalenten. Dus voor toets X vervangen we μ_X door $\mu_{T(X)}$ en σ_X door $\sigma_{T(X)}$; voor toets Y geldt hetzelfde. Dit levert dan

$$\frac{T(\mathrm{X}) - \mu_{T(X)}}{\sigma_{T(X)}} = \frac{T(Y) - \mu_{T(Y)}}{\sigma_{T(Y)}}. \tag{8.6}$$

Merk nu op dat alle termen in (8.6) onbekend zijn. Zowel de ware scores $T(\mathrm{X})$ en $T(\mathrm{Y})$ als ook de parameters $\mu_{T(X)}$, $\sigma_{T(X)}$, $\mu_{T(Y)}$ en $\sigma_{T(Y)}$ van de ware score verdelingen zijn niet bekend. Gelukkig beschikken we voor alle onbekenden over goede schatters. Voor het gemak beperken we ons in de schrijfwijze even tot toets X. We starten met de parameters, daar deze het eenvoudigst zijn. Immers, uit hoofdstuk 3 weten we dat $\mu_{T(X)} = \mathscr{E}(T) = \mathscr{E}(X) = \mu_X$ en $\sigma^2_{T(X)} = \sigma^2_X \rho_{XX'}$. Voor de schatting van de ware scores beschikken we over twee kandidaten: de geobserveerde-score-schatter en de Kelley-schatter. Als we de geobserveerde score nemen als schatter voor

de ware scores, dan vullen we voor $T(X)$ dus X in. Invullen van deze schattingen in (8.6) levert dan

$$\frac{X - \mu_X}{\sqrt{\rho_{XX'}}\,\sigma_X} = \frac{Y - \mu_Y}{\sqrt{\rho_{YY'}}\,\sigma_Y}. \tag{8.7}$$

Herschikking van de termen in (8.7) levert dan de eerste formule voor het lineair equivaleren van twee niet even betrouwbare toetsen:

$$e_X(Y) = \mu_X + \frac{\sigma_X\sqrt{\rho_{XX'}}}{\sigma_Y\sqrt{\rho_{YY'}}}(Y - \mu_Y). \tag{8.8}$$

Als we de Kelley-schatter nemen als schatter van de ware score, dan wordt de schatter van de teller van het linkerlid van (8.6) gegeven door

$$\frac{\sigma^2_{E(X)}}{\sigma^2_{E(X)} + \sigma^2_{T(X)}}\,\mu_{T(X)} + \frac{\sigma^2_{T(X)}}{\sigma^2_{E(X)} + \sigma^2_{T(X)}}\,X - \mu_{T(X)}, \tag{8.9}$$

waarbij $\sigma^2_{E(X)}$ de foutenvariantie weergeeft. Uitwerken van (8.9) geeft $\rho_{XX'}(X - \mu_{T(X)})$. Invullen hiervan en van de bovengenoemde schatters voor de parameters en herschikking van de verschillende termen levert dan de tweede formule voor het equivaleren van twee niet even betrouwbare toetsen:

$$e_X(Y) = \mu_X + \frac{\sigma_X\sqrt{\rho_{YY'}}}{\sigma_Y\sqrt{\rho_{XX'}}}(Y - \mu_Y). \tag{8.10}$$

Merk op dat in de formules (8.8) en (8.10) de ratio tussen de wortels van de beide betrouwbaarheden is omgekeerd. Bovendien geldt voor beide formules dat het verschil met (8.5) alleen zit in de ratio van die wortels. Hieruit lezen we dan ook direct af dat het voor twee bijna even betrouwbare toetsen, het praktisch geen verschil maakt of formule (8.5) dan wel (8.8) of (8.10) gebruikt wordt. Ten overvloede wellicht, zullen in de praktijk zowel in (8.8) als in (8.10) schattingen voor de parameters moeten worden ingevuld. Merk op dat nu ook de verschillende betrouwbaarheden geschat moeten worden. Hoe de betrouwbaarheid van een toets geschat kan worden is reeds uitgebreid behandeld in paragraaf 3.6, we zullen dit hier dan ook niet herhalen.

Regressiemethode

Bij de regressiemethode wordt de equivaleerfunctie tussen de scores bepaald door de regressie van de scores van de ene toets op de andere te bepalen. Voor de lineaire regressie van X op Y volgt dan

$$e_X(Y) = \mu_X + \rho_{XY}\frac{\sigma_X}{\sigma_Y}(Y - \mu_Y), \tag{8.11}$$

waarbij ρ_{XY} de correlatie tussen de scores van de toetsen X en Y is. Merk op dat (8.11) identiek is aan (8.8) op de factor ρ_{XY} na. Om ρ_{XY} te schatten is het noodzakelijk om over een steekproef van leerlingen te beschikken die zowel toets X als toets Y gemaakt hebben. Dit is bijvoorbeeld mogelijk als de data verzameld zijn volgens het eerste basisdesign, het single group design. In (8.11) wordt de equivaleerfunctie bepaald door de regressie van X op Y. Als we de rol van X en Y omdraaien, dat wil zeggen als we de regressie van Y op X bepalen, dan vinden we

$$e_Y(X) = \mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(X - \mu_X). \tag{8.12}$$

Nu is (8.12) niet de inverse van (8.11), hetgeen niet strookt met de symmetrie eis. De equipercentiel en de lineaire methode voldoen wel aan de symmetrie eis, hetgeen direct uit (8.2) en (8.4) kan worden afgelezen. De regressiemethode dient dus altijd met de nodige voorzichtigheid betracht te worden.

We vervolgen nu het voorbeeld van de equivalering van de Eindtoets. Voor de lineaire equivalering hebben we alleen maar de eerste twee momenten nodig. Schattingen van μ_X etcetera worden uiteraard gegeven door de steekproefmomenten, deze zijn $\bar{X} = 41.22$, $s_X = 11.46$, $\bar{Y} = 41.96$ en $s_Y = 10.98$. Invullen van deze schattingen in (8.5) levert dan de equivaleerfunctie $e_X(Y) = 41.22 + 1.04(Y - 41.96) = -2.57 + 1.04\,Y$. Merk op dat Y bij 1992 hoort en X bij 1993. Voor de score van 31 op de Eindtoets van 1992 vinden we dan de lineair geëquivaleerde score van 29.67 op de Eindtoets van 1993, hetgeen redelijk overeenkomt met de score van 29.4 bij het equipercentiel equivaleren. Merk op dat er voor de regressiemethode nooit genoeg gegevens zijn. Er zijn immers geen leerlingen die beide versies van de Eindtoets gemaakt hebben, zodat we ρ_{XY} niet kunnen schatten.

In het bovenstaande hebben we de equivalering van de toetsen X en Y steeds eerst beschouwd op de totale populatie $\mathcal{P}$. We merkten daarbij op dat we in werkelijkheid nooit beschikken over de gehele populatie, doch slechts uit steekproeven hieruit. We moeten dus altijd de data, en daarmee het design meenemen om tot een goede keuze voor de equivaleerprocedure te komen. Bovendien kan er sprake zijn, zoals bijvoorbeeld bij het verticaal equivaleren, van meerdere populaties. Vooral dit laatste is nog

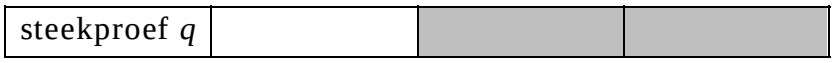
een behoorlijk probleem. Bij de bespreking van het voorbeeld van de equivalering van de Eindtoets hebben we dit probleem een beetje verdoezeld. We hadden daar immers ook twee populaties, die van 1992 en die van 1993, die we samengevoegd hebben tot een (alle leerlingen in groep 8). Dit samenvoegen tot een populatie is statistisch goed gefundeerd (Braun & Holland, 1982), maar conceptueel moeilijk voorstelbaar. Deze populatie heet in de literatuur 'synthetic population'. We zullen in het vervolg dan ook aannemen dat, indien er twee populaties in het geding zijn, deze samengevoegd zijn tot één synthetische populatie. We bespreken nu de equivalering van het in de praktijk vaak voorkomende ankertoetsdesign.

8.2.2 Equivaleren met behulp van het ankertoetsdesign

In deze paragraaf bespreken we het equivaleren indien de data verzameld zijn met een anker- toetsdesign. De nadruk zal hierbij liggen op de meest gebruikte vorm van equivaleren, namelijk lineair equivaleren. Voor de duidelijkheid hebben we het ankertoetsdesign nogmaals weergeven in figuur 8.6. Steekproef *p*, uit populatie 1, maakt toets X en de ankertoets A (X-groep), terwijl steekproef *q* uit populatie 2, toets Y en ankertoets A maakt (Y-groep). De totale steekproef, *p* en *q* samen, zullen we *t* noemen. Populatie 1 en populatie 2 vormen samen de synthetische populatie $\wp$; *t* is een steekproef uit $\wp$.

Allereerst een opmerking over de ankertoets A. Evenals voor de te equivaleren toetsen X en Y, zullen we ook aan de ankertoets psychometrische eisen moeten opleggen. Als we bijvoorbeeld twee toetsen engels willen equivaleren, dan mogen we van de ankertoets op z'n minst verwachten dat deze ook engels meet. Een redelijke eis is dan hier ook dat de ankertoets A congeneriek is met X (en dus ook met Y). Ook hier geldt weer, dat naarmate de eisen sterker worden, de equivalering eenvoudiger wordt. Een overzicht van alle mogelijke psychometrische eisen voor lineair equivaleren die in een ankertoetsdesign gesteld kunnen worden is te vinden in MacCann (1990). Bedenk bovendien dat we steeds veronderstellen dat de ankertoets extern is, zodat de ankertoets niets aan de te equivaleren scores bijdraagt. Uiteraard nemen we weer aan dat toets X en toets Y even betrouwbaar zijn.

	toets X	ankertoets A	toets Y
steekproef *p*			



Figuur 8.6
Ankertoetsdesign

We gaan nu verder met het beschrijven van de equivalering in het ankertoetsdesign. Een belangrijke observatie is nu dat we direct zouden kunnen equivaleren als we over data zouden beschikken in de lege cellen. We zijn dan immers weer terug in de situatie van volledige data uit de vorige paragraaf. Alle equivaleermethoden welke met ontbrekende data werken, vullen dan ook op de een of andere manier deze ontbrekende data in, om zo weer in het volledige data geval terecht te komen. De idee bij dit invullen is natuurlijk om de gegevens van ankertoets A te gebruiken om de scores van leerlingen uit de Y-groep (X-groep) op toets X (Y) te voorspellen. Soms hebben we echter niet de scores op de toetsen nodig, maar kunnen we met minder gegevens volstaan. Als we, bijvoorbeeld, lineair willen equivaleren, dan leert inspectie van (8.5) dat de enige relevante grootheden de gemiddelden en de standaarddeviaties van de scores in de verschillende populaties zijn. Het bepalen van deze gemiddelden en standaarddeviaties, of meer algemeen voor de ingevulde waarden, gebeurt dan uiteraard onder bepaalde aannames. De meest gebruikte aanname is die welke in de literatuur 'constancy of regression' wordt genoemd. Bij deze aanname wordt eerst verondersteld dat de scores op de toetsen X en Y een lineair verband hebben met de ankertoets, zodat lineaire regressie zinvol wordt. Vervolgens veronderstelt men dat de intercept, de regressiecoëfficiënt en de variantie van de schattingsfout van de scores op toets X (Y) op A is gelijk voor de X-groep (Y-groep) en de totale groep (= X-groep + Y-groep). Met andere woorden, als we de totale steekproef zouden hebben geobserveerd, dan zouden we dezelfde schattingen voor alle regressie-parameters gevonden hebben als we nu voor de gedeeltelijke steekproef gevonden hebben.

We zullen nu laten zien waarop de 'constancy of regression' aanname gebaseerd is. Laat daartoe μ_X en σ_X het onbekende gemiddelde en de standaarddeviatie van de scores van toets X zijn in de synthetische populatie Ω. We zullen eerst laten zien hoe we op een eenvoudige manier een goede schatter van μ_X kunnen construeren. Een eerste schatting is simpel te maken. Kies daartoe gewoon het gemiddelde van X in de geobserveerde steekproef p, oftewel $\hat{\mu} = \bar{X}_p$. Het moge duidelijk zijn dat we om deze schatting te kunnen verbeteren op de een of andere manier gebruik zullen moeten maken van de gegevens omtrent A in de totale steekproef t. Daartoe beschouwen we eerst de volledige data (X,A) in steekproef p, waarbij we aannemen dat er een lineair verband is tussen X en A. Stel nu eens dat $X_v = \beta_0 + \beta_1 A_v + \varepsilon_v$ met $\varepsilon_v \sim N(0, \sigma^2)$ in steekproef p. Hierbij, en in het vervolg, staat de subscript v voor een leerling. De

subscripten X, Y, A, p, q en t spreken voor zich; ze verwijzen naar de toetsen en de steekproeven (of bijbehorende populaties). Dan worden de kleinste kwadraten schatters $\hat{\beta}_0$ en $\hat{\beta}_1$ gegeven door

$$\hat{\beta}_0 \equiv b_0 = \bar{X}_p - b_1 \bar{A}_p \quad \text{en}$$
$$\hat{\beta}_1 \equiv b_X = r_{X A_p} s_{X_p} / s_{A_p}, \tag{8.13}$$

waarbij $r_{X A_p}$ de correlatie tussen X en A in steekproef p is. De geschatte waarde van X_v in steekproef p wordt dan, met de gebruikelijke notatie voor gemiddelden, gegeven door

$$\hat{X}_v = \bar{X}_p + b_X (A_v - \bar{A}_p). \tag{8.14}$$

Vervolgens nemen we aan dat deze formule ook geldt voor leerlingen in steekproef q. Met behulp van bovenstaande regressievergelijking kunnen we dus ook voor leerlingen in steekproef q geschatte waarden voor X_v berekenen (imputeren). Merk op dat dit volledig identiek is aan het voorspellen van de waarde van de afhankelijke variabele voor een nieuwe waarde van de onafhankelijke variabele in een eenvoudig regressieprobleem.

Het geschatte gemiddelde in de totale steekproef t wordt gegeven door formule (8.14) te middelen over de totale steekproef t, zodat we vinden

$$\hat{\mu}_X = \bar{X}_p + b_X (\bar{A}_t - \bar{A}_p). \tag{8.15}$$

Dit nieuwe geschatte gemiddelde $\hat{\mu}_X$ is dus verkregen door de gegevens van de steekproeven p en q op een eenvoudige manier samen te nemen. Op dezelfde manier, maar met meer schrijfwerk wat we hier achterwege zullen laten, kunnen we ook een schatting voor σ_X^2 construeren:

$$\hat{\sigma}_X^2 = S_{X_p}^2 + b_X^2 (S_{A_t}^2 - S_{A_p}^2). \tag{8.16}$$

Dit extra schrijfwerk is een rechtstreeks gevolg van het feit dat de standaardfout voor de geïmputeerde waarden anders (en groter) is dan voor de geobserveerde waarden. Op precies dezelfde manier als voor toets X kunnen we ook het (geschatte) gemiddelde en de standaarddeviatie voor toets Y in de totale steekproef t berekenen. Deze worden dan gegeven door

$$\hat{\mu}_Y = \bar{Y}_q + b_Y (\bar{A}_t - \bar{A}_q) \quad \text{en}$$

$$\hat{\sigma}_Y^2 = S_{Y_q}^2 + b_Y^2(S_{A_t}^2 - S_{A_q}^2), \tag{8.17}$$

waarbij b_Y de (geschatte) regressiecoëfficiënt is van Y op A in steekproef q.
Bekijk nu nogmaals de 'constancy of regression' aanname. Als we deze aanname voor toets X in formule vorm opschrijven, dan vinden we

$$\begin{aligned} &\mu_{X_t} - \beta_{XA_t}\mu_{A_t} = \mu_{X_p} - \beta_{XA_p}\mu_{A_p} && \text{intercept} \\ &\beta_{XA_t} = \beta_{XA_p} && \text{regressie-coëfficiënt} \\ &\sigma_{X_t}^2(1 - r_{XA_t}^2) = \sigma_{X_p}^2(1 - r_{XA_p}^2) && \text{foutenvariantie.} \end{aligned} \tag{8.18}$$

Hierbij staan aan de linkerkant steeds de parameters voor de totale steekproef t en aan de rechterkant voor steekproef p. Substitutie van de tweede vergelijking van (8.18) in de eerste en herschikking van de termen levert dan

$$\mu_{X_t} = \mu_{X_p} + \beta_{XA_p}(\mu_{A_t} - \mu_{A_p}). \tag{8.19}$$

Als we wederom in (8.18) de tweede vergelijking in de derde invullen, en bedenken dat $r_{XA_t} = \beta_{XA_t}\sigma_{X_p}/\sigma_{X_t}$, levert herschikken

$$\sigma_{X_t}^2 = \sigma_{X_p}^2 + \beta_{XA_p}(\sigma_{A_t}^2 - \sigma_{A_p}^2). \tag{8.20}$$

Als we nu in de rechterleden van (8.19) en (8.20) de gebruikelijke schattingen voor de parameters substitueren, dan vinden we weer (8.15) en (8.16) terug.
De 'constancy of regression' aanname is dus niets anders dan datgene wat we in een eenvoudig lineair regressieprobleem doen, als we voor het voorspellen van de afhankelijke variabele, waarden van de predictor invullen die niet gebruikt zijn bij het bepalen van de regressievergelijking.

We beschikken nu over de benodigde gegevens om in t tot de eigenlijke equivalering over te gaan. We hebben nu immers voor elke leerling een score (geobserveerd dan wel geïmputeerd) op zowel toets X als op toets Y; bovendien beschikken we nu over (schattingen) van de gemiddelden en van de standaarddeviaties van de scores. In principe kunnen nu alle klassieke equivaleermethoden direct worden uitgevoerd. Voor lineair equivaleren moeten we de gegevens uit de formules (8.15), (8.16) en (8.17) invullen in formule (8.5). Dit levert dan

$$e_X(Y) = \bar{X}_p + b_X\left(\bar{A}_t - \bar{A}_p\right) + \sqrt{\frac{S_{X_p}^2 + b_X^2\left(S_{A_t}^2 - S_{A_p}^2\right)}{S_{Y_q}^2 + b_Y^2\left(S_{A_t}^2 - S_{A_q}^2\right)}}\left(Y - \bar{Y}_q - b_Y\left(\bar{A}_t - \bar{A}_q\right)\right). \qquad (8.21)$$

Bedenk dat we hiervoor steeds aangenomen hebben dat de toetsen X en Y even betrouwbaar zijn. Ook voor toetsen die niet even betrouwbaar zijn, kunnen we, net zoals in paragraaf 8.2.1, een formule voor het lineair equivaleren in het ankertoetsdesign afleiden. Ook dan geldt weer, dat het voor de praktijk weinig verschil uitmaakt of de toetsen even betrouwbaar, danwel bijna even betrouwbaar zijn (MacCann, 1990). Bovendien hebben we aangenomen dat de toetsen X, Y en A congeneriek zijn. Zoals reeds in hoofdstuk 3 is opgemerkt, dient het toetsen op het congeneriek, of het even betrouwbaar, zijn van twee toetsen in een ruimer model plaats te vinden, bijvoorbeeld in een LISREL kader (Jöreskog & Sörbom, 1989). Hiervoor is het echter noodzakelijk om over de covariantie- of correlatiematrix van de toetsscores te beschikken. Omdat in het ankertoetsdesign de toetsen X en Y nooit bij dezelfde leerlingen zijn afgenomen, kunnen we de correlatie tussen X en Y niet schatten. Alleen door extra dataverzameling kunnen we op het congeneriek of even betrouwbaar zijn toetsen. We zullen hier verder niet op ingaan.

We sluiten nu het voorbeeld van de equivalering van de Eindtoets, voor het onderdeel rekenen, af. Daar we over drie verschillende ankertoetsen beschikken, (L, M en N) kunnen we ook op drie verschillende manieren equivaleren. We kunnen namelijk elke ankertoets de rol van A laten spelen in formule (8.21). We zullen de gegevens presenteren voor de ankers L en M. Als we deze formule uitwerken, waarvan we de details hier niet zullen presenteren, dan vinden we voor anker L de equivaleerfunctie $e_X(Y) = 1.04\,Y - 1.82$. Voor anker M wordt de equivaleerfunctie gegeven voor $e_X(Y) = 1.04\,Y - 2.52$. Merk op dat, alhoewel deze formules veel op elkaar lijken, ze toch niet geheel identiek zijn. Het lijkt er dus op dat de invariantie-eis hier geschonden is, daar de equivaleerfuncties voor twee verschillende groepen niet gelijk zijn. Als we echter toetsen of deze twee equivaleerfuncties verschillen, dan blijkt dat ze (statistisch) niet te onderscheiden zijn. Immers, het moge duidelijk zijn dat de standaardfout horende bij (8.21) best behoorlijk groot kan zijn. De equivaleerfunctie is namelijk opgebouwd uit heel veel verschillende elementen, die we allemaal moeten schatten. De fouten die we hierbij maken werken natuurlijk door in het uiteindelijke resultaat. De precieze berekening van de standaardfout van (8.21) is nogal ingewikkeld, en zullen we hier dan ook achterwege laten, zie bijvoorbeeld Braun en Holland (1982). We willen hier nog opmerken dat in de praktijk van de equivalering van de Eindtoets gewerkt wordt met de gemiddelde equivaleerfunctie. Zoals hiervoor al is opgemerkt, hebben we bij de afleiding van (8.21) aangenomen dat de twee te equivaleren toetsen gelijke

betrouwbaarheden hebben. Dit blijkt voor dit voorbeeld redelijk te kloppen. Voor de Eindtoets van 1992 vinden we als schatting van de betrouwbaarheid .918, terwijl we .920 voor die van 1993 vinden, uiteraard steeds voor het onderdeel rekenen.

8.3 Equivaleren met itemresponstheorie

Bij de bespreking van de equivaleermethoden in de KTT hebben we opgemerkt dat het soms problematisch is om de scores van verschillende toetsen op dezelfde schaal uit te drukken, en dus vergelijkbaar te maken, aangezien de moeilijkheid van opgaven of toetsen en de vaardigheid van personen niet gescheiden kunnen worden. In de IRT ligt de zaak heel anders: vaardigheden van personen en kenmerken van items worden middels aparte parameters in een kansmodel aan elkaar gerelateerd. En indien voor een verzameling opgaven in een bepaalde populatie een itemresponsmodel geldt, dan kunnen de vaardigheidsparameters van personen op eenzelfde schaal geschat worden door slechts deelverzamelingen van de betrokken opgaven te beschouwen. Maar dit laatste is nu juist waar het bij de equivalering om gaat. Immers, bij equivalering willen we de scores op verschillende toetsen vergelijkbaar maken. Maar als we de vaardigheidsparameter onafhankelijk van de toetsen kunnen bepalen, hoeven we de scores niet meer vergelijkbaar te maken. Ze liggen immers direct op de vaardigheidsschaal waarop we kunnen weergeven.

Het voorgaande suggereert dat er bij toepassing van de IRT geen equivaleerproblemen zijn. In principe is deze uitspraak juist, maar er zijn in de praktijk nog diverse interessante problemen, die we nu kort aan zullen duiden.

Allereerst moet er voldaan zijn aan de eerste aanname uit de vorige alinea: we moeten een itemverzameling hebben met antwoorden van personen die aan een bepaald itemresponsmodel voldoen. Voordat we in de IRT gaan equivaleren moeten we eerst calibratieproblemen oplossen. Onder calibratie verstaan we het kiezen van een geschikt itemresponsmodel, het afnemen van data volgens een bepaald design, het schatten van de itemparameters en het toetsen op de geldigheid van het model. Calibratie is geen eenvoudige zaak en de problemen ermee in de praktijk moeten zeker niet onderschat worden. Een groot deel van de calibratie is reeds uitgebreid besproken in de hoofdstukken 4, 5 en 6. In paragraaf 8.3.1 zullen we een aantal aspecten nog eens de revue laten passeren. Indien de calibratie succesvol is afgesloten kunnen we de vaardigheid van de personen schatten op de vaardigheidsschaal. Dit onderwerp wordt in paragraaf 8.3.2 besproken. Hiermee zouden we IRT equivaleren kunnen afsluiten. Deze laatste twee paragrafen bespreken namelijk precies het equivaleren als we kunnen

werken met gecalibreerde itembanken: we zorgen voor een goede calibratie en de score op elke toets die we uit de bank samenstellen is automatisch geëquivaleerd middels vaardigheidsschattingen op de vaardigheidsschaal. De schaal waarop deze schattingen liggen kunnen we tenslotte nog transformeren naar een schaal die de gebruiker in staat stelt de resultaten goed te interpreteren. Aangezien dit laatste onderwerp uitgebreid wordt besproken in hoofdstuk 13, zullen we er hier verder geen aandacht aan besteden. De situatie waarin we met gecalibreerde itembanken kunnen werken zouden we actief equivaleren kunnen noemen: we stellen per definitie geëquivaleerde toetsen samen uit de itembank. In paragraaf 8.3.3 bespreken we een concreet voorbeeld van de opbouw en het werken met geëquivaleerde toetsen uit een itembank.

In de praktijk zijn er echter nog veel situaties waarin we passief moeten equivaleren: we beschikken over twee of meer toetsen waarvan de scores geëquivaleerd moeten worden. Van deze bestaande toetsen moet dan nagegaan worden of ze te calibreren zijn onder een IRT-model. Als er een passend IRT-model is gevonden, dan kan het soms nog een probleem zijn dat de resulterende schattingen op de vaardigheidsschaal komen te liggen en niet op een bestaande schaal voor de toets, bijvoorbeeld de ruwe scoreschaal. Een uitweg daarvoor kan bij IRT altijd worden gevonden via het zogenaamde ware score equivaleren, hetgeen we ook in paragraaf 8.3.2 zullen bespreken. Tenslotte zullen we in paragraaf 8.3.4 een mogelijke aanpak bespreken bij het equivaleren van bestaande toetsen als het gewenste IRT-model niet past.

8.3.1 Calibratie

Na de uitvoerige behandeling van de calibratie in de hoofdstukken 4, 5 en 6 zullen we ons hier beperken tot een aantal algemene overwegingen en factoren die direct gevolgen voor de praktijk van het equivaleren (kunnen) hebben. Welke factoren zijn dat nu precies? In de eerste plaats is (uiteraard) het gekozen itemresponsmodel van belang. Ten tweede kan het gebruikte design een rol spelen en ten derde moet er een methode gekozen worden waarmee de itemparameters geschat worden. Tenslotte besteden we ook nog enige aandacht aan het toetsen van het model. Al deze zaken impliceren keuzes en bovendien zijn deze keuzes niet onafhankelijk.

De keuze van het itemresponsmodel

Bij de keuze van het itemresponsmodel spelen vele factoren een rol. De toetsspecificatie, waarmee ondermeer bedoeld wordt het vaststellen van het doel van de toetsing en de keuze van het soort items, zie hoofdstuk 1, beperkt voor een groot deel de keuze uit de grote klasse van de bestaande IRT-modellen. Een paar voorbeelden: worden de items dichotoom dan wel polytoom gescoord; kan gokken een rol kan spelen, zoals bijvoorbeeld bij meerkeuze-items; is de te meten vaardigheid uni- of multidimensionaal. We zullen ons voorlopig beperken tot de unidimensionale modellen. Gegeven de toetsspecificatie moeten we binnen de beschikbare klasse een model kiezen. Een belangrijke overweging bij de keuze kan zijn, dat als we een model kiezen met voldoende statistieken voor de vaardigheidsparameter, dit automatisch leidt tot vaardigheidsparameterschatters die direct gekoppeld zijn aan de in de praktijk vaak gewenste (gewogen) ruwe scores op een toets. De keuze voor een bepaald itemresponsmodel heeft ook de belangrijke consequentie dat voor een deel de schattingsmethode reeds vastligt. Alleen als we kiezen voor een model met voldoende statistieken voor de vaardigheid hebben we, zoals uitvoerig betoogt in hoofdstuk 4 en 5, de voordelige eigenschappen van de CML-schattingsmethode ter beschikking en bovendien hebben we dan modeltoetsen met goede statistische eigenschappen. Een keuze voor bijvoorbeeld het drieparameter logistisch model, zie hoofdstuk 5, sluit de CML-schattingsmethode uit.

De eerste keuze voor een IRT-model wordt bepaald door het afwegen van theoretisch gewenste eigenschappen en praktische wensen en randvoorwaarden, echter deze keuze is soms slechts een voorlopige. Het is immers mogelijk dat tijdens de calibratie blijkt dat we met het gekozen model niet goed overweg kunnen en dat we een ander, vaak een ruimer, model moeten kiezen.

Het design

Het design is binnen de IRT een belangrijke factor. In hoofdstuk 6 hebben we gezien dat het design voor een gedeelte de schattingsmethode vastlegt. Bovendien is daar reeds uiteengezet dat om meer redenen de traditionele omweg van calibreren in volledige deeldesigns en het daarna op dezelfde schaal brengen van de itemparameters, soms het equivaleren van itemparameters genoemd, zo mogelijk vermeden dient te worden. Het schatten van de itemparameters dient in één calibratie plaats te vinden, ook als het design onvolledig is. Bovendien moeten we ons realiseren dat de keuze van een design vooral beperkt wordt door praktische randvoorwaarden, bijvoorbeeld in het geval dat

we twee bestaande toetsen gaan equivaleren. Alleen bij het actief equivaleren, het opbouwen van een itembank, staan doorgaans alle mogelijke designs ter beschikking.

Laten we de drie basisdesigns uit paragraaf 8.1.2 eens nader bekijken. Bij het eerste basisdesign, het single group design, zijn alle schattingsmethoden mogelijk. Bij het random group design, het tweede basisdesign, is er geen overlap tussen de items en ook niet tussen de personen. De extra aanname die bij dit design dan ook gemaakt dient te worden is dat de twee steekproeven uit één populatie getrokken zijn. Als we nu één vaardigheidsverdeling voor deze populatie aannemen, dan kunnen we met MML de itemparameters en ook de parameters van de vaardigheidsverdeling schatten. Merk op dat de CML schattingsprocedure bij het random group design nooit mogelijk is omdat dit design niet verbonden is. Het derde basisdesign, het ankertoetsdesign, heeft in zijn algemeenheid de ruimste toepassings- mogelijkheden en laat daarbij ook altijd nog een keuze voor de schattingsprocedure toe. Voor dit design is MML altijd mogelijk, en, als het model dit toelaat, CML ook.

Zoals eerder reeds opgemerkt is het ankertoetsdesign het enige basisdesign dat verticale equivalering mogelijk maakt. In dit verband moet er op gewezen worden dat in dat geval er wel speciale eisen aan de samenstelling van het anker moeten worden gesteld. We zullen dit met een voorbeeld toelichten. Als men toetsen calibreert die een onderwijstraject over een aantal jaren bestrijken en waarmee men de vorderingen van de leerlingen in kaart wil brengen, kan men niet met een vaste ankertoets werken. Vooruitgang op de ankertoets is namelijk bepalend voor de mogelijk te meten vooruitgang van de leerlingen over de jaren. In dit geval zal men per meetmoment ankers moeten kiezen die de vooruitgang kunnen weergeven. Zonder zorgvuldige analyse van het vaardigheidsdomein in de tijd en relevante keuzes voor de afnamemomenten kan het verticaal geëquivaleerde instrument mogelijk irrelevante veranderingen in de vaardigheid weergeven. In hoofdstuk 10 zal op dit onderwerp nog worden teruggekomen. Als algemene aanbeveling voor de samenstelling voor een ankertoets kan gesteld worden dat de inhoud ervan en ook de psyschometrische eigenschappen representatief moeten zijn voor de toetsen die het anker verbindt, zoals we ook al in paragraaf 8.2.2 zagen. Bij verticale equivalering impliceert dit dus ook een goede spreiding van de items qua moeilijkheid.

Toetsing van het model

Daar de modeltoetsing reeds uitgebreid behandeld is in hoofdstuk 4, volstaan we hier met het maken van een tweetal opmerkingen. De eerste opmerking betreft de calibratie

voor het verticaal equivaleren. Om verticaal te kunnen equivaleren zal, daar de vaardigheids- verdelingen flink kunnen verschillen, de verbondenheid uit de items moeten komen. Dat wil dus zeggen dat de ankeritems door personen met flink uiteenlopende vaardigheden gemaakt zullen gaan worden. Een belangrijke vraag in dit verband is dan: meten deze items wel hetzelfde in de verschillende populaties? Naast de gebruikelijke toetsing van het IRT-model, zullen we hierop speciaal moeten toetsen. Hoe hierop getoetst moet worden is het onderwerp van hoofdstuk 9, dat het onderwerp itemonzuiverheid behandeld. We zullen hier dan ook niet verder op ingaan.

De tweede opmerking heeft te maken met slecht passende items. Bij de calibratie zullen er, zoals de ervaring leert, naar alle waarschijnlijkheid items verwijderd moeten worden die om de een of andere reden niet aan het gekozen itemresponsmodel voldoen. Als de calibratie dient om een itembank te construeren, dat wil zeggen om een verzameling van items te vinden die op dezelfde schaal liggen, dan is er geen probleem. Tenminste, als de domeinomschrijving van de overgebleven items nog voldoende dekking geeft zodat we nog steeds hetzelfde meten. Anders is het als de equivalering plaats dient te vinden op bestaande toetsen, eerder passieve equivalering genoemd. We kunnen de equivalering dan uitvoeren met de overgebleven items. Een nadeel hiervan kan zijn dat de leerlingen slechts op een gedeelte van de werkelijk gemaakte toets worden beoordeeld. Dit kan problematisch en oneerlijk zijn, denk hierbij bijvoorbeeld aan de eindexamens. In dat geval zullen we óf een itemrespons-model moeten kiezen waarbij géén items meer verwijderd hoeven te worden óf we zullen moeten equivaleren met behulp van de KTT.

8.3.2 Verschillende vormen van equivalering in de itemresponstheorie

Binnen de IRT zijn er, net zoals in de KTT, in principe, twee methoden in gebruik om te equivaleren. De eerste methode, die het vaakst wordt gebruikt, is het equivaleren via het schatten van de vaardigheid. Hierbij wordt voor elke persoon op basis van zijn antwoord-patroon een schatting $\hat{\theta}$ van zijn of haar latente vaardigheid θ berekend. Deze schattingen zijn dan gelijk geëquivaleerd, daar ze op dezelfde schaal liggen. De tweede methode, die met name in de Amerikaanse literatuur veel wordt besproken, zie bijvoorbeeld Lord (1980), is het ware score equivaleren. Deze methode, die met name gebruikt wordt bij het equivaleren van bestaande toetsen, gebruikt ook schattingen van θ en transformeert deze naar een schaal die past bij de oorspronkelijke ruwe (en ware) score schaal van de toets. Alvorens deze methoden te bespreken merken we op dat beide methoden ervan uitgaan dat calibratie van alle items succesvol is verlopen. We

beschikken dan dus over schattingen van de itemparameters, die daarna als vast verondersteld worden. Bij het berekenen van de vaardigheidsschattingen gaan we er dan eigenlijk ten onrechte van uit dat de itemparameters geen schattingsfout hebben. Over het precieze effect van deze benadering is nog slechts weinig bekend. Dit effect wordt uiteraard geringer naarmate de schattingsfouten van de itemparameters kleiner zijn. De grootte van de steekproef en het afnamedesign zijn hiervoor bepalend.

Het schatten van de vaardigheid

In hoofdstuk 4 zijn drie methoden voor het schatten van de vaardigheid behandeld te weten de ML, WML en de bayesiaanse schattingsmethode EAP. De eigenschappen en respectieve-lijke voor- en nadelen van deze methoden zijn daar reeds uitgebreid besproken. Een voorbeeld met een vergelijking van schattingen met deze methoden staat in tabel 4.13. Hier volstaan we met een aantal opmerkingen over de keuze van een schatter voor de vaardigheid in relatie tot de schattingsmethode die bij de calibratie is gevolgd. Voor de keuze van een methode voor het schatten van de vaardigheid is het van belang of het itemresponsmodel wel of geen voldoende statistieken voor de vaardigheid heeft. In modellen zonder voldoende statistieken voor de vaardigheidsparameter moet de calibratie, als we de JML-methode vanwege het niet consistent zijn van de itemparameterschatters buiten beschouwing laten, altijd met de MML of andere in dit boek niet besproken bayesiaanse methoden worden uitgevoerd. Het is een gemeenschappelijk kenmerk van deze methoden dat het gebruikte itemresponsmodel wordt aangevuld met een (of meer) verdeling(en) voor de vaardigheid. Laten we even aannemen dat we beschikken over slechts één populatie. De aanname van een vaardigheidsverdeling voor deze populatie betekent eigenlijk dat de vaardigheid van de personen niet meer vast of fixed is, maar random, dat wil zeggen getrokken uit een bepaalde, al dan niet compleet gespecificeerde, vaardigheidsverdeling. Tijdens de calibratie moeten dan zowel de itemparameters als de (eventuele) parameters van de vaardigheidsverdeling gezamenlijk geschat worden. Het model geldt dus alleen onder de extra aanname van deze vaardigheidsverdeling. Aan de ene kant kunnen we nu stellen dat we bij de schatting van de vaardigheid van individuele personen rekening dienen te houden met het feit dat ze getrokken zijn uit een bepaalde populatie met een onderliggende verdeling. Maar dit betekent dat we de vaardigheid met een bayesiaanse methode moeten bepalen. De EAP-methode komt dan in aanmerking. Als we namelijk bij de schatting van de vaardigheidsparameter géén gebruik maken van deze onderliggende verdeling, dan gebruiken we niet alle beschikbare informatie, zodat deze

schatting statistisch niet optimaal kan zijn. Aan de andere kant kunnen we ook stellen dat de calibratie alleen maar dient om de itemparameters te schatten. De aanname van een vaardigheidsverdeling was alleen maar noodzakelijk om de schaal vast te leggen. Bij de schatting van de vaardigheid hoeven we hier dus geen rekening meer mee te houden. In de praktijk wordt bijna altijd gekozen voor de tweede optie. Er wordt dan dus géén rekening gehouden met de onderliggende vaardigheidsverdeling en het informatieverlies wordt op de koop toe genomen. In concreto betekent dit dat de vaardigheidsparameter θ gewoon met de ML- of WML-methode geschat wordt. In modellen met voldoende statistieken voor de vaardigheid kan de calibratie uitgevoerd worden met zowel CML als MML. Als we gecalibreerd hebben met CML, een methode die steekproefonafhankelijk is, kunnen we de vaardigheid schatten met de ML- of WML-methode. Als de calibratie met MML is geschied, geldt hetzelfde als in modellen zonder voldoende statistieken, zoals hiervoor uiteengezet. Ook dan worden ML- of WML-schattingen voor de vaardigheid gebruikt.

Als we bij de schatting van de vaardigheidsparameter géén gebruik (wensen te) maken van populatiegegevens, dan gaat, voor elk itemresponsmodel, de voorkeur uit naar WML-schatters, daar deze, bij benadering, zuivere schatters van de vaardigheid opleveren (zie hoofdstuk 4). Zoals bekend zal de nauwkeurigheid van deze schatters (standaardfout kleiner) en dus van de equivalering toenemen naarmate de moeilijkheid van de toets dichter bij de te schatten vaardigheid ligt.

Ware score equivalering

Bij het equivaleren van bestaande toetsen, en soms ook als men toetsen samenstelt uit een itembank, wenst men na equivalering te rapporteren naar de gebruiker op de (eventueel nog te transformeren) ruwe score schaal, dat wil zeggen het aantal items goed. Schattingen op de vaardigheidsschaal hebben daar niet altijd een direct verband mee. Als we toetsen beschouwen met dichotome items en als IRT-model het twee- of drieparametermodel, dan levert elk verschillend antwoordpatroon een verschillende schatting van de vaardigheid op. Ter illustratie beschouwen we een voorbeeld. We hebben de gegevens geanalyseerd van een subtoets van de zogenaamde Scholastic Apptitude Test (LSAT-6), die vermeld staan in Mislevy en Bock (1986). Deze subtoets bestaat uit vijf items. Met de antwoorden van 1000 personen werd een calibratie uitgevoerd met het tweeparametermodel en met het Raschmodel. Vervolgens werden de vaardigheden van deze personen geschat met de EAP-methode. Een deel van de

resultaten staat in tabel 8.1, en wel de EAP-schattingen voor personen die 3 of meer scoorden op deze toets.

Tabel 8.1
EAP-vaardigheidschattingen tweeparametermodel en Raschmodel LSAT-6

Tweeparametermodel				Raschmodel		
Patroon	score	aantal	EAP	score	aantal	EAP
00111	3	4	-.314	3	237	-.331
01011	3	16	-.395			
01101	3	3	-.296			
01110	3	2	-.275			
10011	3	81	-.366			
10101	3	28	-.266			
10110	3	15	-.245			
11001	3	56	-.347			
11010	3	21	-.326			
11100	3	11	-.226			
01111	4	15	.062	4	357	.063
10111	4	80	.093			
11011	4	173	.008			
11101	4	61	.112			
11110	4	28	.134			
11111	5	298	.498	5	298	.477

We zien dat, als we het tweeparametermodel gebruiken, voor elk antwoordpatroon een andere schatting voor de vaardigheid volgt. Dit in tegenstelling tot als we het Raschmodel gebruiken: in dat model is immers de somscore een voldoende statistiek voor θ, en krijgen we alleen voor verschillende somscores verschillende vaardigheids-schattingen. Voor de volledigheid zij vermeld dat de schattingen in tabel 8.1 gerapporteerd staan op een schaal, die genormeerd is op de vaardigheidsverdeling. Deze verdeling heeft een gemiddelde van 0 en een standaarddeviatie van .075.

Bij het tweeparametermodel, en in het algemeen met modellen die geen voldoende statistiek voor θ hebben, is er dus geen directe relatie tussen de geschatte vaardigheden en de (eventueel gewogen) ruwe score schaal. Deze schattingen hebben dus ook geen

directe relatie met de ruwe scores van de te equivaleren toetsen. Als men de te equivaleren toetsen op de ruwe score schaal zou willen rapporteren, komt men met de geschatte vaardigheden niet verder. Een werkwijze die men dan kan toepassen is ware score equivalering, die als volgt werkt.

Men definieert de ware score op een toets, vergelijkbaar met de ware score in de KTT, als de verwachtingswaarde van de ruwe score:

$$\tau_X = \mathscr{E}(X) = \mathscr{E}(\sum_{i \in X} X_i) = \sum_{i \in X} \mathscr{E}(X_i) = \sum_{i \in X} P_i(\theta), \tag{8.22}$$

waarbij $P_i(\theta)$ de kans op een goed antwoord onder het gebruikte IRT-model is. Het is eenvoudig in te zien, dat bij dichotome items de ware score precies het bereik heeft van de ruwe score schaal. De ware score (8.22) als functie van θ beschouwd, wordt ook wel de toetskarakteristieke functie genoemd en is de som van de itemresponsfuncties van de items waaruit de toets bestaat. Een schatting van de ware score van een persoon op een toets verkrijgt men door het evalueren van (8.22) in het punt van de schatting van de persoon op de vaardigheidsschaal $\hat{\theta}: \hat{\tau}_X = \Sigma_{i \in X} P_i(\hat{\theta})$.

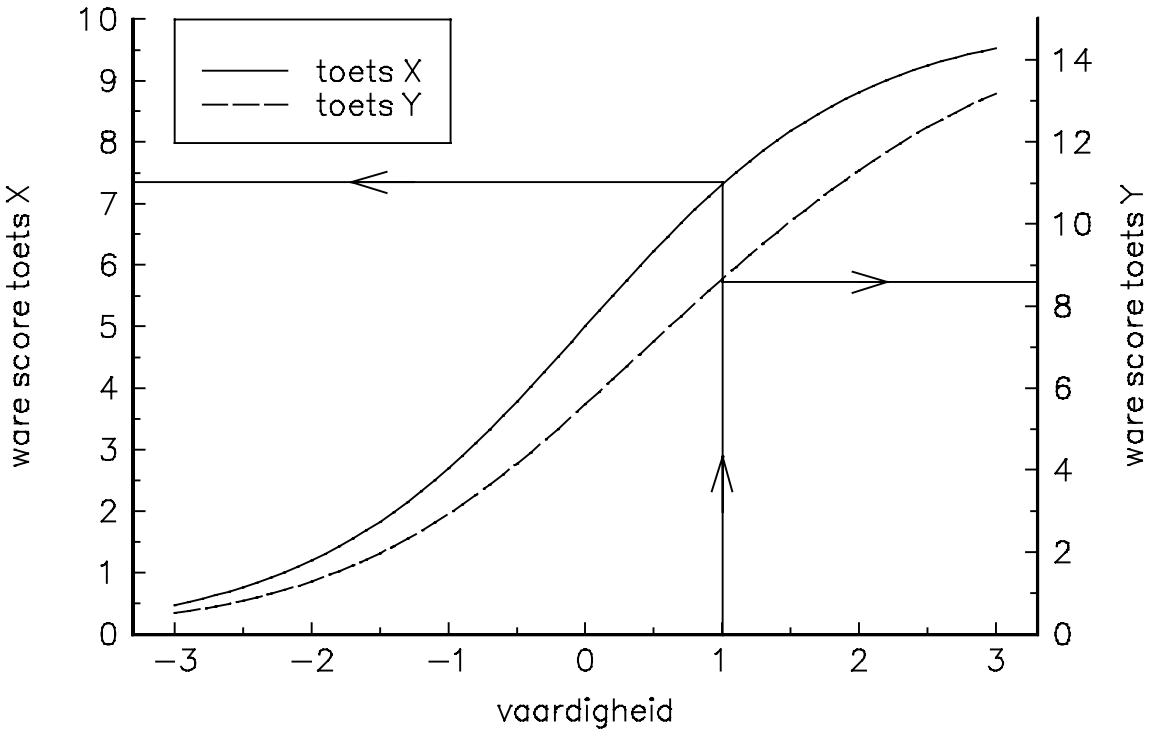
Als we nu twee toetsen X en Y hebben die gecalibreerd zijn onder een IRT-model, dan kan men de geschatte ware scores op beide toetsen die horen bij een bepaalde θ als geëquivaleerde scores beschouwen. Voor de te equivaleren toetsen X en Y zijn de ware scores als functie van θ gegeven door

$$\begin{aligned} \tau_X &= \sum_{i \in X} P_i(\theta) \quad \text{en} \\ \tau_Y &= \sum_{j \in Y} P_j(\theta). \end{aligned} \tag{8.23}$$

Voor elke θ en dus ook voor elke schatting van $\hat{\theta}$ van θ zijn dan de ware scores en dus ook de geschatte ware scores $\hat{\tau}_X$ en $\hat{\tau}_Y$ equivalent. Met een voorbeeld zullen we dit toelichten. In figuur 8.7 staan de toetskarakteristieke functies van toets X, bestaande uit 10 items, en toets Y, die uit 15 items bestaat. Als voorbeeld is aangegeven dat bij $\theta = 1$ de ware score op toets X gelijk is aan 7.35 is en voor toets Y gelijk aan 8.29, de equivalente scores op deze toetsen bij deze waarde van θ. Voor elke θ kunnen we op deze manier equivalente scores op de toetsen vinden.

In de praktijk gebruikt men ware score equivalering ook nog wel eens op de volgende manier. Stel dat men toets Y wil equivaleren met een vroegere versie toets X en men wil weten wat de equivalente score is van een ruwe geobserveerde score op toets Y op de ruwe score schaal van toets X. Men wil dan dus ruwe geobserveerde scores equivaleren. In plaats van de ware score op toets Y gebruikt men dan de

geobserveerde ruwe score en zoekt daarbij de bijpassende score op de schaal van toets X. Als voorbeeld in figuur 8.7 vinden we dan bij een score 6 op toets Y een score van 5.2 op toets X. Alhoewel er theoretisch geen recht- vaardiging is voor het op deze manier equivaleren van geobserveerde scores, blijkt het in de praktijk redelijke resultaten op te leveren (Lord & Wingersky, 1983). Merk op dat voor het Raschmodel ware score IRT equivalering identiek is aan deze vorm van geobserveerde score IRT equivalering. Bij elke geobserveerde ruwe score hoort in het Raschmodel immers maar één schatting $\hat{\theta}$.



Figuur 8.7
Ware score equivalering van twee toetsen X en Y

8.3.3 Equivaleren met behulp van een itembank

In deze paragraaf behandelen we een voorbeeld van de opbouw van een itembank, dat wil zeggen het calibreren en het samenstellen van geëquivaleerde toetsen uit de bank. Dit concrete voorbeeld betreft de schaal vorderingen in spellingvaardigheid (SVS; Van den Bosch, Gillijns, Krom & Moelands, 1991). De SVS is een instrument om (vorderingen in) spellingvaardigheid te meten voor de groepen drie en vier van het

basisonderwijs. Na proefafnames zijn er negen verschillende modules samengesteld, elk van ongeveer 20 items. Daarna zijn deze modules afgenomen bij een landelijke steekproef middels het (longitudinale) design zoals gegeven in figuur 8.8. Boekje 1 bijvoorbeeld, dat is samengesteld uit de modules 1 en 2, is afgenomen op tijdstip m3 (medio groep 3) bij sag a. Een sag is een school afname groep en dient ter vereenvoudiging van de afname procedure; elke school in een sag maakt per afnametijdstip één boekje. Merk op dat binnen elk tijdstip het design verbonden is. Bovendien is het design over de tijdstippen heen verbonden en is het afnameschema zo geconstrueerd dat geen enkele leerling twee maal dezelfde module maakt, waardoor herinneringseffecten vermeden worden. Module 3 bijvoorbeeld, is op het eerste tijdstip (m3) gemaakt door leerlingen uit sag b en sag c, en een tijdstip later (e3, eind groep 3) door leerlingen uit sag a. Of, andersom bekeken, leerlingen

			Module								
			1	2	3	4	5	6	7	8	9
boekje	sag	tijd									
1	a	m3	■	■							
2	b			■	■						
3	c		■		■						
4	a	e3			■	■					
5	b					■	■				
6	c			■		■					
7	a	m4					■	■			
8	b							■	■		
9	c						■		■		
10	a	e4							■	■	
11	b									■	■
12	c							■		■	

Figuur 8.8
Calibratiedesign Spellingvaardigheid

uit sag a maken op de verschillende afnametijdstippen achtereenvolgens de modules 1+2, 3+4, 5+6 en 7+8, nooit dezelfde dus. Merk bovendien op dat een module die het design voor twee aanliggende tijdstippen verbindt, alleen op die twee tijdstippen is ingezet. Er is dus geen vast anker gebruikt (zie ook paragraaf 8.3.1). Omdat het voor rapportage- en onderwijskundige doeleinden het noodzakelijk was om over genoeg

gegevens omtrent de spelling van allochtone leerlingen te beschikken, zijn binnen elke sag de scholen met relatief veel allochtone leerlingen oververtegenwoordigd. Dit heeft als belangrijke consequentie dat voor een willekeurig gekozen tijdstip de steekproef niet meer representatief is voor de populatie op dat tijdstip. Bepaalde groepen zijn oververtegenwoordigd en de leerlingen zijn ook nog eens in clusters (scholen) getrokken. Uit de proefafname was bovendien bekend dat een goede beschrijving van de antwoorden op de items mogelijk was als we gebruik maakten van het OPLM. Om dezelfde reden als in paragraaf 7.1, geven we dan de voorkeur aan een calibratie met de CML-methode, deze methode is immers steekproefonafhankelijk. Alle (173) afgenomen items bleken op de SVS schaal te passen. In deze schaal zitten dus bijvoorbeeld geen items meer die tijdstip-onzuiverheid vertonen. Voor elke leerling die een bepaald boekje gemaakt heeft, kunnen we nu aan de hand van zijn toetsscore een schatting van zijn vaardig- heid maken. Deze vaardigheidsschattingen gebruiken we op verschillende manieren. De eerste, en meest belangrijke, is voor de bepaling van referentiegegevens. Deze referentie- gegevens worden per tijdstip zowel voor de totale populatie als ook voor de subpopulatie van allochtonen bepaald; de procedure hiervoor staat beschreven in hoofdstuk 10. Merk op dat bij de bepaling van de referentiegegevens op populatieniveau, er rekening mee gehouden dient te worden dat de allochtonen in de steekproef oververtegenwoordigd waren. Bovendien worden de vaardigheidsschattingen van de leerlingen naar de scholen die aan de calibratie hebben deelgenomen gerapporteerd.

Nadat de itembank SVS was geconstrueerd, zijn er voor elk afnametijdstip modules op maat samengesteld. Hiermee kan de leerkracht een leerling een toets voorleggen die meer toegespitst is op zijn of haar vaardigheid. De minder goede leerling krijgt dan een makkelijke en de goede leerling een moeilijke module. De belangrijkste reden voor dit toetsen op maat is dat de schattingsfouten van de vaardigheid flink kleiner worden. Bij WML, bijvoorbeeld, worden de schattingsfouten gemiddeld ongeveer dertig procent kleiner. Omdat de itembank gecalibreerd is, zijn de vaardigheidsschattingen op de verschillende modules gelijk geëquivaleerd. Bovendien kunnen deze geëquivaleerde scores direct gerelateerd worden aan de referentiegegevens: we kunnen nu immers de relatieve positie van de leerling in de betrokken populatie bepalen (zie ook hoofdstuk 10). Ook kan de vaardigheid van de leerling gerelateerd worden aan relevante onderwijskundige criteria (Van den Bosch e.a., 1991).

Een laatste opmerking. Omdat we werken met OPLM, zullen voor een juiste afspiegeling van de vaardigheid gewogen scores gebruikt moeten worden. In de praktijk wordt er door de leerkracht, voor wie de SVS als hulpmiddel dient, voornamelijk gebruik gemaakt van ongewogen (ruwe) scores. Er is daarom dan ook een procedure

ontwikkeld die aan dit probleem tegemoet komt. We zullen hier verder echter niet op ingaan.

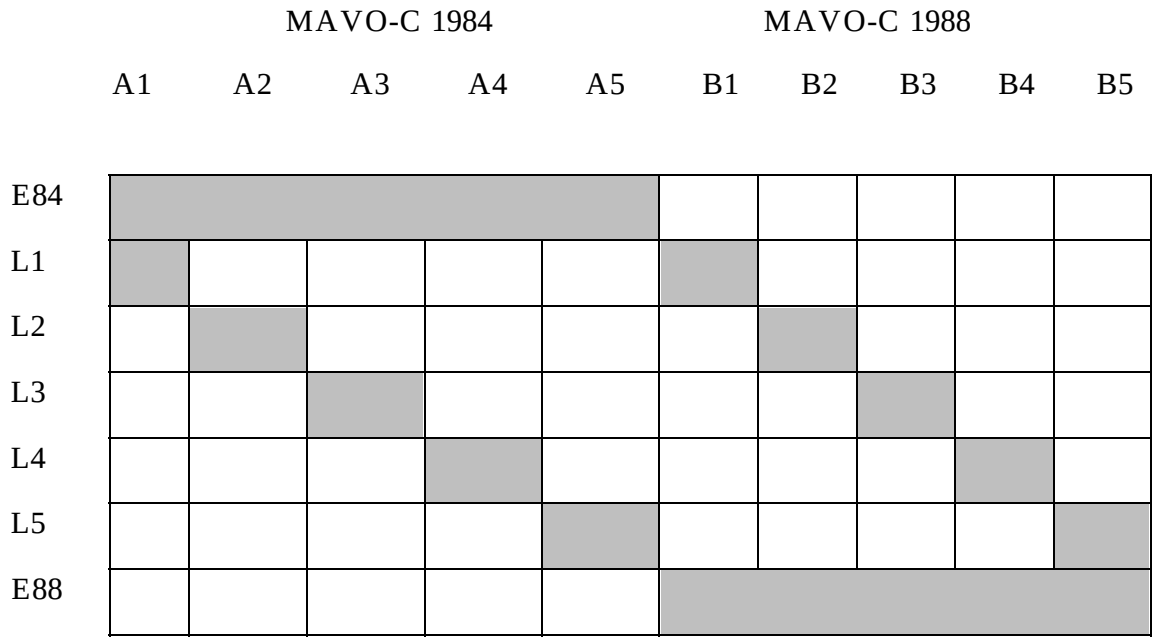
8.3.4 Quasi-multidimensionaal IRT-equivaleren

Zoals reeds in de inleiding is opgemerkt worden elk jaar de twee tijdvakken van een aantal centraal schriftelijke examens geëquivaleerd. Maar hoe zit dat nu met de examens over de jaren heen? Is het eindexamen van 1992, zeg, vergelijkbaar met dat van 1993? Dit is niet alleen een moeilijk maar ook, zeker voor belanghebbenden zoals leerlingen en onderwijsgevenden, een belangrijk probleem. In het vervolg zullen we ons voor het gemak beperken tot examens waarbij de items dichotoom gescoord worden. Een eerste opmerking die hier van belang is, betreft de scoringsregel die bij de examens gehanteerd wordt. Bij de examens moet de behaalde score een functie zijn van het aantal goed gemaakte opgaven. Bovendien moet elke opgave 'even zwaar' meetellen in het eindresultaat. Dit heeft als belangrijkste consequentie dat er een beperking op het te kiezen IRT-model ligt: alleen modellen met gelijke discriminatie-parameters komen in aanmerking. Het enige model dat dan nog over blijft is het Raschmodel. Voor de calibratie-methode komen dan zowel MML als CML in aanmerking. Bovendien zijn we bij examens behalve in equivalente scores over verschillende jaren ook in het slagingspercentage geïnteresseerd. Dit betekent dat we graag willen weten hoeveel procent van de kandidaten uit 1993 zou geslaagd zijn als ze het examen van 1992 gemaakt hadden. Daar dit laatste een kenmerk van de populatie is, ligt het voor de hand om de calibratie uit te voeren met MML.

Hoe de equivalering van twee examens uitgevoerd kan worden, zullen we demonstreren aan de hand van een voorbeeld. Als voorbeeld nemen we de examens frans van de jaren 1984 en 1988 voor MAVO-C. Eerst zijn beide examens in vijf delen geknipt. Voor het 1984 examen noemen we deze delen A1 tot A5 en voor het examen van 1988 duiden we deze delen aan met B1 tot B5. Vervolgens zijn deze delen, net na de afname van het examen in 1988, volgens het design in figuur 8.8 afgenomen bij een steekproef van leerlingen uit klas 3 van het VWO. De groepen L1 tot L5, allen uit klas 3 van het VWO, maken dus steeds een gedeelte van het 1984 en een gedeelte van het 1988 examen. Het ligt namelijk in de lijn der verwachting dat de vaardigheid van deze leerlingen vergelijkbaar is met de vaardigheid van de eindexamen kandidaten in MAVO-C (Glas, 1989).

Nu valt het niet te verwachten valt dat beide examens op een unidimensionale schaal liggen, omdat examens immers van de kandidaten diverse 'vaardigheden' vragen. Dit betekent dan ook dat het Raschmodel voor de totale itemverzameling naar verwachting niet zal passen, wat in werkelijkheid ook zo bleek te zijn. Daarom is gezocht naar een multi-dimensionale oplossing voor het equivaleerprobleem. Het bleek namelijk dat de totale itemverzameling op te splitsen was in een aantal subschalen die alle aan het Raschmodel voldeden. De gebruikte procedure om tot deze subschalen te komen werkt als volgt. Eerst moeten de vaardigheids-verdelingen gespecificeerd worden. Voor elk van de drie onderscheiden groepen, te weten de examen kandidaten van 1984 (E84), leerlingen uit klas 3 van het VWO (L1-L5) en de examen kandidaten van 1988 (E88) nemen we een normale verdeling aan. De schaal wordt vastgelegd door het gemiddelde van de vaardigheidsverdeling van de 1984 examinandi gelijk aan nul te stellen.

We gaan nu de eerste subschaal zoeken. Dit doen we door uit de totale set van items die items te verwijderen die op basis van de itemgerichte R_{1m} toets niet blijken te passen. Dit doen we net zo lang totdat er een schaal gevonden is. Bij deze schaal kunnen dus geen items meer verwijderd worden op basis van de R_{1m} toets. Deze unidimensionale Raschschaal noemen we subschaal 1. Vervolgens zoeken we de tweede subschaal op precies dezelfde



Figuur 8.8
Equivaleerdesign MAVO-eindexamen

manier als hierboven uit de overgebleven items, dat wil zegen uit de totale set van items behalve de items uit subschaal 1. Uiteindelijk werden er drie subschalen gevonden en bleken slechts vier items (alle uit 1984) van de in totaal 100 items op geen enkele subschaal te passen. Het blijkt dus dat we zelfs met dit multidimensionale itemresponsmodel niet alle items kunnen schalen. We zouden dus nu eigenlijk een ruimer IRT model moeten kiezen. Dit is mogelijk, daar er voor dit soort items modellen bestaan waarbij een item op meerdere vaardigheidsdimensies laadt, zie bijvoorbeeld paragraaf 5.5. Voor de beschrijving van dit voorbeeld zullen we echter aannemen dat de calibratie met succes is afgesloten, de vier niet passende items ten spijt. We beschikken nu over drie subschalen met per subschaal drie vaardigheidsverdelingen, voor elk van de onderscheiden groepen leerlingen één. De linking groepen, dat wil zeggen de leerlingen uit klas 3 van het VWO, zijn nu verder niet meer van belang, daar deze alleen maar dienden om het design te verbinden.

Uiteindelijk hebben we op deze manier nu precies een model zoals beschreven in paragraaf 5.5. Merk op dat elk examen uit drie subschalen bestaat, een leerling heeft op elke subschaal een vaardigheid. Laten we eens aannemen dat een leerlinge 43 items goed beantwoord heeft van het 1984 examen. Deze score van 43 kan op zeer veel verschillende manieren tot stand gekomen zijn. De leerlinge kan bijvoorbeeld van de eerste subschaal 20 items goed hebben, van de tweede 17 en van de laatste subschaal 6. Bij deze combinatie horen uiteraard drie vaardigheidsschattingen, op elke subschaal een. Omdat we bij de examens niet op de vaardigheidsschaal werken, moeten we dus deze vaardigheidschattingen gebruiken om op elke subschaal een equivalente score op dezelfde subschaal van 1988 examen te zoeken. Of, met andere woorden, op elke subschaal passen we ware score equivalering toe. Tenslotte berekenen we de equivalente score van deze leerlinge op het totale 1988 examen door de som van de drie geëquivaleerde scores (op de subschalen) te nemen. Het is eenvoudig in te zien dat voor een andere leerling met 43 items goed in 1984, best een andere geëquivaleerde score in 1988 gevonden kan worden.

Een van de belangrijkste waarden bij een examen is de cesuur, dat wil zeggen de score, waar de grens tussen een onvoldoende en een voldoende ligt. We kunnen nu de cesuur voor het 1988 examen berekenen op grond van de populatie uit 1984. Hiermee kunnen we dan gelijk de vraag beantwoorden hoeveel kandidaten uit 1984 voor het 1988 examen geslaagd zouden zijn. Daarvoor schatten we eerst voor elke 1984 leerling de vaardigheidsparameters $\hat{\theta}_{84q}$, $q = 1,...,3$, waarbij q de subschaal weergeeft. De somscore op het examen van 1988, r^*_{88}, wordt vervolgens geschat door

$$r^*_{88} = \sum_{q=1}^{3} \sum_{i \in I_q} \mathscr{E}(X_i \mid \hat{\theta}_{84q}, \hat{\delta}_q), \tag{8.23}$$

waarbij δ_q de itemparameters van het 1988 examen zijn en I_q die items die op subschaal q van het 1988 examen liggen. Bovenstaande formule geeft dus de verwachting van de score van een 1984 examinandus op het 1988 examen. Als we voor elke leerling (8.23) berekenen, en de cesuur van 1988 toepassen, kunnen we dus gelijk vaststellen hoeveel procent van de 1984 populatie in 1988 geslaagd zou zijn.

8.4 De kwaliteit van de equivaleermethoden vergeleken

Bij de beschrijving van de equivaleermethoden in dit hoofdstuk zijn soms voor- en nadelen genoemd. Dit is één bron om de kwaliteit van de methoden te vergelijken. De tweede is om terug te grijpen op de zeer omvangrijke psychometrische literatuur die de laatste jaren is verschenen en nog verschijnt over studies die tot doel hadden equivaleermethoden te vergelijken. Het is in dit verband niet zinvol om uitvoerig op deze studies in te gaan. Op de eerste plaats heeft dit te maken met de enorme hoeveelheid artikelen die over het onderwerp verschijnen; het volledig bespreken zou zeer veel tijd kosten. In de tweede plaats zijn deze studies vaak zeer specifiek toegespitst op één bepaald aspect van één equivaleermethode, zodat ze slechts geringe generalisatiemogelijkheden hebben. In de derde plaats is de kwaliteit van de artikelen vaak matig. De voorwaarden en aannamen waaronder een bepaalde techniek geldig is, worden zelden expliciet genoemd. Een veel voorkomende fout is bijvoorbeeld dat de kwaliteit van IRT equivalering als slecht wordt beoordeeld, terwijl het gehanteerde model niet past. In dit geval kan echter geen oordeel over de kwaliteit plaatsvinden, daar de equivalering slechts bij modelpassing kan worden uitgevoerd.

Een integratie van beide bronnen leidt tot de volgende conclusies. De eerste en belang- rijkste conclusie is dat equivaleren met behulp van de IRT in het algemeen de voorkeur heeft boven equivaleren met behulp van de KTT. Uiteraard moet dan bij het gebruik van een bepaald itemresponsmodel allereerst de modelgeldigheid nagegaan worden. De strenge eisen die bij de modeltoetsing worden opgelegd hebben als rechtstreeks gevolg dat de equivalering eenvoudig wordt. Als we over IRT equivaleren praten, zullen we steeds aannemen dat de calibratie met succes is afgesloten. Indien het gekozen itemresponsmodel echter niet past, en een ruimer model ook geen oplossing geeft, dan kunnen we altijd terugvallen op de KTT, welke immers minder stringente eisen aan de data stelt. In dat geval moeten we er ons echter wel bewust van zijn dat we nu meestal enkele niet toetsbare aannames en vooronderstellingen moeten maken.

De tweede conclusie is dat IRT equivaleermethoden eerder werken naarmate het aantal parameters groter is, omdat dan de modellen eerder passen. Het blijkt echter,

dat er voor itemresponsmodellen met veel parameters, zoals bijvoorbeeld het 3PL, geen goede toetsen beschikbaar zijn, behalve hele strenge toetsen. Denk hierbij bijvoorbeeld aan toetsen die met behulp van kruisvalidatie-technieken geconstrueerd kunnen worden (zie ook hoofdstuk 5).

De derde conclusie slaat alleen op equivaleermethoden binnen de KTT. Hier blijkt dat bij het gebruik van het single group design of het random group design alle equivaleermethoden, binnen praktisch relevante marges, overeen komen. Bij het ankertoetsdesign gelden ongeveer dezelfde conclusies, mits het anker aan de in dit hoofdstuk reeds besproken (psychometrische) voorwaarden voldoet en het aantal ankeritems groot genoeg is.

Tenslotte nog een laatste opmerking. In dit gehele hoofdstuk zijn schattingsfouten doorgaans buiten beschouwing gelaten. Enerzijds is dit gebeurt om het niet nodeloos ingewikkeld te maken, anderzijds omdat er slechts weinig analytische resultaten bekend zijn. In de literatuur worden de equivaleerfouten meestal gekarakteriseerd als systematisch en random. De systematische fouten zijn dan het rechtstreekse gevolg van het schenden van de assumpties. Als we bijvoorbeeld het random group design bekijken, dan kan het zo zijn dat de verschillende groepen niet vergelijkbaar zijn. Het moge duidelijk zijn dat systematische fouten ten alle tijden zoveel mogelijk vermeden dienen te worden. Daaruit volgt logischerwijs dat de assumpties op de een of andere manier getoetst moeten worden. Hoe deze assumpties, indien mogelijk, getoetst kunnen worden is beschreven bij de bespreking van de verschillende methoden. Merk op dat het toetsen van de assumpties voornamelijk een groot probleem is bij equivaleren in de KTT. Omdat we in de praktijk altijd met steekproeven werken waarmee populatie kenmerken geschat moeten worden, zullen we altijd statistische fouten maken (random equivaleerfouten). Om deze zo klein mogelijk te maken is het een eerste vereiste dat de steekproef voldoende groot is. Bovendien verdient het uiteraard aanbeveling om de steekproef af te stemmen op de te equivaleren toetsen. Dit laatste is voornamelijk een groot voordeel bij equivaleren in de IRT, bijvoorbeeld bij 'toetsen op maat'. Voor meer informatie omtrent (statistische) schattingsfouten als we equivaleren in de KTT, verwijzen we naar Braun en Holland (1982), Lord (1950) en Angoff (1971).

9

Vraagonzuiverheid

Onzuiverheid van vragen (in het Engels 'item bias' of 'differential item functioning', afgekort DIF) en onzuiverheid van tests of toetsen ('test bias') vormen in Amerika sinds het midden van de jaren 60 een belangrijk thema in 'educational measurement'. Door een aantal juridische zaken is dit onderwerp in Amerika in de jaren 80 ook sterk in de publieke belangstelling komen te staan. Een belangrijk geval daarbij vormt de rechtszaak die verzekeringsmaatschappij Golden Rule in 1976 tegen Educational Testing Service (ETS) aanspande. De aanklacht had betrekking op de negatieve gevolgen voor kleurlingen van het gebruik van bepaalde door ETS geconstrueerde toetsen voor het diploma van verzekeringsagent. In 1984 werd tussen ETS en de betreffende verzekeringsmaatschappij een schikking getroffen. Een belangrijk punt daarin was dat voor de constructie van twee specifieke toetsen uit dit examen bij de selectie van vragen zoveel mogelijk de voorkeur zou worden gegeven aan vragen die zo klein mogelijke verschillen in moeilijkheidsgraad vertoonden tussen de meerderheidsgroep en de verschillende ethnische groepen. Daarbij zou men vooral verschillen ten nadele van minderheidsgroepen trachten te voorkomen.

In Nederland werd in 1987 naar aanleiding van verschillende klachten door het Landelijk Bureau Racismebestrijding (LBR) een onderzoeksproject 'Psychologische tests en allochtonen' gestart. Gebleken was dat een aantal allochtone sollicitanten, die gekwalificeerd waren voor een functie waarnaar zij solliciteerden, door negatieve resultaten op bepaalde psychologische tests waren afgewezen. Uit een symposium van experts dat in dat jaar georganiseerd werd, kwam de volgende aanbeveling naar voren: "Psychologische tests moeten, willen ze gehanteerd worden in een selectieprocedure, gescreend zijn op 'cultural bias' en cultuurgebonden en racistische items" (LBR, 1988). Naar aanleiding hiervan werd door de Commissie Testaangelegenheden (COTAN) van het Nederlands Instituut van Psychologen en het LBR een commissie samengesteld met als taak om de twintig meest gebruikte tests op deze punten te screenen. In 1990 volgde het rapport van deze commissie waarin twintig van de in Nederland meest gebruikte psychologische tests voor de selectie voor opleiding en beroep op deze punten werden

doorgelicht (LBR, 1990). De belangrijkste conclusie uit dit rapport was dat: "alle gescreende tests sterk beperkt toepasbaar zijn bij allochtonen" en de commissie beval voor veel van de tests een "grondige revisie aan vanwege hun ethnocentristische inhoud" aan. Verder constateerde de commissie een "ernstige achterstand in Nederland op het gebied van onderzoek naar test en item bias".

Onder andere op grond van de hierboven genoemde overwegingen wordt er op het Cito de nodige aandacht besteed aan onderzoek naar onzuiverheid. Een andere overweging is dat in verschillende onderzoeken bij examens en toetsen opvallende verschillen tussen sociale groepen en geslachtsverschillen gevonden zijn, hetgeen de vraag naar de rol van de toetsen of toetsvragen zelf daarin relevant maakt. Zo zijn er verschillende onderzoeken naar vraagonzuiverheid uitgevoerd met betrekking tot allochtonen bij de Eindtoets Basisonderwijs (Uiterwijk, 1990) en bij de eindexamens voortgezet onderwijs met betrekking tot sexe (Bügel, 1993).

Onzuiverheid van tests of vragen hoeft niet alleen betrekking te hebben op bepaalde sociale groepen maar kan ook als onderdeel van een meer algemeen probleem beschouwd worden. In het kader van het meten van leerprestaties kan men bijvoorbeeld ook de onzuiverheid van toetsen of toetsvragen ten opzichte van verschillende onderwijsmethoden beschouwen.

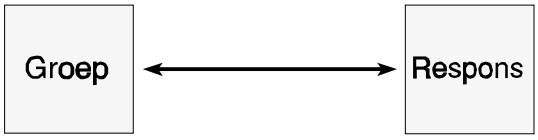
Hoewel in dit hoofdstuk ook enige aandacht aan testonzuiverheid zal worden besteed, vormt vraagonzuiverheid het belangrijkste onderwerp. In de literatuur zijn verschillende onderzoeksmethoden voor het opsporen van vraagonzuiverheid beschreven. Bij de bespreking van dergelijke methoden zullen we ons in dit hoofdstuk voornamelijk concentreren op onderzoek met behulp van IRT-modellen.

Dit hoofdstuk is als volgt opgebouwd. In paragraaf 9.1 wordt een definitie van het begrip onzuiverheid gegeven. In paragraaf 9.2 wordt deze definitie vertaald naar een aantal technieken voor het opsporen en aantonen van vraagonzuiverheid. In paragraaf 9.3 zal de toepassing van deze technieken aan de hand van een voorbeeld worden geïllustreerd.

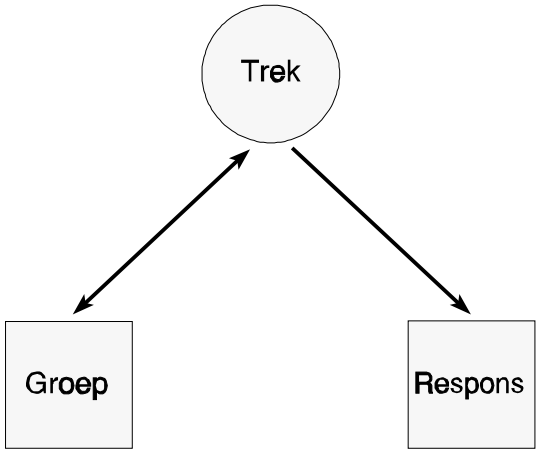
9.1 Definitie van onzuiverheid

Een algemene omschrijving van het begrip onzuiverheid, die zowel van toepassing is op het niveau van tests als van vragen, wordt gegeven door Mellenbergh (1985). In deze omschrijving wordt uitgegaan van een samenhang tussen groepslidmaatschap en de respons op een vraag of de score op een test. Men kan hierbij bijvoorbeeld denken aan het verband tussen het al dan niet behoren tot de groep autochtone leerlingen en de

score op een schooltoets. De relatie tussen groepslidmaatschap en de respons op een item of een toetsscore wordt in figuur 9.1 schematisch weergegeven, waarbij de geobserveerde variabelen (groepslidmaatschap en de respons) zijn aangegeven als blokken en de samenhang tussen die variabelen is aangeduid als een pijl met twee punten. Deze pijl geeft aan dat er sprake is van een samenhang tussen de variabelen en niet van een specifieke invloed van de ene variabele op de andere.



Figuur 9.1
Samenhang tussen groepslidmaatschap en respons



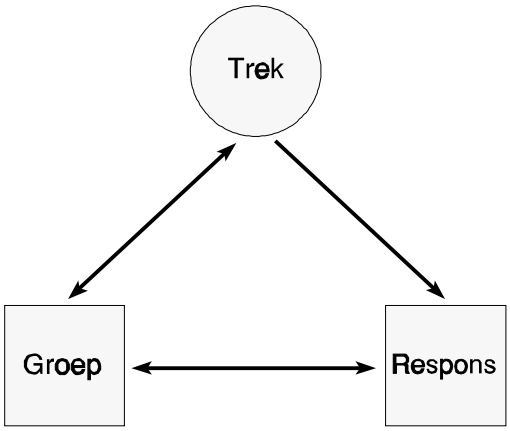
Figuur 9.2
Een zuivere vraag of test

Een dergelijke samenhang tussen groepslidmaatschap en respons hoeft niet te duiden op onzuiverheid van de vraag of de test, maar kan ook het gevolg zijn van werkelijke niveauverschillen tussen de betreffende groepen. Dit wordt weergegeven in figuur 9.2. Daar wordt de samenhang tussen het groepslidmaatschap en de respons geheel verklaard door een latente, niet direct geobserveerde variabele, een latente trek. De latente variabele is weergegeven als een cirkel en de invloed van deze variabele op de respons met een pijl met één punt. Omdat de verschillen op de vraag of de test veroorzaakt zijn door werkelijke vaardigheidsverschillen spreekt men van een zuivere vraag of test.

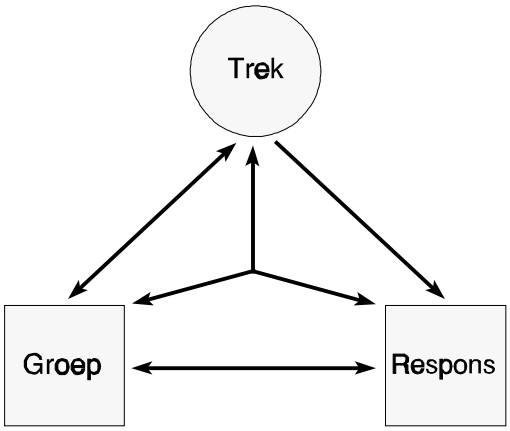
Er is sprake van een onzuivere vraag of test als de verschillen tussen de groepen niet helemaal verklaard kunnen worden door verschillen op de latente vaardigheidsdimensie. Dit wordt weergegeven in figuur 9.3, waar naast de samenhang tussen het groepslidmaatschap en de latente trek en de invloed van de latente trek op de respons nog steeds een directe samenhang blijft bestaan tussen het groepslidmaatschap en de

respons. Wanneer deze verschillen in prestaties tussen de groepen voor alle niveaus van de latente trek gelijk zijn, spreekt men van een uniform-onzuivere vraag of test.

Het is echter ook mogelijk dat de verschillen tussen de groepen variëren over de verschillende niveaus van de latente trek. Dit is bijvoorbeeld het geval als bij een laag vaardigheidsniveau de ene groep leerlingen hoger scoort terwijl bij een hoog vaardigheidsniveau de andere groep leerlingen hoger scoort. In deze situatie spreekt men van een niet-uniform onzuivere vraag. Niet-uniforme onzuiverheid wordt weergegeven in figuur 9.4, waarbij de drie pijlen vanuit het midden aangeven dat er sprake is van een samenhang tussen groepslidmaatschap en de respons welke gerelateerd is aan het niveau van de latente trek (een samenhang tussen de drie variabelen samen).



Figuur 9.3
Een uniform onzuivere vraag of test

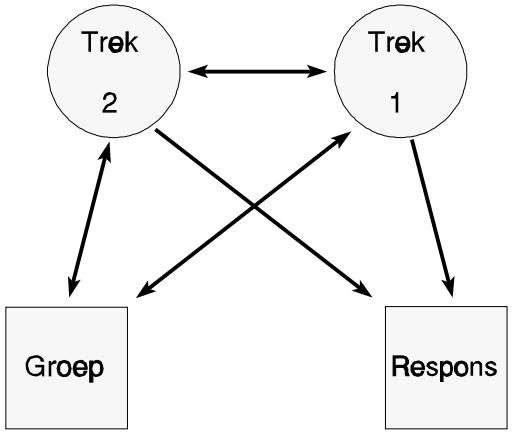


Figuur 9.4
Een niet-uniform onzuivere vraag of test

Tenslotte wordt in figuur 9.5 de situatie weergegeven waarbij de onzuiverheid verklaard wordt door het beschouwen van een tweede latente variabele, die niet tot de oorspronkelijke meetpretentie van het betreffende instrument hoort. Deze tweede latente variabele veroorzaakt de samenhang tussen het groepslidmaatschap en de respons. Na het toevoegen van deze trek is de samenhang tussen de geobserveerde variabelen, het groepslidmaatschap en de respons, verdwenen.

Wat betreft de hiervoor gegeven algemene beschrijving van het begrip onzuiverheid is het niet van belang of de geobserveerde respons op één of enkele vragen van een test, of op een hele test betrekking heeft. Bij het ontwikkelen van een methodologie voor het opsporen en aantonen van onzuiverheid is het daarentegen wel relevant of een test in zijn geheel onzuiver is, of dat slechts enkele vragen onzuiver zijn. Als een test

in z'n geheel onzuiver is, moet men om het groepseffect te kunnen evalueren namelijk over een additionele meting beschikken die wel zuiver is. Bij deze additionele meting moeten de groepsverschillen voldoende verklaard worden door verschillen op de latente trek. Wanneer de assumptie van normaliteit van de testscores aannemelijk kan worden gemaakt doordat bijvoorbeeld het scorebereik van de test voldoende groot is zodat de variabelen bij benadering continu zijn, kunnen variantie- of factoranalytische modellen worden toegepast. In het geval van één of enkele onzuivere vragen ligt het probleem anders, omdat daar naast de onzuivere ook zuivere vragen in de test aanwezig zijn. Aangezien de scores op testvragen echter meestal dichotoom of polytoom zijn, zal de assumptie van normaliteit per vraag meestal niet aannemelijk kunnen worden gemaakt. De itemresponstheorie levert in dat geval een meer geëigende context voor het ontwikkelen van een methodologie voor het opsporen en aantonen van onzuiverheid.



Figuur 9.5

Een onzuivere vraag of test waarbij onzuiverheid veroorzaakt wordt door één extra latente variabele

9.2 Methoden voor het bepalen van vraagonzuiverheid

In het onderzoek naar onzuiverheid is het gebruikelijk onderscheid te maken tussen een referentiegroep, zeg de meerderheidsgroep, en de potentieel benadeelde groep, die wordt aangeduid als de doelgroep. Wanneer bijvoorbeeld onzuiverheid als gevolg van culturele verschillen onderzocht wordt, bestaat de referentiegroep over het algemeen

uit autochtone en de doelgroep uit allochtone leerlingen. Deze terminologie zal ook in het vervolg van dit hoofdstuk gehanteerd worden.

Als we de theorie uit de vorige paragraaf vertalen naar dichotome items, is vraagonzuiverheid of DIF te definiëren als de omstandigheid dat bij een gegeven vaardigheidsniveau twee willekeurige leden van twee verschillende populaties niet dezelfde kans hebben om een vraag goed te maken. De statistische technieken voor het opsporen van DIF zijn dan ook alle gebaseerd op het evalueren van verschillen tussen de groepen in de kansen op een goed antwoord, conditioneel op een of andere maat voor vaardigheid. Meestal neemt men als maat voor de vaardigheid de somscore van de leerlingen. De meest algemeen toegepaste technieken zijn gebaseerd op de Mantel-Haenszel-toets (Holland & Thayer, 1988) of op IRT-modellen (Hambleton & Rogers, 1989; Kelderman, 1989). In de volgende twee paragrafen worden deze twee benaderingen toegelicht, in de daaropvolgende paragraaf worden zij met elkaar vergeleken. Daarna zal een concreet voorbeeld van het opsporen van vraagonzuiverheid met een itemresponsmodel worden gegeven.

9.2.1 De Mantel-Haenszel-procedure

Holland en Thayer (1988) stellen de volgende procedure voor om vast te stellen of de verschillen tussen de groepen in de moeilijkheidsgraad van een item, conditioneel op de somscores van de leerlingen, statistisch significant zijn. Voor elke niveaugroep, dat wil zeggen voor elke groep leerlingen met een score in een bepaald bereik, wordt een 2x2-tabel van itemscore bij groepslidmaatschap opgesteld. De tabel is weergegeven in figuur 9.6, waarbij in de cellen de aantallen personen staan aangegeven.

	Score op item i		
	1 (goed)	0 (fout)	Totaal
Referentiegroep	a_q	b_q	n_{1q}
Doelgroep	c_q	d_q	n_{2q}
Totaal	m_{1q}	m_{0q}	n_q

Figuur 9.6
2x2-tabel van niveaugroep q

Betekenis van de symbolen in figuur 9.6:

n_q totaal aantal kandidaten in niveaugroep q;

a_q personen in de referentiegroep bij niveaugroep q die item i juist beantwoord hebben;

b_q personen in de referentiegroep bij niveaugroep q die item i onjuist beantwoord hebben;

c_q personen in de doelgroep bij niveaugroep q die item i juist beantwoord hebben;

d_q personen in de doelgroep bij niveaugroep q die item i onjuist beantwoord hebben.

De door Holland en Thayer voorgestelde procedure is gebaseerd op een zogenaamde 'odds-ratio' (ratio van kansen) α_q. Deze wordt geschat door

$$\hat{\alpha}_q = \frac{p_{1q}/(1-p_{1q})}{p_{2q}/(1-p_{2q})} = \frac{a_q d_q}{b_q c_q}, \tag{9.1}$$

waarbij p_{1q} de kans op een goed antwoord is van de referentiegroep en p_{2q} de kans op een goed antwoord van de doelgroep. Wanneer de prestaties van beide groepen niet verschillen, is $\hat{\alpha}_q$ gelijk aan 1. In het geval de twee groepen verschillende antwoordpatronen vertonen, is $\hat{\alpha}_q$ groter dan 1 wanneer de referentiegroep een grotere kans op een goed antwoord heeft en $\hat{\alpha}_q$ kleiner dan 1 wanneer dit voor de doelgroep geldt. Voor de Mantel-Haenszel-toets worden de Mantel-Haenszel-statistieken van alle niveaugroepen gecombineerd tot

$$\hat{\alpha}_{MH} = \frac{\sum_q a_q d_q / n_q}{\sum_q b_q c_q / n_q}. \tag{9.2}$$

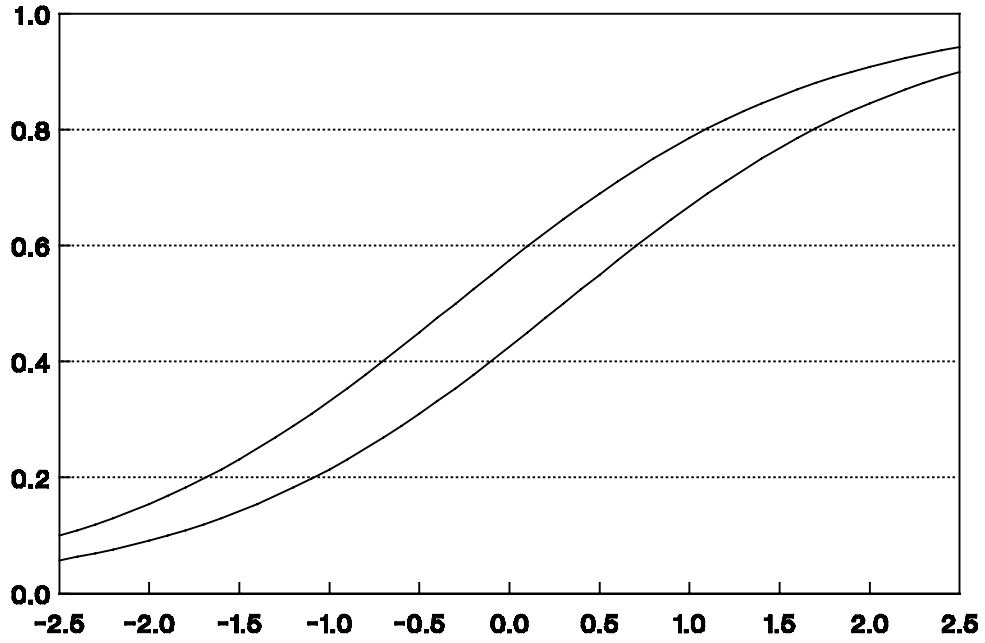
Indien er in de populaties geen DIF voorkomt en dus $\alpha_{MH} = 1$, kan aangetoond worden dat $\log\hat{\alpha}_{MH}$ normaal verdeeld is met een gemiddelde nul en standaarddeviatie $\mathrm{SE}(\log\hat{\alpha}_{MH})$, zodat de gestandaardiseerde log-odds-ratio $z = \log\hat{\alpha}_{MH}/\mathrm{SE}(\log\hat{\alpha}_{MH})$ bij benadering standaard-normaal is verdeeld. Bij een significantie-niveau van 1%, zijn de kritische waarden $z \geq 2.58$ als het item gemakkelijker is in de referentiepopulatie en $z \leq -2.58$ als het item moeilijker is in de referentiepopulatie.

De aanwezigheid van items met DIF doet afbreuk aan de waarde van de somscore als indicator van de vaardigheid van de leerlingen. De somscore wordt immers mede bepaald door items die voor de twee groepen een verschillende moeilijkheidsgraad hebben. Daarom is het zoeken naar DIF een iteratief proces. Eerst wordt een analyse

uitgevoerd waarbij de antwoorden op alle items worden opgenomen in de somscore. Vervolgens wordt er een analyse uitgevoerd waarbij de items die in de eerste analyse een significante uitkomst van de Mantel-Haenszel-toets hadden niet meer in de somscore worden opgenomen. Nu is het enerzijds mogelijk dat er nieuwe items met significante DIF bijkomen, anderzijds is het mogelijk dat de significante DIF verdwijnt bij items die in de eerste analyse wel een significante uitkomst van de Mantel-Haenszel opleverden. Het iteratieve proces gaat door tot er een verzameling items zonder DIF gevonden wordt waarmee de somscore berekend kan worden en een verzameling items met een significante uitkomst van de Mantel-Haenszel-toets die niet in de berekening van de somscore zijn betrokken.

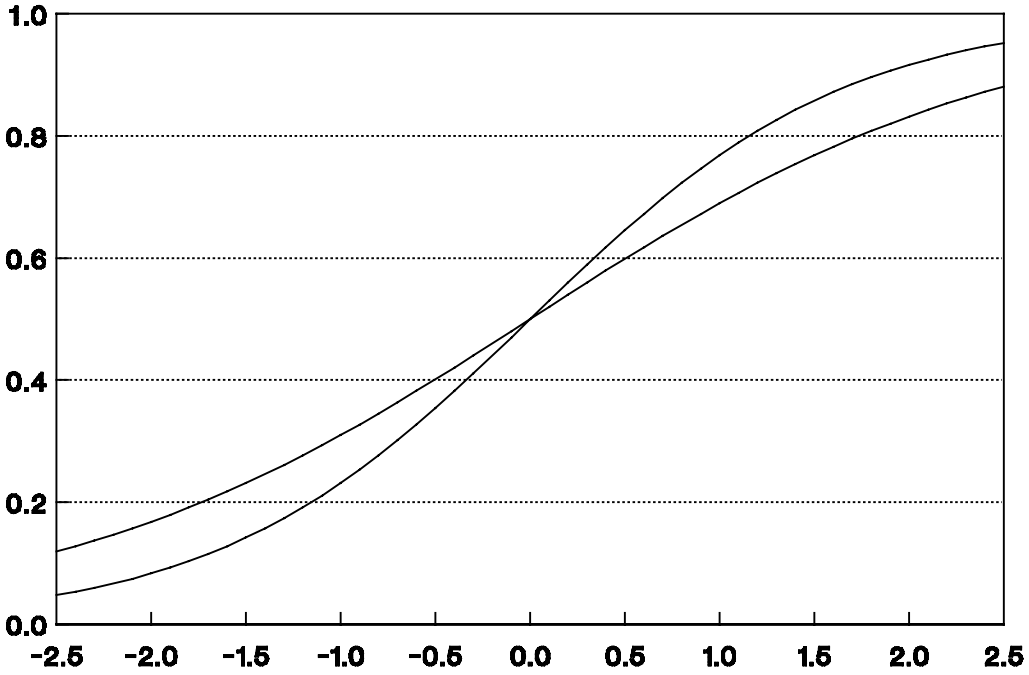
9.2.2 Procedure met IRT-modellen

In de itemresponstheorie wordt de kans op een goed antwoord op een item beschreven als een functie van persoonsparameters en itemparameters. Deze eigenschap maakt de klasse van IRT-modellen bijzonder geschikt voor het onderzoeken van DIF: conditioneren op het vaardigheidsniveau van respondenten is hier niets anders dan het constant houden van de persoonsparameters. Individuen met gelijke persoonsparameters moeten, ongeacht de populatie waartoe ze behoren, dezelfde kans op een goed antwoord hebben. Items kunnen verschillen in moeilijkheidsgraad en groepen kunnen verschillen in hun bekwaamheid om een juist antwoord op een item te geven, maar dat is op zich nog geen vraagonzuiverheid. Een item wordt alleen als onzuiver beschouwd als de moeilijkheidsgraad ervan varieert tussen personen van eenzelfde vaardigheidsniveau die tot verschillende populaties behoren. De generalisatie van DIF naar polytome items volgt eenvoudig uit de definitie voor dichotome items: een polytoom item is onzuiver als de verzameling van kansen om in één van de categorieën van het item te scoren, conditioneel op het vaardigheidsniveau, verschilt tussen groepen. Bij deze definities is niet van belang welk itemresponsmodel bij de data past. De term vaardigheidsniveau kan bijvoorbeeld betrekking hebben op een multidimensionale vaardigheidsparameter θ, zoals die voorkomt in het Raschmodel met een multivariate vaardigheidsverdeling dat behandeld is in hoofdstuk 5. Een unidimensionaal IRT-model maakt de problematiek conceptueel echter een stuk eenvoudiger.



Figuur 9.7
Responsfunctie van een uniform onzuiver dichotoom item

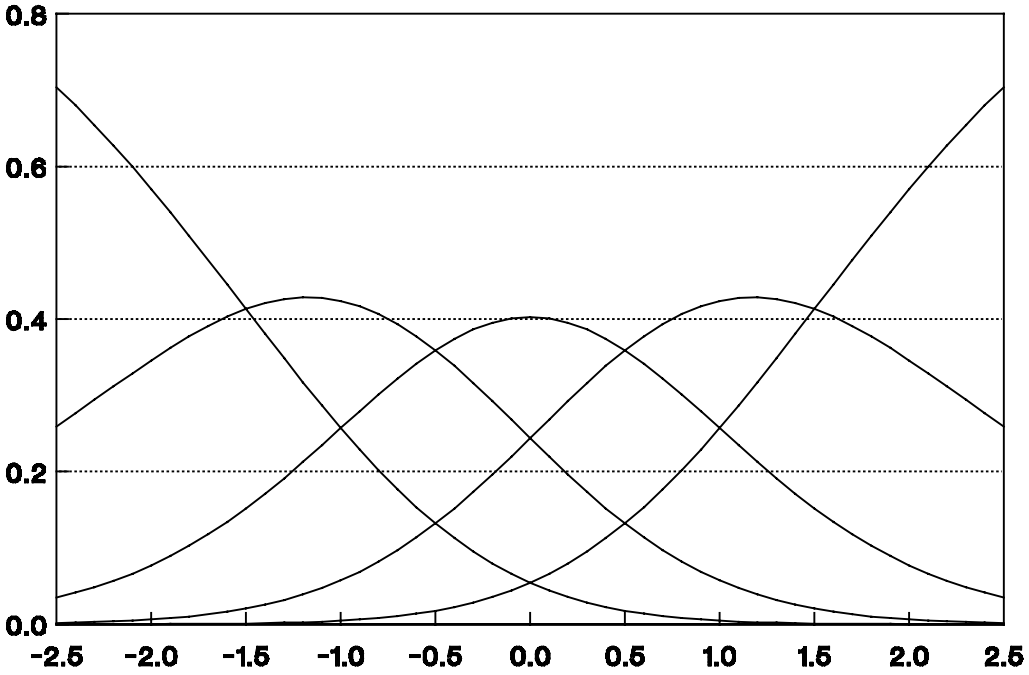
In paragraaf 9.1 is een onderscheid gemaakt tussen uniforme en niet-uniforme onzuiverheid. Een dichotoom item is uniform onzuiver als de kans op een juist antwoord in de doelpopulatie voor alle vaardigheidsniveaus lager is dan in de referentiepopulatie, of als het omgekeerde het geval is. Een voorbeeld van een dergelijk item wordt gegeven in figuur 9.7. Een item is niet-uniform onzuiver als de kans op een juist antwoord voor verschillende vaardigheidsniveaus nu eens in het voordeel is van de referentiepopulatie en dan weer in het voordeel is van de doelpopulatie. Een voorbeeld daarvan wordt gegeven in figuur 9.8.



Figuur 9.8
Responsfunctie van een niet-uniform onzuiver dichotoom item

In het voorbeeld doet de ene groep het op een laag vaardigheidsniveau beter dan de andere, terwijl dit op een hoog vaardigheidsniveau precies omgekeerd is. De systematische patronen van figuur 9.7 en 9.8 kunnen goed gemodelleerd worden door de locatie- en discriminatieparameters van het item te variëren over de groepen. In de praktijk kan het patroon van onzuiverheid veel onregelmatiger zijn en is het expliciet modelleren van de responsen van beide groepen niet altijd mogelijk.

De generalisatie van de concepten van uniforme- en niet-uniforme vraagonzuiverheid van dichotome naar polytome items is gecompliceerd omdat er in dat geval meer dan één itemresponsfunctie per item aanwezig is. In het voorbeeld van figuur 9.9 geeft de monotoon dalende curve links de kans op een score in de categorie nul aan, terwijl de monotoon stijgende curve rechts de kans op een score in de hoogste categorie aangeeft. De resterende eentoppige curven geven de kansen aan om in de overige categorieën te scoren. De itemresponscurven voldoen aan het partial credit model (PCM), maar aangezien slechts één item wordt beschouwd voldoen ze tevens aan het één-parameter logistische model (OPLM). In het PCM zijn de parameters β_{ij}, $j = 1,...,m_i$ de grenswaarden waar de kansen om in de categorie $j-1$ en de categorie j te scoren, gelijk zijn. Dat wil zeggen, de parameters geven de positie op de x-as aan waar de curven van categorie $j-1$ en j elkaar snijden.



Figuur 9.9

Itemresponsfunctie in het partial credit model

Het onderscheid tussen uniforme en niet-uniforme vraagonzuiverheid is intuïtief gezien bij dichotome items gerelateerd aan het al dan niet elkaar snijden van de itemkarakteristieke curven voor de verschillende populaties. In het geval van polytome items is een dergelijk eenvoudige definitie door het aantal karakteristieke curven en hun onderlinge afhankelijkheid niet mogelijk. Voor unidimensionale polytome modellen, zoals het PCM, het rating scale model of het OPLM kan men een item uniform onzuiver noemen wanneer de verwachte score op het item gegeven θ in de doelpopulatie systematisch hoger of lager is dan in de referentiepopulatie.

Onderzoek naar vraagonzuiverheid met behulp van IRT

Zoals hiervoor in termen van IRT is aangegeven, is een item onzuiver als de kansen op de responsen in de categorieën van het item, conditioneel op het vaardigheidsniveau, tussen groepen verschillen. De procedure voor het aantonen van dit verschijnsel bestaat uit twee stappen:

(1) het zoeken naar een passend IRT model voor de data van de referentiegroep en, voor zover mogelijk, de doelgroep,

(2) het evalueren van de verschillen in responskansen tussen de referentie- en de doelgroep in homogene subgroepen van gelijke vaardigheid.

Indien onzuivere items gevonden worden, kan men nog twee bijkomende stappen zetten: (3) het modelleren van de responsen van de doelpopulatie op de onzuivere items,

(4) het evalueren van de consequenties van de aanwezigheid van DIF, door het schatten van de resultaten (bijv. de scoreverdeling) van de doelpopulatie voor het geval geen DIF aanwezig zou zijn.

Met betrekking tot de eerste stap is allereerst de keuze van een itemresponsmodel van belang. Bij veel toetsen wordt de meting uitgevoerd door gebruik te maken van een ongewogen somscore. Dit betekent dat men de leerlingen ordent op een unidimensionaal vaardigheidscontinuüm en dat de persoonsparameter unidimensionaal is. Fischer (1974, pp. 193-203) heeft aangetoond dat onder de assumptie dat de somscore een voldoende steekproefgrootheid is voor een unidimensionale vaardigheidsparameter, en een paar technische assumpties (lokale stochastische onafhankelijkheid, een strikt monotoon stijgende kans op een goed antwoord die nergens gelijk aan nul of een is), het Raschmodel noodzakelijkerwijze volgt. Met andere woorden, het gebruik van de somscore als uitkomst van de met het toetsinstrument uitgevoerde meting impliceert dat de resultaten van de meting in feite aan het Raschmodel zouden moeten voldoen. Vaak voldoen de data echter niet aan het Raschmodel en moet men gebruik maken van andere modellen zoals het OPLM of een model met een multivariate vaardigheidsverdeling. Dit betekent dat de responskansen op de items conditioneel op de door deze modellen voorgeschreven steekproefgrootheden voor de vaardigheidsparameters moeten worden geëvalueerd. Met andere woorden, de rol van het IRT-model is het leveren van een adequate beschrijving van de vaardigheid van de leerlingen. In dit verband zullen we hier kort ingaan op een door Bügel en Glas (1991) gerapporteerd onderzoek naar vraagonzuiverheid bij examens tekstbegrip voortgezet onderwijs. Voor de eerste stap van het onderzoek, het zoeken naar een passend IRT-model voor de data van de referentiegroep en, voor zover mogelijk, de doelgroep, maakten zij gebruik van een variant van het model met een multivariate vaardigheidsverdeling dat beschreven is in hoofdstuk 5. Om zo dicht mogelijk bij de uiteindelijke resultaatbepaling van de examens te blijven, werd door de onderzoekers in de verzameling opgaven van het complete examen eerst gezocht naar een aantal Rasch-homogene subsets van items. Voor ieder van die subschalen is de somscore een voldoende grootheid voor de vaardigheidsparameter. In de examensituatie worden, voor de uiteindelijke resultaatbepaling, de somscores op de subschalen opgeteld tot een totaalscore als eindwaardering. Dit impliceert in feite een, meestal arbitraire, waardering voor de verschillende vaardigheidsdimensies: bij een andere combinatie van deelscores tot een

eindwaardering ontstaat namelijk een andere ordening van leerlingen. Overigens is de correlatie tussen de vaardigheidsdimensies hoog (altijd groter dan .85) zodat de afwijking ten opzichte van het Raschmodel niet bijzonder groot is en men zeker niet mag concluderen dat een examen een aantal scherp afgebakende vaardigheidsdimensies meet. Men zou de gevonden multidimensionaliteit eerder kunnen kenschetsen als additionele ruis bij een unidimensionaal Raschmodel. Het door Bügel en Glas gekozen IRT-model is niet per definitie het enig juiste. De essentie van de eerste stap is het zoeken van een passend IRT-model om een adequate maat voor de vaardigheid van de leerlingen te construeren. Zo zal voor het voorbeeld in dit hoofdstuk een andere keuze gemaakt worden, en zal gebruik worden gemaakt van het OPLM. Voor meer informatie over de procedure met het Raschmodel met een multivariate vaardigheidsverdeling zij men verder verwezen naar Bügel en Glas (1991).

De tweede stap van het onderzoek naar onzuiverheid is het evalueren van de verschillen in responskansen tussen de referentie- en doelgroep in subgroepen van gelijke vaardigheid. Hieronder zal worden beschreven hoe dit, in het kader van het OPLM, kan worden uitgevoerd. Hiertoe zullen twee toetsen voor het OPLM, de R_{1c}- en de S_i-toets, worden aangepast voor het opsporen van vraagonzuiverheid.

Om het zoeken van een passend IRT model niet te laten beïnvloeden door eventueel aanwezige onzuivere items, is het verstandig in eerste instantie alleen de gegevens van de referentiegroep te gebruiken. Voor het evalueren van de modelpassing kan men gebruik maken van de in de hoofdstukken 4 en 5 beschreven toetsen. Als een voor de referentiegroep passend model gevonden is, breidt men de analyse uit naar beide groepen. Stel dat groepslidmaatschap wordt aangeduid met het subscript g, waarbij de referentiegroep wordt geïndiceerd met $g = 1$ en de doelgroep met $g = 2$. Zoals bij de eerder geïntroduceerde versies van de R_{1c}- en S_i-toets (zie formule 5.44 en 5.45) worden homogene niveaugroepen, geïndexeerd met q, gevormd op basis van de voldoende statistieken s voor de persoonsparameters. Dus net als in de hoofdstukken 4 en 5 bestaat niveaugroep q uit alle leerlingen die een score s in een scorebereik G_q hebben. Beide toetsen zijn gebaseerd op het verschil tussen de proportie antwoorden in categorie j van item i in scoregroep s, $p_{ij|s}$ en de onder het model geschatte kans op een antwoord in categorie j van item i in scoregroep s, $\hat{\pi}_{ij|s}$. Voor het evalueren van vraagonzuiverheid worden deze proporties en kansen voor iedere groep g afzonderlijk uitgerekend, dus de toets zal nu gebaseerd zijn op proporties $p_{ij|sg}$ en geschatte kansen $\hat{\pi}_{ij|sg}$. De CML schattingen van de itemparameters worden berekend met behulp van de gegevens van zowel de referentie- als de doelgroep. Er wordt dus verondersteld dat voor beide groepen hetzelfde model geldt.

Om de relatie met de Mantel-Haenszel-procedure wat duidelijker te kunnen maken zullen we de veralgemening van de R_{1c}- en S_i-toets in termen van tellingen geven. Daartoe definiëren we de stochastische variabele $M_{ij|sg}$, met realisatie $m_{ij|sg}$, als het aantal antwoorden in categorie j van item i gegeven door personen van groep g en scoregroep s. De passing van het model voor beide groepen zal dus geëvalueerd worden met behulp van de verschillen tussen de geobserveerde en verwachte waarden van $M_{ij|sg}$. Deze verschillen zijn gegeven door

$$d_{ij|sg} = m_{ij|sg} - \mathcal{E}(M_{ij|sg} \mid \hat{\beta}) \tag{9.3}$$

waarbij $\mathcal{E}(M_{ij|sg} \mid \hat{\beta})$ de verwachte waarde is van $M_{ij|sg}$, uitgerekend met CML schattingen van de itemparameters β. Er geldt dat $m_{ij|sg} = n_{sg} p_{ij|sg}$ en $\mathcal{E}(M_{ij|sg} \mid \hat{\beta}) = n_{sg} \hat{\pi}_{ij|sg}$, met n_{sg} het aantal personen in groep g dat score s haalt. Naar analogie van (5.44) kan de globale modelpassing worden geëvalueerd met behulp van de asymptotisch chi-kwadraat verdeelde toetsingsgrootheid R_{1c}. Deze wordt benaderd door

$$R_{1c}^{*} = \sum_{g=1}^{2} \sum_{q=1}^{r} \sum_{i=1}^{k} \sum_{j=1}^{m_i} \frac{\left[\sum_{s \in G_q} d_{ij|sg}\right]^2}{\sum_{s \in G_q} var(d_{ij|sg})}, \tag{9.4}$$

waarbij $var(d_{ij|sg})$ de variantie van het verschil $d_{ij|sg}$ is.

Merk op dat in het geval van dichotome items het aggregatieniveau van de data waarop de verschillen $d_{ij|sg}$, met $j = 1$, gebaseerd zijn, hetzelfde is als bij de Mantel-Haenszel-toets. Met de verschillen $d_{ij|sg}$ gaat men na of de proportie goede antwoorden voor de referentie- en doelgroep conform de voorspellingen van het model zijn en, omdat voor beide groepen hetzelfde model geldt, of deze proporties gelijk zijn. Als de toetsingsgrootheid significant is, is door inspectie van de verschillen $d_{ij|sg}$ na te gaan of de verwerping toe te schrijven is aan systematische verschillen tussen de twee groepen in de kans op het produceren van een goed antwoord. Per item kan men de verschillen $d_{ij|sg}$ ook combineren tot een toetsingsgrootheid die is op te vatten als een veralgemening van de itemgerichte S_{ij}-toets. De benaderende toetsingsgrootheid gedefinieerd door (5.45) wordt daartoe veralgemeniseerd tot

$$S_{ij}^{*} = \sum_{g=1}^{2} \sum_{q=1}^{r} \frac{\left[\sum_{s \in G_q} d_{ij|sg}\right]^2}{\sum_{s \in G_q} var(d_{ij|sg})}, \qquad (j = 1,...,m_i). \tag{9.5}$$

Als is aangetoond dat één of meer items in een toets onzuiver zijn, is de derde stap in het onderzoek naar DIF mogelijk. Deze stap heeft betrekking op de vraag of het antwoordgedrag van de doelgroep adequaat kan worden beschreven door een itemresponsmodel. Inzicht in de aard van de onzuiverheid is uiteraard essentieel voor het voorkomen ervan. Zowel bij dichotome als bij polytome items kan het variëren van locatie- en discriminatieparameters van het item soms voldoende zijn om het antwoordgedrag van de verschillende populaties te modelleren. Een voorbeeld hiervan wordt in paragraaf 9.3 gegeven. Er zijn echter uiteraard ook vormen van DIF denkbaar waarbij de onzuiverheid complexer van aard is. Zo is het bijvoorbeeld mogelijk dat onzuiverheid ten nadele van de doelgroep alleen bij lage vaardigheidsniveaus voorkomt, en dat bij hogere vaardigheidsniveaus de doelgroep zijn achterstand volledig weet te compenseren.

De vierde mogelijke stap in het onderzoek naar vraagonzuiverheid is het evalueren van de invloed van de onzuiverheid op de verdeling van zowel de gewogen als de ongewogen somscores van de respondenten. Daarvoor moet eerst de vaardigheidsverdeling van de referentiegroep en de vaardigheidsverdeling van de doelgroep geschat worden. Hiertoe kan men bijvoorbeeld het OPLM uitbreiden met de veronderstelling dat de vaardigheidsparameters in beide groepen, overigens verschillende, normale verdelingen hebben. Vervolgens kan men de parameters in dit uitgebreide model met behulp van MML schatten. Het is echter ook mogelijk de CML schattingen van de itemparameters als constanten te beschouwen en alleen ML-schattingen van de populatieparameters te maken. In beide gevallen is het echter wel noodzakelijk dat de passing van het uitgebreide model aannemelijk wordt gemaakt. De effecten van de aanwezigheid van DIF zijn nu als volgt te evalueren.

Stel dat N_{sg} het aantal respondenten van groep g is dat een gewogen of ongewogen score s haalt. Gegeven nu de schattingen $\hat{\beta}$ van de itemparameters en $\hat{\mu}_g$ en $\hat{\sigma}_g$, voor $g = 1$ en 2, van de populatieparameters, kan men voor alle mogelijke scores s de verwachte waarde $\mathscr{E}(N_{sg} \mid \hat{\beta},\hat{\mu}_g,\hat{\sigma}_g)$ berekenen. Dit is overigens geen triviale aangelegenheid. Stel dat $\{\boldsymbol{x} \mid s\}$ de verzameling is van alle mogelijke antwoordpatronen $\boldsymbol{x}$ die resulteren in een score s. Dan berekent men deze verwachte waarden als

$$\mathscr{E}(N_{sg} \mid \hat{\beta},\hat{\mu}_g,\hat{\sigma}_g) \;=\; N_g \sum_{\{\boldsymbol{x} \mid s\}} P(\boldsymbol{x} \mid \hat{\beta},\hat{\mu}_g,\hat{\sigma}_g). \qquad (9.6)$$

Met andere woorden, men moet de kansen op antwoordpatronen sommeren over alle antwoordpatronen die resulteren in score s. Doordat ook hier echter symmetrische basisfuncties een rol blijken te spelen (zie Glas, 1991) is dit echter minder bewerkelijk

dan het lijkt. Het gaat er nu om, de resultaten van de doelpopulatie te schatten als de toets geen onzuivere items had gehad, met andere woorden, als de itemparameters voor beide groepen gelijk zouden zijn geweest. Daartoe kan men de verwachte frequentieverdeling van de doelpopulatie $\mathscr{E}(N_{sg} \mid \hat{\beta},\hat{\mu}_g,\hat{\sigma}_g)$ berekenen met voor de onzuivere items parameterwaarden die gevonden zijn bij de referentiepopulatie.

9.2.3 De relatie tussen de Mantel-Haenszel-procedure en de IRT-procedure

Een speciaal geval van de hierboven geschetste procedure met behulp van itemresponsmodellen is die welke gebaseerd is op het Raschmodel voor dichotome items. Zowel deze procedure als die met de Mantel-Haenszel-toets zijn allebei gebaseerd op hetzelfde principe, namelijk het toetsen of de kans op een goed antwoord gegeven een somscore of een bereik van somscores hetzelfde is voor de referentie- en de doelgroep. Beide technieken hebben voordelen en hun beperkingen.

Bij de Mantel-Haenszel-procedure is de somscore, in tegenstelling tot bij het Raschmodel, niet gevalideerd als maat voor de vaardigheid van de respondenten. Het gebruik van de ongewogen somscore is overigens niet essentieel voor de Mantel-Haenszelprocedure. Ook is het mogelijk om de niveaugroepen voor de toets op basis van een andere statistiek voor vaardigheid te vormen. Hierbij kan men bijvoorbeeld denken aan een gewogen somscore zoals bij OPLM gebruikt wordt. Ook hier blijft echter de kritiek dat deze maat voor het vaardigheidsniveau eerst gevalideerd zou moeten worden.

Een andere nadeel van de Mantel-Haenszel-procedure is dat niet alle vormen van onzuiverheid gedetecteerd kunnen worden. In het geval van uniforme onzuiverheid is de kans op een goed antwoord voor één van de groepen over het hele scorebereik systematisch hoger. In het geval van niet-uniforme onzuiverheid zijn er niveaus waarop de ene groep en niveaus waarop de andere groep beter scoort. De Mantel-Haenszelprocedure is alleen gevoelig voor de eerste vorm van onzuiverheid, in het tweede geval vallen de effecten in de toetsstatistiek tegen elkaar weg. De toetsingsgrootheden voor het Raschmodel en OPLM leiden niet aan dit euvel omdat hier de verschillen tussen verwachte en geobserveerde proporties gekwadrateerd worden.

Het toepassen van het Raschmodel of OPLM heeft echter als nadeel dat de parameterschatting leidt tot restricties op de toetsingsgrootheden, waardoor een item met DIF ten nadele van de ene groep kan resulteren in één of meer items die schijnbaar DIF vertonen ten nadele van de andere groep. Dit ongewenste effect ontstaat doordat de CML schattingsvergelijkingen voor de itemparameters te schrijven zijn als

$$\sum_g \sum_s m_{ij|sg} = \sum_g \sum_s \mathscr{E}(M_{ij|sg} \mid \beta), \tag{9.7}$$

zodat, na invulling van de schattingen geldt dat $\Sigma_{g,s}\ d_{ij|sg} = 0$. Met andere woorden, voor ieder item is de som over groepen respondenten van de verschillen tussen verwachte en geobserveerde frequenties nul. Dit betekent dat door de schattingsmethode, vraagonzuiverheid die de ene groep benadeelt altijd samengaat met een bevoordeling van de andere groep. Restrictie (9.7) geldt voor ieder item afzonderlijk. Er ontstaan door de schattingsmethode echter ook afhankelijkheden die betrekking hebben op alle items. Na CML schatting geldt namelijk ook dat $\Sigma_i\ d_{ij|sg} = 0$, met $j = 1$. Dus voor iedere groep respondenten is de som over items van de verschillen tussen verwachte en geobserveerde frequenties ook nul. Voor iedere groep respondenten wordt de aanwezigheid van benadelende items hierdoor vertaald in de aanwezigheid van bevoordelende items, vice versa.

Gezien deze overwegingen is het raadzaam de beide technieken zo veel mogelijk in elkaars verlengde te hanteren. Zo kan men bijvoorbeeld eerst Rasch-homogene subschalen of een passend OPLM zoeken en op de aanwezigheid van DIF toetsen met het IRT model, om vervolgens voor iedere subschaal de Mantel-Haenszel-techniek toe te passen. Door deze vorm van kruisvalidatie kan men artefacten die samenhangen met de gebruikte methode zoveel mogelijk vermijden.

9.2.4 Een voorbeeld van het bepalen van vraagonzuiverheid met behulp van OPLM

Het voorbeeld dat gegeven zal worden betreft een deel van het eindexamen HAVO voor het vak economie. Dit voorbeeld vormde een onderdeel van een groter onderzoek naar geslachtsgebonden vraagonzuiverheid bij de eindexamens in het voortgezet onderwijs. Aangezien het hier de bedoeling is om statistische procedures te illustreren en niet om inhoudelijk op de uitkomsten van het onderzoek naar vraagonzuiverheid in te gaan, zullen geen voorbeelden van onzuivere items getoond worden of conclusies worden getrokken over de mate waarin het verschijnsel voorkomt.

De analyses werden uitgevoerd op een steekproef van 1000 jongens en 1000 meisjes uit de totale examenpopulatie. Voor de eenvoud van de presentatie zal het voorbeeld tot tien polytoom gescoorde items beperkt worden.

De eerste stap van de procedure bestond uit het zoeken van een passend OPLM. Dit gebeurde door een iteratieve procedure van het postuleren van discriminatie-indices, het berekenen van CML schattingen, het toetsen en bijstellen van de hypothesen met betrekking tot de discriminatie-indices. Om het zoeken naar een geschikt model niet

te laten beïnvloeden door mogelijk aanwezige DIF, zijn eerst alleen de data van de referentiegroep gebruikt. De analyses werden uitgevoerd met het computerprogramma OPLM (Verhelst, Glas & Verstralen, 1993). In tabel 9.1 wordt een overzicht gegeven van de uitkomsten van de toetsen voor het definitieve model. In de kolom "A" worden de discriminatie-indices weergegeven.

Tabel 9.1

Overzicht van passingstoetsen voor de referentiegroep

Item	A		S	df	P	M	M2	M3
1	2	[:1]	11.724	7	.110	-.294	-.648	-.039
		[:2]	6.685	7	.462	-.460	.098	-.584
2	3	[:1]	5.918	6	.432	-1.390	.716	.587
		[:2]	6.346	7	.500	-.195	.554	.029
		[:3]	4.025	5	.546	.003	.512	.878
3	4	[:1]	9.685	5	.085	1.543	2.476	3.615
		[:2]	1.624	6	.951	.893	.750	.167
4	2	[:1]	4.054	7	.774	.578	.423	.163
		[:2]	10.543	7	.160	.238	-.309	-1.202
		[:3]	3.582	5	.611	.472	.010	-.634
5	2	[:1]	9.124	6	.167	1.408	1.601	1.888
		[:2]	2.208	7	.947	.284	.837	-.631
		[:3]	5.140	7	.643	-1.064	.494	-.928
6	3	[:1]	6.090	7	.529	.743	.761	.006
		[:2]	4.065	7	.772	.315	.836	.414
7	3	[:1]	5.873	7	.555	-.063	-.961	.286
		[:2]	15.456	6	.017	.528	-.645	1.892
8	3	[:1]	6.971	5	.223	-.687	-.361	-1.348
		[:2]	15.915	6	.014	-1.473	-.427	-2.709
		[:3]	6.283	6	.392	.010	-.002	-.141
9	4	[:1]	6.359	6	.384	.120	-.930	-.779
		[:2]	1.958	6	.923	-1.202	-.913	-.386
10	4	[:1]	2.321	4	.677	-.187	-1.186	-.158
		[:2]	2.575	5	.765	-1.126	-.794	-1.339
		[:3]	5.503	5	.358	-.653	-1.213	.532

R_{1c} = 75.182; df = 72; p = .3757

De splitsing van het scorebereik van een item in de scores $0,...,j$ en $j+1,...,m_i$ kan in verkorte notatie worden weergegeven als $[:j+1]$, voor $j = 0,...,m_i-1$. Het programma OPLM berekent de S_{ij}- en M-toetsen voor alle dichotomisaties [:1],...,$[:m_i]$. In de

kolom "S" worden de waarden van de S_{ij}-toetsen weergegeven, de volgende twee kolommen geven respectievelijk het aantal vrijheidsgraden en de overschrijdingskansen. In de laatste drie kolommen worden de waarden van de drie versies van de M-toets gegeven, deze toetsen zijn asymptotisch normaal verdeeld. Aan de hand van de waarde van de R_{1c}-toets die onderaan de tabel staat afgedrukt, kan men zien dat de passing van het model aanvaardbaar is. In de daarop volgende twee analyses werden de discriminatie-indices die voor de referentiegroep waren gevonden niet veranderd. In de eerste analyse werden CML schattingen berekend en modeltoetsingen uitgevoerd op de doelpopulatie. In de tweede analyse werden CML parameterschattingen en modelpassing berekend op beide groepen tegelijk. De resultaten van de daarbij behorende R_{1c}-toetsen staan vermeld in tabel 9.2 in de rijen genummerd twee en drie. Het blijkt dat het model in beide gevallen verworpen moest worden. De resultaten van de tweede analyse laten zien dat de discriminatie-indices van de referentiepopulatie niet passen in de doelpopulatie, zelfs wanneer de schattingen van de itemparameters in deze laatste groep verkregen zijn.

Tabel 9.2
Hypothesetoetsing

analyse	model	R_{1c}	df	prob
1.	referentiegroep	75.182	72	.3757
2.	doelgroep	127.283	72	.0001
3.	gecombineerde groepen	356.747	168	.0000
4.	doelgroep, 9 aangepaste index	59.982	72	.8430
5.	gecombineerde groepen, 3 gesplitst	258.614	166	.0000
6.	gecombineerde groepen, 9 gesplitst	379.550	166	.0000
7.	gecombineerde groepen, 3 en 9 gesplitst . .	154.301	164	.6971

De resultaten van de derde analyse geven ook aan dat de gecombineerde data van beide groepen tegelijk, niet goed door hetzelfde model beschreven kunnen worden. Om na te gaan of dit laatste resultaat een gevolg is van DIF wordt in tabel 9.3 een overzicht gegeven van de passingstoetsen voor beide groepen samen. De tabel heeft hetzelfde formaat als tabel 9.1. Het blijkt dat de items drie en negen in belangrijke mate bijdragen aan het niet passen van het model. Onderaan de tabel staat de bijdrage van de twee groepen aan de uitkomst van de R_{1c}-toets. De bijdrage van de doelgroep (een χ^2 van 212.64) is veel groter dan de bijdrage van de referentiegroep (een χ^2 van 144.11).

Gezien het feit dat de discriminatie-indices bepaald zijn op de referentiegroep is dit niet verwonderlijk.

Om de hypothese van DIF verder te onderzoeken, kunnen bijvoorbeeld de verschillen tussen geobserveerde en verwachte frequenties behorend bij de R_{1c}-toets geïnspecteerd worden. Voor het berekenen van deze toets zijn de respondenten van zowel de referentie- als van de doelgroep, op basis van hun gewogen somscores, opgedeeld in vier subgroepen. Deze subgroepen werden zodanig samengesteld dat ze ongeveer hetzelfde aantal respondenten bevatten. De gekozen scoreniveaus en de resulterende aantallen respondenten per subgroep staan vermeld in de eerste twee regels van tabel 9.4. Verder worden voor alle items en alle categorieën de gestandaardiseerde afwijkingen tussen de verwachte en de geobserveerde frequenties in de subgroepen getoond. Voor de interpretatie van deze getallen is het belangrijk in gedachte te houden dat het realisaties van bij benadering standaard normaal verdeelde variabelen zijn.

Tabel 9.3

Overzicht van passingstoetsen voor de doel- en referentiegroep samen

Item	A		S	df	P	M	M2	M3
1	2	[:1]	28.189	14	.013	-.864	-.791	-1.121
		[:2]	12.748	14	.546	.067	.517	-1.236
2	3	[:1]	7.399	11	.766	-.070	.183	1.079
		[:2]	13.011	14	.526	-.838	-.625	-1.210
		[:3]	4.268	10	.934	.795	.755	.024
3	4	[:1]	107.862	12	.000	2.315	.658	2.771
		[:2]	37.500	12	.000	-1.438	.548	-1.787
4	2	[:1]	8.121	14	.883	-.721	-.351	-.338
		[:2]	15.971	14	.315	-.131	-.475	-.610
		[:3]	15.665	10	.110	-.137	-1.084	-1.317
5	2	[:1]	11.393	12	.496	1.428	-.339	.395
		[:2]	15.399	14	.351	-1.318	-1.453	-1.701
		[:3]	10.520	14	.723	-1.997	-1.455	-1.384
6	3	[:1]	10.486	14	.726	.358	1.505	.543
		[:2]	11.375	14	.656	.442	1.264	.518
7	3	[:1]	18.279	14	.194	-.438	-1.395	-.066
		[:2]	18.005	12	.116	1.376	-1.221	1.179
8	3	[:1]	9.410	10	.494	-1.049	-.234	-.955
		[:2]	19.127	13	.119	-1.341	-.615	-1.566
		[:3]	8.080	12	.779	-.322	.173	-.609
9	4	[:1]	113.760	12	.000	4.025	4.297	4.614
		[:2]	35.874	12	.000	2.657	2.655	3.173
10	4	[:1]	14.893	9	.094	-1.120	-1.083	-1.070
		[:2]	16.264	10	.092	-1.642	-1.612	-2.343
		[:3]	24.262	11	.012	-2.164	-2.123	-.712

groep	#items	#subgr.	#deviaties	R_{1c}

1	10	4	96	144.11
2	10	4	96	212.64

R_{1c} = 356.747; df = 168; p = .0000

Aan het teken kan men zien of er meer of minder observaties waren dan voorspeld door het model. In de kolommen "SS" worden de kwadratensommen van de afwijkingen vermeld, voor alle combinaties van items en categorieën. Merk op dat met name de kwadratensommen van item 3 groot zijn vergeleken met de kwadratensommen van de andere items. Verder vallen de geschaalde afwijkingen voor de referentiegroep over het algemeen positief uit, terwijl de afwijkingen bij dit item voor de doelgroep negatief zijn.

Tabel 9.4

Geschaalde afwijkingen op grond van CML schattingen, verkregen in beide groepen tegelijk

		referentiegroep					doelgroep				
		1	2	3	4		1	2	3	4	
Range →		1-20	21-37	38-52	53-73		1-20	21-36	37-52	53-73	
#obs →		228	237	253	263	SS	240	246	239	252	SS
Item	cat										
1	1	-.1	-1.5	.6	-.6	3.2	1.3	-.9	.3	1.1	4.0
	2	-1.5	.4	-.9	-.1	3.5	.5	.2	.7	.3	1.0
2	1	-.9	-.1	.8	.4	1.7	1.9	-.2	-.2	-2.0	7.8
	2	-.4	-.5	-1.2	-.0	2.1	-.2	-.0	.8	1.3	2.5
	3	-1.4	.9	.4	-.8	3.8	-.1	.3	-.0	.2	.1
3	1	5.1	2.8	.0	-1.0	35.7	-2.9	-3.4	.5	-1.0	21.6
	2	.9	2.4	1.7	.4	10.4	-1.8	-2.0	-3.2	1.0	18.9
4	1	1.9	-.0	.5	.0	4.1	-.4	-.6	-1.7	.6	4.0
	2	-2.0	-.8	.1	-.3	4.8	-.2	2.2	1.8	-1.3	10.3
	3	-.8	-.2	-1.0	-.0	1.8	-.8	-.2	-.8	2.0	5.6
5	1	1.8	.4	1.0	-1.1	6.0	-.8	-.1	-.8	-1.4	3.4
	2	.0	-.1	-.8	-.0	.7	-.6	.8	-.1	1.1	2.4
	3	-.8	-1.1	-.2	.5	2.4	.7	-.8	1.4	.5	3.4
6	1	-.3	.6	.4	.7	1.2	-.4	.2	-.2	-1.3	2.1
	2	-.4	-.4	-1.3	-.3	2.3	.4	1.7	-.1	.7	3.7
7	1	-.7	.0	-1.0	1.1	2.8	-.9	.6	1.0	-.5	2.6
	2	1.5	-2.8	.6	-1.1	11.8	-.2	.8	.8	.6	1.9
8	1	1.8	-.5	.5	-1.1	5.2	.2	-.9	-1.0	.6	2.4
	2	-2.3	1.0	.4	-.0	6.6	-.1	.2	.6	-.7	1.0
	3	-.7	.5	-1.3	.7	3.2	1.2	-.4	.0	.2	1.7
9	1	1.1	.4	.6	-1.0	3.0	1.4	-.2	-1.9	.1	6.0
	2	.7	.4	1.0	1.7	4.8	3.3	.7	-1.0	-4.2	30.6
10	1	.4	1.1	.6	-.6	2.4	-.2	-.4	-1.6	.0	3.0
	2	-1.3	-.0	.2	.8	2.6	-1.5	1.9	-.4	-1.5	8.6
	3	-1.0	-2.0	-.5	-.4	5.8	.6	-.0	1.9	1.5	6.6
SS →		63.0	36.8	18.3	15.0	133.1	39.2	34.4	36.8	45.8	156.3

$R_{1c}\rightarrow$	58.5	45.2	20.4	19.9	144.1	39.2	45.3	43.8	84.2	212.6

Dat wil zeggen dat dit item de referentiegroep bevoordeelt, aangezien deze groep meer responsen in de categorieën $h>0$ vertoont dan op grond van een in beide groepen samen gecalibreerd model verwacht zou kunnen worden. Op dezelfde wijze is het item nadelig voor de doelgroep, aangezien deze groep minder responsen in de categorieën $j > 0$ vertoont, en dus meer responsen in categorie $j = 0$. Voor item 9 is het patroon veel minder duidelijk.

Op grond van de analyse die in tabel 9.2 met een 2 genummerd is, zou verwacht kunnen worden dat de discriminatie-index voor item 9 in beide groepen verschillend zou zijn. Daartoe werd de analyse uitgevoerd die in tabel 9.2 met een 4 genummerd is. Voor deze analyse, waarbij alleen de gegevens van de doelgroep gebruikt weren, werd de discriminatie-index voor dit item van 4 in 2 veranderd. In tabel 9.2 is te zien dat deze aanpassing inderdaad resulteerde in een goede modelpassing: de uitkomst van R_{1c} is 59.982 bij 72 vrijheidsgraden.

In de laatste drie analyses waarvan de resultaten van de hypothesetoetsing in tabel 9.2 vermeld staan, is getracht om een model te construeren wat voor de data van beide groepen tegelijk zou passen. In analyse 5 is toegelaten dat de parameters van item 3 voor de referentie- en de doelgroep verschillend zouden kunnen zijn, waarbij de discriminatieparameter constant is gehouden. Dit resulteerde echter niet in een acceptabele modelpassing. In analyse 6 werd dezelfde procedure toegepast voor item 9, met dit verschil dat de discriminatie-index in de referentiegroep op vier werd gezet en in de doelgroep op twee. Opnieuw waren de resultaten onbevredigend. Tenslotte werd in analyse 7 voor beide items toegelaten dat de moeilijkheidsparameter tussen de groepen zouden kunnen verschillen en dit bleek, zoals te zien in de laatste regel van tabel 9.2, in een acceptabele modelpassing te resulteren. Resumerend kan men stellen dat item 3 uniform onzuiver is, omdat de itemparameters per groep verschillen, terwijl de discriminatie per groep gelijk is, terwijl item 9 niet-uniform onzuiver is, omdat ook de discriminatie-index aangepast moest worden. Overigens werden item 3 en 9 ook in de Mantel-Haenszel-procedure als onzuiver geïdentificeerd. Hiermee is de derde stap in het onderzoek, het modelleren van de responsen van de doelpopulatie afgesloten.

Tot slot werd de vierde stap van het onderzoek naar vraagonzuiverheid gezet door het evalueren van de invloed van de onzuiverheid op de verdeling van zowel de gewogen als de ongewogen somscores van de respondenten. Als eerste stap werd daartoe de passing van het model uit analyse 7, uitgebreid met normale vaardigheidsverdelingen voor de referentie- en doelgroep, onderzocht. De

itemparameters β en populatieparameters μ_g en σ_g, voor $g = 1$ en 2, werden geschat met behulp van MML. Berekening van de R_0-toets (zie hoofdstuk 4) resulteerde in een waarde van 121.79 (df: 138, p: .83), terwijl het berekenen van R_{1m} een waarde 267.82 opleverde (df: 303, p: .92), zodat dit uitgebreide model niet verworpen hoefde te worden. Hierna werd voor de doelpopulatie de frequentieverdeling $\mathscr{E}(N_{sg} \mid \hat{\beta}, \hat{\mu}_g, \hat{\sigma}_g)$ berekend met de parameters van de items 3 en 9 gelijk aan de waarden die gevonden werden bij de referentiepopulatie en de schattingen van de populatieparameters van de doelpopulatie. Op deze wijze worden de resultaten van de doelpopulatie op een zuivere toets geschat, dat wil zeggen, de resultaten voor het geval de itemparameters voor de referentie- en doelpopulatie gelijk zouden zijn geweest. Deze geschatte frequentieverdeling op een zuivere toets kan men dan vervolgens vergelijken met de gerealiseerde frequentieverdeling. Voor het bovenstaande voorbeeld werden de berekeningen uitgevoerd voor zowel de gewogen als de ongewogen scores. In beide gevallen bleek het gemiddelde van de verwachte frequentieverdeling voor de doelpopulatie lager voor de onzuivere test. Het verschil bedroeg overigens in beide gevallen minder dan één scorepunt. Met andere woorden de onzuiverheid had inderdaad een bescheiden negatieve invloed op het gemiddelde resultaat van de doelpopulatie.

9.3 Conclusie

Itemresponstheorie biedt een goed gefundeerd kader voor het opsporen van vraagonzuiverheid. Hierbij is het echter belangrijk dat de hulpmiddelen die de IRT ons aanreikt ook zorgvuldig worden gebruikt. In de eerste plaats dient een passend IRT-model te worden gevonden. Hierbij spelen twee aspecten een rol: de data en de mate waarin de passing van de verschillende IRT-modellen statistisch goed gefundeerd te evalueren zijn. Het OPLM beschikt enerzijds over een goed uitgerust toetsingsarsenaal en blijkt anderzijds in veel gevallen goed bij de data te passen. Daar komt bij dat de statistische toetsen voor dit model zo zijn te generaliseren, dat ze gevoelig zijn voor vraagonzuiverheid. Door parameterschatting en andere oorzaken kan de informatie die de toetsen opleveren enigszins vertroebelen. Daarom is het aan te bevelen de resultaten te kruisvalideren door het uitvoeren van een Mantel-Haenszel-procedure, waarbij de niveaugroepen gevormd worden op basis van de afdoende statistieken van het passende IRT-model. Tenslotte is een niet onaantrekkelijk aspect van het werken met een IRT-model dat men het niet hoeft te laten bij het opsporen van vraagonzuiverheid, maar dat men ook de effecten hiervan op de toetsresultaten kan schatten.

10

Het meten van veranderingen

In het onderwijs kan een groeiende belangstelling bespeurd worden voor systemen die de vorderingen van individuele leerlingen kunnen meten. Zulke systemen noemt men leerlingvolgsystemen (LVS). Daarbij gaat het om de volgende vragen. Hoeveel beter kan een leerling technisch lezen na drie maanden onderwijs? In welke mate is de leerling het afgelopen half jaar vooruitgegaan in rekenen? Deze vragen refereren aan veranderingen in individuele vaardigheidsniveaus. We proberen dan individuele groei, op basis van meetresultaten op verschillende tijdstippen, te kwantificeren. In het verleden was de gangbare praktijk groei te meten met veranderingsscores, het verschil tussen twee meetresultaten, meestal binnen het kader van de klassieke testtheorie. Het meten van groei met veranderingsscores was echter geen succes. Vandaar dat wij in dit hoofdstuk een meer modelmatige benadering kiezen, veranderingsscores blijven buiten beschouwing.

We gaan na wat de meetmodellen die in de hoofdstukken 3 en 4 zijn besproken, de klassieke testtheorie en de itemresponstheorie te bieden hebben voor het volgen van individuele vaardigheden. In principe zijn deze meetmodellen statisch, dat wil zeggen: ontworpen voor metingen op één bepaald tijdstip. Een meetmodel beschrijft de relatie tussen het meetresultaat en de te meten vaardigheid op één tijdstip, bijvoorbeeld de relatie tussen observatie en ware score (klassieke testtheorie) of latente vaardigheid (itemresponstheorie). Bij het meten van veranderingen beschikken we over meetresultaten van hetzelfde individu op verschillende tijdstippen. Toepassing van een statisch meetmodel op de meetresultaten resulteert dan in een aantal momentopnamen van de te meten vaardigheid, zonder er rekening mee te houden dat de metingen betrekking hebben op hetzelfde individu. Modellen die metingen aan hetzelfde individu op meer dan een tijdstip beschrijven, worden aangeduid als dynamische of tijdsafhankelijke modellen. Dynamische modellen onderscheiden zich van statische modellen door expliciet de relatie te leggen tussen metingen op verschillende tijdstippen.

In dit hoofdstuk ligt de nadruk op modellen die de vorderingen in leerresultaten van individuele leerlingen kunnen beschrijven of voorspellen. In de eerste paragraaf wordt

de problematiek van het meten van veranderingen in het algemeen besproken. De bepaling van individuele vorderingen wordt, met als uitgangspunt een simpel lineair groeimodel, in de tweede paragraaf uitgewerkt, waarbij als meetmodel de klassieke testtheorie wordt gehanteerd. Hetzelfde doen we in de derde paragraaf, maar nu met een itemresponsmodel als meetmodel. Het accent in de paragrafen 10.2 en 10.3 ligt op de vergelijking van een statische en een dynamische aanpak bij de modellering en de consequenties daarvan voor de bepaling van individuele vorderingen. Tenslotte wordt in de laatste paragraaf de problematiek van het meten van veranderingen in een breder perspectief geplaatst en wordt nader ingegaan op alternatieve benaderingen en verwachtingen over mogelijke ontwikkelingen.

10.1 Individuele groei

De problematiek van het meten van veranderingen, het volgen van leerresultaten, of meer algemeen het vaststellen van groei, is geen sinecure. In het verleden zijn sommige auteurs (Cronbach & Furby, 1970) zo pessimistisch geworden dat zij hebben voorgesteld de hele kwestie van veranderingsscores maar te vergeten en de onderzoeksvragen zo te formuleren dat er geen veranderingsscores aan te pas komen (zie ook Jansen, 1979). Uit het aantal verwijzingen naar het werk van Cronbach en Furby in recenter literatuur blijkt echter dat door de jaren heen de kwestie van het meten van veranderingen de wetenschap is blijven boeien.

In deze paragraaf onderzoeken we waar de problemen zitten bij het meten van verande- ringen. Eerst kijken we naar de relatie tussen model en data in een longitudinaal onderzoek. Daarna worden aan de hand van een concreet voorbeeld enkele problemen bij het meten van veranderingen geïllustreerd. De paragraaf wordt besloten met een korte verhandeling over de methodologische aspecten bij het meten van veranderingen, maar dan specifieker gericht op het volgen van individuele leerresultaten.

10.1.1 Longitudinale data en modellering

Als over een longitudinale gegevensverzameling wordt gesproken, wordt daarmee bedoeld dat men beschikt over meetresultaten van hetzelfde object met betrekking tot een bepaald attribuut op verschillende tijdstippen. In het onderwijs resulteert dit

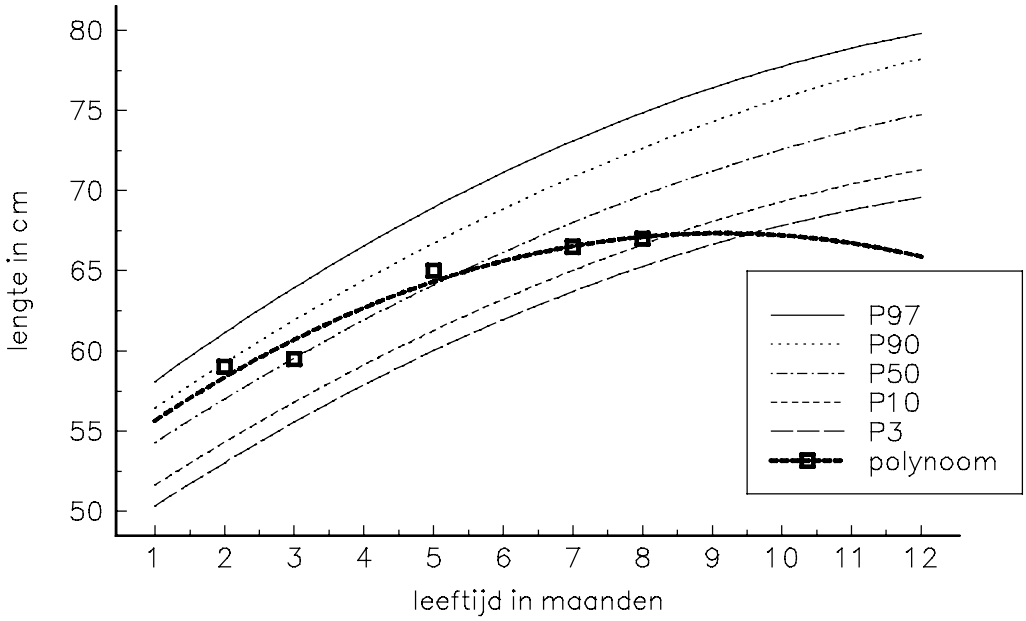
meestal in een gegevensver- zameling die betrekking heeft op de interactie tussen toetsen en individuen op verschillende tijdstippen.

Als men beschikt over een longitudinale dataset, is dat geen garantie dat die gegevens daad- werkelijk dynamisch gemodelleerd worden, dat wil zeggen dat de interactie tussen toetsen, individuen en tijdstippen simultaan wordt beschouwd. De gangbare praktijk is om longi- tudinale meetresultaten te reduceren tot statische momentopnamen op afzonderlijke tijdstippen. Eigenlijk beschouwen we dan de afzonderlijke meetmomenten los van elkaar. De afzonderlijke meetmomenten in de longitudinale gegevensverzameling duidt men aan met de term cross-secties. Wordt er al gebruik gemaakt van een dynamisch model, dan heeft dit vaak alleen betrekking op geaggregeerde gegevens op populatieniveau. Een voorbeeld van zo'n gangbare praktijk is het verticaal equivaleren (zie hoofdstuk 8). Binnen de itemresponstheorie is het mogelijk, onder zekere condities, een longitudinale dataset met een statisch meetmodel te analyseren. Feitelijk wordt de longitudinale gegevensverzameling opgedeeld in afzonderlijke cross-secties (individuen × toetsen) welke vervolgens worden gecombi- neerd in een onvolledig design tot één dataverzameling, die traditioneel met een statisch meetmodel geanalyseerd kan worden (zie bijvoorbeeld figuur 8.8 in hoofdstuk 8). Deze benadering is prima zolang zij schattingen van itemparameters en dergelijke betreft en we ons realiseren dat deze schattingen betrekking hebben op de onderhavige populatie. Bovendien geldt dat bij het analyseren van cross-secties van de data met een statisch model een mogelijke samenhang tussen de individuele meetresultaten in de tijd niet naar voren gehaald en belicht wordt. Veel van de door Cronbach en Furby (1970) gesignaleerde problemen bij het meten van veranderingen zijn dan ook artefacten van de gekozen benadering. Concluderend kan gezegd worden dat longitudinale gegevens in principe om een dynamisch model vragen.

10.1.2 Het vaststellen van de individuele groei bij zuigelingen

Op het consultatiebureau van de Kruisvereniging houdt men periodiek, naast andere zaken als gewicht en hoofdomtrek, de lichaamslengte van zuigelingen bij. Het doel hiervan is het tijdig signaleren van stagnaties in de groei zodat, indien gewenst, passende maatregelen genomen kunnen worden. De vraag rijst dan uiteraard wanneer er actie ondernomen dient te worden. We zullen hier niet de medische aspecten doch de methodologische aspecten beschouwen. De zuigeling wordt bij alle gelegenheden gemeten met dezelfde schuifmaat met een schaal in centimeters. Laten we aannemen dat bij de metingen de meetfout verwaarloosbaar is. Het is duidelijk dat bij alle

gelegenheden een en het zelfde attribuut, lichaamslengte in cm, bij de zuigeling gemeten wordt. In figuur 10.1 hebben we voor een hypothetische zuigeling de gemeten lichaamslengte uitgezet tegen de leeftijd in maanden. De open vierkantjes zijn de waarnemingen bij de leeftijden: 2, 3, 5, 7 en 8 maanden. De mate van groei kunnen we direct aflezen als het verschil tussen twee metingen. Na drie maanden meet de zuigeling 59.5 cm en na vijf maanden 65 cm: in twee maanden tijd is de zuigeling 5.5 cm gegroeid. Zou de medicus over absolute criteria beschikken, bijvoorbeeld dat na zeven maanden elke zuigeling 60 cm moet zijn, dan is het mogelijk op grond hiervan te beslissen of voor een specifieke zuigeling hulp nodig is. Aangezien absolute criteria meestal niet voorhanden zijn, gebruikt men relatieve. Men zou bijvoorbeeld de populatie zuigelingen in Nederland kunnen beschouwen en met behulp van een steekproef kunnen vaststellen hoe de ontwikkeling in de populatie van zuigelingen er uit ziet. De ontwikkeling in de populatie kan men dan per tijdstip met referentiegegevens beschrijven, bijvoorbeeld door per tijdstip decielen of percentielen (zie paragraaf 13.4.1) te bepalen. Het signaleren van stagnatie in de groei kan dan relatief plaatsvinden, een afwijking van twee of meer decielen naar beneden zou men als ongewenst kunnen bestempelen. In figuur 10.1 zijn als referentiegegevens vijf percentiellijnen getrokken. De percentiellijn P50 bijvoorbeeld geeft aan waar het vijftigste percentiel voor een bepaalde leeftijd ligt. Met behulp van deze lijnen is het mogelijk de relatieve positie van de zuigeling aan te geven. In het voorbeeld bevindt de zuigeling zich na vijf maanden tussen de P50 en P90, na zeven maanden tussen de P10 en P50.



Figuur 10.1
Groeicurve voor een hypothetische zuigeling met referentiegegevens

De positie van de zuigeling in de Nederlandse populatie van zuigelingen is dus veranderd. Immers, na vijf maanden behoorde de zuigeling tot de 'groten', terwijl na zeven maanden de zuigeling bij de 'kleintjes' gerekend mag worden. Of deze ontwikkeling ongewenst is, is een medische vraag. Verder, maar meer discutabel op grond van het geringe aantal waarnemingen, is het mogelijk de groei van de zuigeling op de een of andere manier te modelleren. De meetpunten in figuur 10.1 zijn benaderd met een polynoom. Deze is zichtbaar als de dikke lijn. Het is nu mogelijk met behulp van dit polynoom, dat we kunnen opvatten als een groeimodel, predicties te doen. Op grond van dit simpele groeimodel is de verwachting dat de lichaamslengte van de onderhavige zuigeling na tien maanden ongeveer 67.5 centimeter is. Met behulp van predicties is het mogelijk reeds vooraf iets te signaleren: gegeven de curve tot nu verwachten we dat na tien maanden de zuigeling in de gevarenzone komt.

Er blijven nog genoeg vragen over. Bijvoorbeeld: is de Nederlandse populatie wel geschikt als referentiepunt? Denkbaar is dat een opdeling van de populatie naar geslacht of gewichts- klasse zeer zinvol zou kunnen zijn. Met andere woorden, niet één maar verschillende popu- laties worden beschouwd. Een complicerende factor in het voorbeeld is het feit dat groei bij de individuele zuigeling niet vloeiend, maar schoksgewijs verloopt. Voorstelbaar is dus dat ogenschijnlijke stagnatie, door het slecht kiezen van tijdstippen, ten onrechte tot de conclusie leidt, dat hulp geboden is. Iets

dergelijks zou men kunnen observeren in het voorbeeld: de lengte na twee en drie maanden is nagenoeg gelijk, terwijl we na vijf maanden een aanzienlijke groei zien.

Dit voorbeeld illustreert dat het vaststellen van (stagnaties in de) groei bij zuigelingen, ook al beschikken we over metingen met te verwaarlozen meetfouten, niet geheel vrij van proble- men is.

10.1.3 Problemen bij het volgen van individuele leerlingen

Waar gaat het nu precies om bij het volgen van de vaardigheid van individuele leerlingen? In eerste instantie proberen we de ontwikkeling van een vaardigheid, bijvoorbeeld het spellen van woorden, van een leerling in kaart te brengen. Afhankelijk van de resultaten kan men dan, net als in het voorbeeld bij de zuigeling, bepalen of deze ontwikkeling al dan niet voorspoedig verloopt en, zo nodig, proberen deze ontwikkeling bij te sturen. De ontwikkeling van de vaardigheid kan men opvatten als een gestructureerd proces waarvan de structuur nog gemodelleerd dient te worden. Modellen voor een gestructureerd proces worden aangeduid als groei-, proces-, tijdreeks- of structuurmodellen. In het onderwijs zal een groeimodel veelal op het niveau van de (sub)populatie geformuleerd zijn, daar we op het individuele niveau te weinig gegevens hebben om het proces te modelleren, dat wil zeggen een model te specifi- ceren, te schatten en te toetsen. Dit is het gevolg van het feit dat in het onderwijs het volgen van leerresultaten zich meestal beperkt tot twee à drie meetmomenten per jaar. Fraaier zou het zijn een leerling frequenter te toetsen. Het mag voor een ieder duidelijk zijn dat dit praktisch niet haalbaar en zelfs niet wenselijk is. In het meest extreme geval zou een leerling bij voortduring getoetst worden, van onderwijs zou dan geen sprake meer zijn. De dagelijkse evaluering van de ontwikkeling van de leerlingen moet hoe dan ook voorbehouden blijven aan de leerkracht. De consequentie hiervan is dat de toepassing van tijdreeksmodellen voor een individuele leerling niet mogelijk zal zijn. Immers, om tijdreeksmodellen zinvol te kunnen toepassen, moet de reeks een zekere minimale lengte hebben: bijvoorbeeld 50 waarnemingen. In het onderwijs, met twee à drie toetsmomenten per jaar, komen we vaak niet verder dan 10 à 15 waarnemingen per leerling gedurende de hele schooltijd. Als bij onderwijsdata de informatie voor een individuele leerling niet uit de lengte van de tijdreeks kan komen dan moet het maar uit de breedte komen! Gelukkig is dit mogelijk door individuele tijdreeksen te beschouwen als replicaties van een onderliggende tijdreeks op populatieniveau. Dit resulteert in een opzet met herhaalde metingen op het individuele niveau met replicaties op het niveau van de populatie.

In het voorbeeld van de lichaamslengte bij zuigelingen kan men de lengte direct waar- nemen. Bovendien kan de vergelijking van de lengte van twee zuigelingen zonder omweg plaatsvinden: leg ze naast elkaar. Om de groei van een zuigeling vast te stellen, een vergelijking van dezelfde zuigeling op twee tijdstippen, zullen we een meetinstrument moeten gebruiken. De keuze van een instrument om lengte te bepalen is niet problematisch. Voor de meting van lengte kunnen we terugvallen op internationaal gemaakte afspraken: lengte meten we in meters en de lengte van een meter ligt vast. Als de meeteenheid vastligt, resteert alleen nog de keuze van een adequaat meetinstrument. Dit meetinstrument moet geijkt zijn aan de standaardmeter, geschikt zijn voor de te meten objecten en zodanig zijn dat de afleesfout beperkt blijft. Voor de meting van lichaamslengte bij baby's kunnen we dan bijvoorbeeld een schuifmaat met een verdeling in centimeters nemen. Nu is het mogelijk de lichaamslengte van dezelfde baby in de tijd te vergelijken. In wezen zijn het meetprobleem, het nauwkeurig be- palen van de lengte op een tijdstip, en het groeiprobleem, de verandering van de lengte van een object tussen twee tijdstippen, gescheiden. Dit wil zeggen dat de meetfout die we maken geen systematische componenten bevat die afhankelijk zijn van het te meten object of de te meten grootheid.

De te modelleren processen in het onderwijs hebben meestal een latente structuur, daar de vaardigheden niet direct waarneembaar zijn. Bij latente vaardigheden als spellingvaardig- heid, zullen het meet- en het groeimodel in de regel niet gescheiden zijn. Allereerst dienen we indirect vast te stellen wat spellingvaardigheid is. Stel dat we beschikken over een valide meetinstrument, toets A, voor meetmoment 1. De vraag rijst hoe we kunnen weten of we op een later tijdstip nog dezelfde spellingvaardigheid meten als bij de eerdere afname. Afgezien van de vraag of we een leerling twee keer dezelfde toets kunnen voorleggen (denk bijvoorbeeld aan geheugeneffecten) is het evident dat we niet hetzelfde dictee kunnen afnemen bij groep 3 en groep 8. Een voor groep 3 geschikt dictee zal in groep 8 naar we hopen door een ieder foutloos gemaakt worden. Met andere woorden, we kunnen niet met één toets volstaan maar we zullen een hele batterij van toetsen moeten hebben. Problematisch is het nu deze toetsen aan elkaar te ijken. We beschikken namelijk niet, zoals bij de lengtemeting, over een standaardspellingvaardigheidsmeter. Het ijken van de toetsen zal nu expliciet in een meetmodel moeten gebeuren. Afhankelijk van het gekozen meetmodel en de daarin gehanteerde schattingsmethode, zal het niet altijd mogelijk zijn het meet- en het groeimodel gescheiden aan te pakken. Voordat we aan de modellering van groei toekomen, dienen er dus nog enkele problemen opgelost te worden met betrekking tot de validering en de ijking van de meetinstrumenten. In de eerste plaats: hoe kunnen we weten of we met verschillende toetsen dezelfde latente vaardigheid meten, zowel cross-

sectioneel als longitudinaal? En in de tweede plaats: hoe kunnen de behaalde resultaten bij die toetsen met elkaar vergeleken worden?

Een ander probleem bij de vaststelling van vorderingen in leerresultaten betreft in de termen van Bock (1976), de typische onbetrouwbaarheid van leerresultaten voor een individuele leerling. Als het gaat om groepsvergelijkingen of de normering van toetsen speelt deze onbetrouwbaarheid ons geen parten, maar op het individuele niveau des te meer. Als illustratie kan de standaardmeetfout in de klassieke testtheorie dienen. Bezien we de meet- resultaten van een leerling op twee tijdstippen en zetten we met behulp van de standaardmeetfout rond deze meetresultaten een betrouwbaarheidsinterval af, dan zien we dat deze inter- vallen elkaar zeer vaak overlappen, ook als het betrouwbare toetsen betreft. Statistisch gezien is er dan geen sprake van groei.

Gezien bovenstaande problemen zal het geen sinecure zijn om individuele groei vast te stellen. Om deze problemen te overwinnen is het nodig, zoals Bock al in 1976 constateerde, de aandacht in de psychometrie te verleggen. De aandacht zal verlegd moeten worden van statische momentopnames, de relatieve positie van leerlingen in een bepaalde groep, naar methoden en modellen die op adequate wijze de groei van individuele leerlingen kunnen beschrijven en voorspellen. Het gaat er om veranderingen in het traject dat een individuele leerling aflegt te detecteren.

Drie methodologische problemen bij het volgen van individuele leerlingen verdienen gerichte aandacht. In de eerste plaats is dat de formulering van adequate meetmodellen. Deze meetmodellen moeten in ieder geval informatie leveren over de precisie van een meetresultaat. Verder is het wenselijk dat de mate van precisie kan variëren over meetresultaten. Daarnaast moet het meetmodel de koppeling kunnen verzorgen tussen groeimodel en observaties. Een tweede aandachtspunt betreft de keuze van een geschikt groeimodel. Het is wenselijk dat het groeimodel flexibel is, in die zin dat groei voor individuen of groepen van individuen verschillend kan verlopen. Het derde aandachtspunt betreft de specificatie van wat in de literatuur een verfijnd referentiekader genoemd wordt. Hiermee bedoelen we dat het mogelijk moet zijn veranderingen in individuele groei af te zetten tegen relevante andere individuen, groepen en populaties en bovendien ook tegen nader te formuleren onderwijsinhoudelijke criteria.

In dit hoofdstuk zullen we het bepalen van individuele leerresultaten in de tijd uitwerken voor de twee meest gangbare meetmodellen in de psychometrie, te weten de klassieke testtheorie en de itemresponstheorie. We zullen daarbij rekening houden met de in deze paragraaf gesignaleerde problemen. Omwille van de eenvoud beperken we ons voor het groeimodel tot een lineair model voor één populatie. Verder blijven vragen aangaande validiteit nagenoeg buiten beschouwing, ervan uitgaande dat deze reeds elders beantwoord zijn.

10.2 Klassieke testtheorie en groeiscores

In deze paragraaf werken we de bepaling van groeiscores nader uit, waarbij we het model van de klassieke testtheorie als meetmodel hanteren. Aan de hand van gesimuleerde longitudinale data zal de schattingsproblematiek van de ware score doorlopen worden. Om voor deze data de groeiscores te bepalen worden twee benaderingen gebruikt: een statische en een dynamische. Recapitulerend luidt de vraagstelling: hoe schatten we de ware score als men de data behandelt als afzonderlijke momentopnamen en welke schatters komen voor de ware score in aanmerking als we de dynamiek in de data gebruiken?

10.2.1 Artificiële longitudinale data

Stel dat de heer Knikker over de uitzonderlijke gave beschikt om knikkervaardigheid bij kinderen direct en feilloos te kunnen vaststellen. Deze heer besluit te onderzoeken inhoeverre de psychometrici dat ook kunnen. Knikker is zich bewust van het unieke van zijn gave en begrijpt dat hij de psychometrici iets concreets in handen moet geven. Hij besluit daarom een experiment te doen. Op vier momenten in een leerjaar stelt hij bij een aselecte steekproef van 1000 kinderen uit groep drie van de basisschool de knikkervaardigheid vast. Deze ware knikkervaardigheidsscores houdt hij angstvallig geheim. Knikker is bekend met het feit dat psychometrici zich meestal met toetsscores moeten behelpen, daarom genereert hij op de vier momenten voor alle kinderen in de steekproef toetsscores volgens het klassieke meetmodel:

$$y_t = \eta_t + \varepsilon_t \qquad t = 1,2,3,4 \qquad \text{(meetvergelijking klassieke testtheorie)}$$

waarbij t het meetmoment aanduidt, y_t de toetsscore op meetmoment t, η_t de ware knikker- score op meetmoment t en ε_t een door de heer Knikker toegevoegde meetfout. Merk op dat wij hier voor een andere notatie van het klassieke meetmodel dan die in hoofdstuk 3 kiezen. Om verwarring te voorkomen tussen de in hoofdstuk 3 gebruikte letter T voor de ware score en de nu geïntroduceerde tijdstipindicator, t duiden we de ware score op tijdstip t in het vervolg aan met η_t. In tegenstelling tot hoofdstuk 3 worden de toetsscore X en de meetfout e nu aangeduid met respectievelijk y en ε. De gevolgde notatie is nu in overeenstemming met de gangbare notatie in lineaire structurele modellen (Jöreskog & Sörbom, 1989). De op deze manier gegenereerde toetsscores stelt Knikker beschikbaar. Om het de psychometrici

makkelijker te maken, laat hij weten dat de toetsscores zijn gegenereerd volgens bovenstaande meetvergelijking. Verder geeft hij aan dat de meetfouten onafhankelijk zijn van de knikkervaardigheidsscores, tussen meetmomenten ongecorreleerd zijn en bovendien normaal verdeeld zijn met verwachting 0 en gelijke variantie voor alle meetmomenten. Bovendien wordt de meetfoutvariantie gegeven, $\sigma_\varepsilon^2 = 6.25$. Verder wordt ook nog bekend gemaakt dat de ware knikkervaardigheid $\eta = (\eta_1, \eta_2, \eta_3, \eta_4)'$, multivariaat normaal $N(\boldsymbol{\mu}_\eta, \Sigma_\eta)$ verdeeld is met

$$\boldsymbol{\mu}_\eta = \begin{pmatrix} 20 \\ 30 \\ 40 \\ 50 \end{pmatrix} \quad \text{en} \quad \Sigma_\eta = \begin{pmatrix} 25 & & & \\ 20 & 25 & & \\ 16 & 20 & 25 & \\ 12.8 & 16 & 20 & 25 \end{pmatrix}.$$

De vraag die de heer Knikker de psychometrici voorlegt is nu: wat zijn de ware knikkervaardigheidscores van deze kinderen op de vier meetmomenten? Twee teams van psychometrici, team A en team B, buigen zich over het probleem. Hierbij hanteert team A een statische benadering en team B een dynamische benadering. We zullen zien waarin het een en ander resulteert.

10.2.2 Statische benadering

De benadering van het probleem door team A is als volgt: men beschouwt de toetsscores op de afzonderlijke momenten als cross-secties. De longitudinale gegevensverzameling wordt opgedeeld in vier afzonderlijke delen. Elke cross-sectie kan op analoge wijze geanalyseerd worden, men besluit daarom de schattingsproblematiek allereerst alleen voor het eerste tijdstip te doorlopen (de tijdstipindex kan voorlopig achterwege blijven). Team A beheerst de theorie van hoofdstuk 3 goed en komt op grond van de klassieke testtheorie tot de volgende globale conclusies. In de eerste plaats constateert men dat de gekwadrateerde correlatie tussen de geobserveerde scores en de ware scores in de populatie, de betrouwbaarheid, wordt gegeven door

$$\rho^2_{Y\eta} = \frac{\sigma^2_\eta}{\sigma^2_Y} = \frac{\sigma^2_\eta}{\sigma^2_\eta + \sigma^2_\varepsilon} = \frac{25}{25 + 6.25} = .8. \qquad (10.1)$$

In de tweede plaats geldt dat de regressie van de geobserveerde toetsscore op de ware score,

$$\mathscr{E}(Y \mid \eta) = \eta, \tag{10.2}$$

lineair is. Men haalt opgelucht adem, uit (10.2) kan men concluderen dat Y, de geobserveerde score, een zuivere schatter voor η is. Hoe goed die schatter is, wordt gegeven door de betrouwbaarheid (10.1) en de schattingsfoutvariantie zal gelijk zijn aan de meetfoutvariantie σ_ε^2. Team A geeft in eerste instantie hoog op van de kwaliteiten van Y als schatter van η; deze schatter zullen in het vervolg aangeven met $\hat{\eta}$. Na enige overpeinzingen is men toch niet helemaal tevreden met deze schatter. Wat heeft men eigenlijk aan de conditionele verwachting, $\mathscr{E}(Y \mid \eta)$, als Y bekend en η onbekend is? Eigenlijk zou men de conditionele verwachting van η gegeven Y willen hebben. Verder geldt dat voor de schatting van de ware score van een individuele leerling op een meetmoment men niet over replicaties beschikt, slechts één waarneming is beschikbaar. Dit impliceert dat de zuiverheid van de geobserveerde score als schatter, op het individuele niveau bezien, niet bar veel betekent. Bij de bepaling van de verwachting, $\mathscr{E}(Y \mid \eta)$, introduceren we als gevolg van de kleine steekproef (één waarneming voor een leerling), een onzuiverheid die gelijk is aan de meetfout ε voor die ene waarneming. Ook denkt men dat er schatters te vinden zijn die een kleinere schattingsfoutvariantie hebben daar men meer informatie kan gebruiken. De verwaarloosde informatie betreft de a priori kennis met betrekking tot η, η is immers getrokken uit een bekende verdeling.

Men besluit verder te zoeken. Het punt van de verwaarloosde informatie levert gelijk een andere schatter van η op: het gemiddelde μ_η van de (marginale) verdeling van η. De schattingsfoutvariantie van deze schatter, $\tilde{\eta}$, is dan gegeven door de variantie van de (marginale) verdeling, σ_η^2. Meer algemeen kan de a priori informatie geschreven worden als

$$\eta = \mu_\eta + \zeta, \qquad \text{(a priori informatie)}$$

waarbij is ζ een meetfoutvariabele is met verwachting 0 en variantie σ_η^2.

Al snel concludeert men dat dit geen groot succes is: onzuiverheid en schattingsfoutvariantie zijn voor de a priori schatter groter dan voor de geobserveerde score schatter. Nader onderzoek leert dat deze twee schatters onafhankelijk zijn en bovendien allebei zuiver zijn in de populatie, dat wil zeggen

$$\mathscr{E}_Y(\hat{\eta}) = \mathscr{E}_Y(\tilde{\eta}) = \mu_\eta.$$

Het ligt nu voor de hand deze schatters te combineren. De optimale combinatie van twee zuivere schatters, zeg $\hat{\eta}_1$ en $\hat{\eta}_2$ met bijbehorende schattingsfoutvarianties P_1 en P_2 wordt gegeven door

$$\eta^* = P(P_1^{-1}\,\hat{\eta}_1 + P_2^{-1}\,\hat{\eta}_2), \tag{10.3}$$

waarbij P, de schattingsfoutvariantie van deze schatter, gegeven wordt door

$$P = (P_1^{-1} + P_2^{-1})^{-1}. \tag{10.4}$$

Substitutie van de a priori schatter (μ_η) en de geobserveerde score schatter (Y) en bijbehorende schattingsfoutvarianties respectievelijk σ_ε^2 en σ_η^2 in (10.3) en (10.4) levert dan

$$\eta^* = \frac{\sigma_\varepsilon^2}{\sigma_\eta^2 + \sigma_\varepsilon^2}\,\mu_\eta + \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma_\varepsilon^2}\,y, \tag{10.5}$$

en

$$P = \frac{\sigma_\varepsilon^2\,\sigma_\eta^2}{\sigma_\eta^2 + \sigma_\varepsilon^2}. \tag{10.6}$$

Deze resultaten in ogenschouw nemend herkent men hierin de Kelley-schatter voor de ware score (de kleinste-kwadratenschatter $\mathscr{E}(\eta \mid Y)$), waarmee men al bekend was uit de klassieke testtheorie (zie hoofdstuk 3). Kelley vond dit al een interessante schatter voor de ware score, daar deze schatter de gewogen som is van twee afzonderlijke schatters, één gebaseerd op de geobserveerde score van de persoon en de ander op het gemiddelde van de groep waartoe deze persoon behoort. Als de betrouwbaarheid van de toets hoog is, wordt deze schatter voornamelijk bepaald door de toetsscore Y, bij een lage betrouwbaarheid voornamelijk door het groepsgemiddelde μ_η (Lord & Novick, 1968, p. 65).
Team A is tevreden. Voor de duidelijkheid zet men de drie schatters met bijbehorende varianties van de schattingsfout nog eens onder elkaar:

$\tilde{\eta}_t = \mathscr{E}(\eta_t) = \mu_{\eta_t}$ $\qquad$ $P_t = \sigma_{\eta_t}^2$ $\qquad$ a priori schatter

$\hat{\eta}_t = \mathscr{E}(Y_t \mid \eta_t) = y_t$ $\qquad$ $P_t = \sigma_\varepsilon^2$ geobserveerde-score-schatter

$$\eta_t^* = \mathscr{E}(\eta_t \mid Y_t) = \frac{\sigma_\varepsilon^2}{\sigma_{\eta_t}^2 + \sigma_\varepsilon^2}\,\mu_{\eta_t} + \frac{\sigma_{\eta_t}^2}{\sigma_{\eta_t}^2 + \sigma_\varepsilon^2}\,y_t \qquad P_t = \frac{\sigma_\varepsilon^2\,\sigma_{\eta_t}^2}{\sigma_{\eta_t}^2 + \sigma_\varepsilon^2} \text{ Kelley-schatter}$$

Om de berekening van de schattingen van de ware scores voor de 1000 leerlingen in de steekproef op de vier tijdstippen te vereenvoudigen, maakt men gebruik van tabel 10.1.

Tabel 10.1
Schatters en schattingsfoutvarianties voor de vier tijdstippen

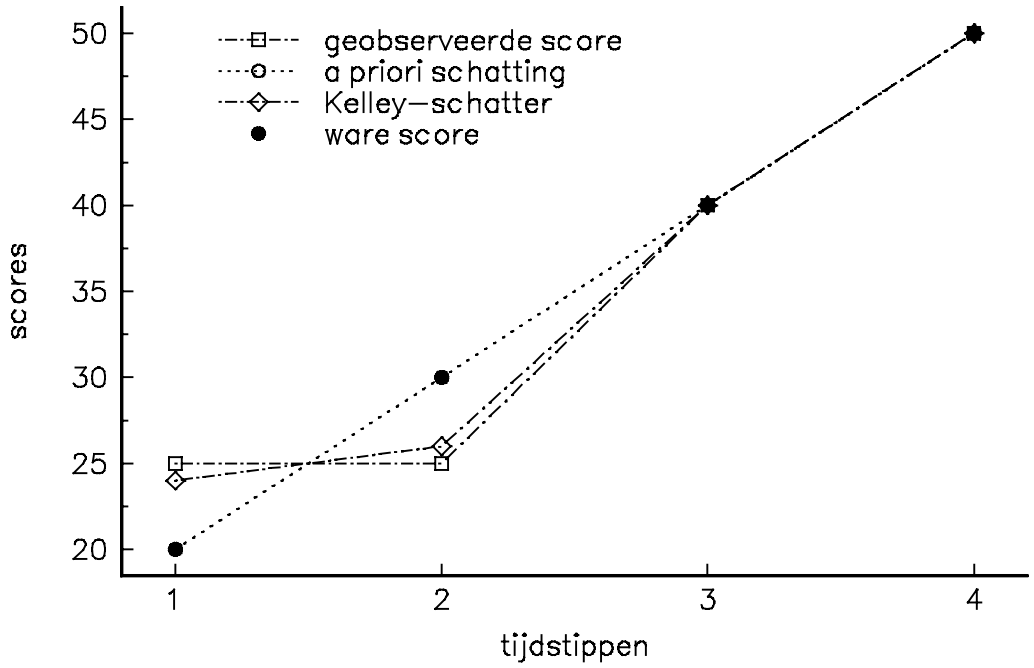
	a priori		geobserveerde score		Kelley	
tijdstip	$\tilde{\eta}$	P	$\hat{\eta}$	P	η^*	P
1	20	25	y_1	6.25	$4+.8y_1$	5
2	30	25	y_2	6.25	$6+.8y_2$	5
3	40	25	y_3	6.25	$8+.8y_3$	5
4	50	25	y_4	6.25	$10+.8y_4$	5

Om enig inzicht te verkrijgen in het functioneren van deze drie schatters, besluit men om voor twee leerlingen het gedrag van deze schatters te onderzoeken. Er van uitgaande dat leerling A op alle vier de tijdstippen een ware score heeft die gelijk is aan het populatiegemiddelde (ware scores: 20, 30, 40 en 50), creëert men de volgende observaties voor de vier tijdstippen: 25, 25, 40 en 50. De toegevoegde meetfout is dus respectievelijk: 5, -5, 0 en 0. In figuur 10.2 zijn de ware scores en de drie besproken schattingen van de ware scores voor leerling A weergegeven voor de vier tijdstippen.

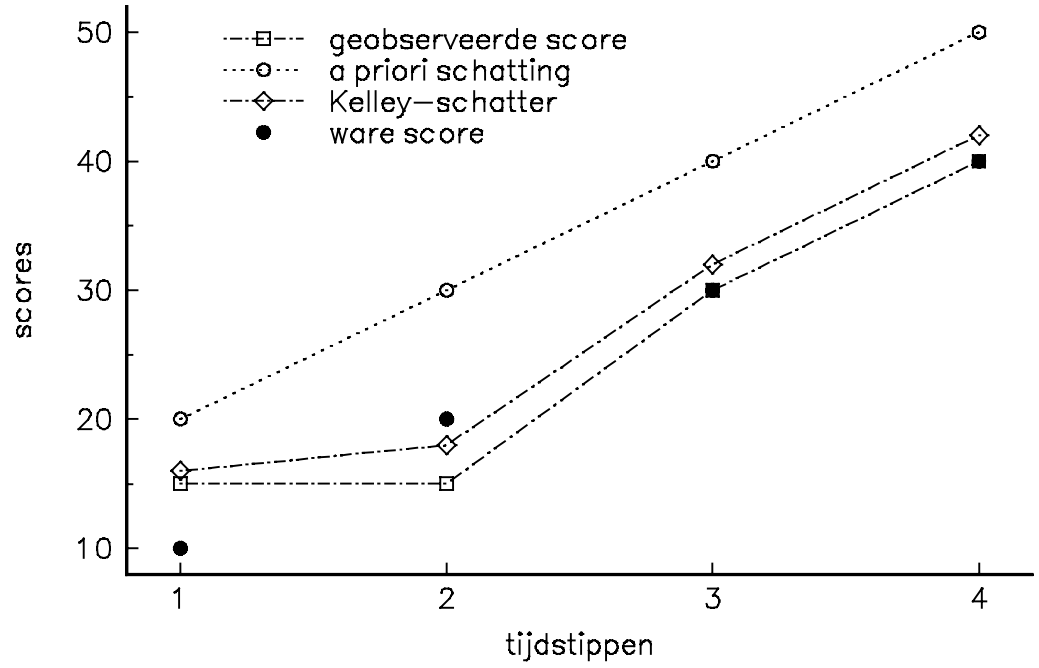
In de eerste plaats kunnen we in figuur 10.2 constateren dat de a priori schatting op alle tijdstippen samenvalt met de ware score, niet zo verwonderlijk als men zich realiseert dat de a priori schatting de gemiddelde ware score in de populatie is. Op tijdstip 3 en 4 vallen ook de geobserveerde score schattingen samen met de respectievelijke ware scores, ook niet opzienbarend daar de toegevoegde meetfout op dat tijdstip 0 was. Omdat de a priori schatting en geobserveerde-score-schatting voor tijdstip 3 en 4 samenvallen, resulteren ook de Kelley-schattingen in de ware scores voor leerling A. De geobserveerde-score-schattingen op tijdstip 1 en 2 zitten er behoorlijk naast, de mate waarin is bepaald door de toegevoegde meetfout, dat is respectievelijk plus en minus 2 × de standaardafwijking van de meetfout. Op tijdstip 1 en 2 functioneert de Kelley-schatter beter dan de geobserveerde-score-schatter, de Kelley-schatter duwt (Engels: 'shrinkage') de geobserveerde scores in de richting van de a priori schatter en

komt zodoende dichter in de buurt van de ware scores. Hoe hard de Kelley-schatter duwt, wordt bepaald door de betrouwbaarheid van de observaties (zie tabel 10.1).

De ware scores voor leerling B zijn respectievelijk 10, 20, 30 en 40. De toegevoegde meetfout is respectievelijk: 5, -5, 0 en 0, hetgeen resulteert in de geobserveerde scores 15, 15, 30 en 40. In figuur 10.3 zijn de ware score schattingen weergegeven voor leerling B.



Figuur 10.2
Schattingen van de ware scores voor de 'gemiddelde' leerling A



Figuur 10.3 Schattingen van de ware scores voor leerling B

De a priori schattingen zitten er behoorlijk naast, en wel 10 scorepunten. Het verschil tussen de geobserveerde scores en de ware scores bij leerling B is hetzelfde als bij leerling

A en is gelijk aan de toegevoegde meetfout op de 4 momenten, respectievelijk 5, 5,0 en 0. Ook hier duwt de Kelley-schatter de geobserveerde scores in de richting van de a priori schat- ter. Op tijdstip 1, 3 en 4 is het effect hiervan dat de afstand tussen de ware score en de Kelley-schatting groter is dan die tussen de ware score en de geobserveerde score. Op tijdstip 2 geldt het omgekeerde.

Uit deze twee voorbeelden kunnen we concluderen dat geen van de drie besproken schatters het onder alle omstandigheden goed doet. Afhankelijk van de relatieve positie van een leerling in de populatie en de grootte van de meetfout, gaat de voorkeur uit naar een van de drie schatters. Welke schatter over individuen heen het predikaat 'best' verdient, zullen we bespreken nadat de dynamische benadering besproken is.

10.2.3 Dynamische benadering

Ook team B begint met een inspectie van de meetvergelijking in de klassieke testtheorie, maar beperkt zich in eerste instantie tot één meetmoment. Men realiseert

zich dat de meetvergelijking in de klassieke testtheorie de relatie beschrijft tussen toevalsvariabelen in de populatie. Met deze constatering als uitgangspunt gaat men het schattingsprobleem van de ware score voor een bepaald individu specificeren. De meetvergelijking in de klassieke testtheorie beschrijft niets anders dan de relatie tussen de toevalsvariabelen Y en η in een populatie, met een niet gespecificeerde gezamenlijke verdeling. De observeerbare variabele Y is in dit geval behept met een meetfout. Intuïtief is het duidelijk dat de meting van Y ons iets kan leren over η. Of, anders gezegd, stel dat we over a priori informatie over η beschikten, dan zou kennis van Y deze informatie omtrent η moeten verbeteren. De volgende vraag is nu relevant: "Gegeven de observatie $Y = y$, wat is dan de beste schatting van η?" Eerst geven we inhoud aan het concept 'best'. Een veel gebruikt criterium hiervoor is dat van de kleinste-kwadraten. Hierbij wordt gezocht naar een schatter $\eta^*(Y)$ die een functie is van de meting $Y = y$ zodanig dat

$$\mathcal{E}[\eta - \eta^*(Y)]^2 \leq \mathcal{E}[\eta - g(Y)]^2, \tag{10.7}$$

voor elke functie g. De oplossing van (10.7) wordt gegeven door

$$\eta^*(Y) = \mathcal{E}(\eta \mid Y).$$

Merk nu op dat $\eta^*(Y)$ een toevalsvariabele is, in tegenstelling tot de realisatie $\eta^*(y)$ daar- van voor observatie $Y = y$. Problematisch is dat $\eta^*(Y)$ meestal niet een lineaire functie van Y is. Daarnaast beschikken we in de klassieke testtheorie meestal niet over de gezamenlijke verdeling van η en Y, zodat het onmogelijk is om $\mathcal{E}(\eta \mid Y)$ te bepalen. Daarom zullen we een extra aanname maken. We veronderstellen namelijk dat $\eta^*(Y)$ een lineaire functie van Y is,

$$\eta^*(Y) = aY + b \tag{10.8}$$

waarbij a en b te bepalen constanten zijn. De oplossing van (10.8), onder de restrictie van vergelijking (10.7), is gegeven door:

$$a = \frac{\sigma_{Y\eta}}{\sigma_\eta^2} \tag{10.9}$$

en

$$b = \mu_\eta - \frac{\sigma_{Y\eta}}{\sigma_Y^2}\,\mu_Y, \tag{10.10}$$

waarbij $\sigma_{Y\eta}$ de covariantie tussen Y en η is. Substitutie van (10.9) en (10.10) in (10.8) levert dan de beste lineaire schatter van η gebaseerd op Y:

$$\eta^*(Y) = \mu_\eta - \frac{\sigma_{Y\eta}}{\sigma_Y^2}\mu_Y + \frac{\sigma_{Y\eta}}{\sigma_Y^2} Y. \tag{10.11}$$

De variantie van de schattingsfout is gegeven door

$$P = \mathscr{E}[\eta - \eta^*(Y)]^2 = \sigma_\eta^2 - \frac{\sigma_{Y\eta}^2}{\sigma_Y^2}. \tag{10.12}$$

Het geoefende oog van team B herkent in (10.11) en (10.12) natuurlijk de Kelley-schatter met bijbehorende schattingsfoutvariantie (herschrijf (10.5) en (10.6) en maak hierbij gebruik van de formules uit de klassieke testtheorie). Daar in dit voorbeeld de ware vaardigheidsverdeling multivariaat normaal en de meetfout normaal verdeeld is, is ook de conditionele verdeling van η gegeven Y normaal verdeeld, waarbij het gemiddelde gegeven wordt door (10.11) en de variantie door (10.12).
Nu men het schattingsprobleem in essentie voor twee toevalsvariabelen heeft opgelost gaat men dit toepassen in een longitudinale context. De subscripten bij de variabelen die in het vervolg gebruikt worden geven nu de tijdstippen weer. Op het eerste meetmoment lijkt de Kelley-schatter en schattingsfoutvariantie de aangewezen keus, dus

$$\begin{aligned} \eta_1^* &= \mu_{\eta_1} - \frac{\sigma_{Y_1\eta_1}}{\sigma_{Y_1}^2}\mu_{Y_1} + \frac{\sigma_{Y_1\eta_1}}{\sigma_{Y_1}^2} Y_1, \\ P_1 &= \sigma_{\eta_1}^2 - \frac{\sigma_{Y_1\eta_1}^2}{\sigma_{Y_1}^2}. \end{aligned} \tag{10.13}$$

In tegenstelling tot team A onderkent team B dat, gegeven de knikkervaardigheidsverdeling in de populatie, het mogelijk is η_2 te voorspellen met η_1. Inmiddels weten we hoe dat moet en de oplossing wordt gegeven door

$$\eta_{2|1}^* = \mathscr{E}(\eta_2 \mid \eta_1) = \mu_{\eta_2} - \frac{\sigma_{\eta_1\eta_2}}{\sigma_{\eta_1}^2}\mu_{\eta_1} + \frac{\sigma_{\eta_1\eta_2}}{\sigma_{\eta_1}^2}\eta_1. \tag{10.14}$$

In de praktijk beschikken we niet over η_1; we zullen ons tevreden moeten stellen met een schatting hiervan, zeg η_1^*. Voorspellen is nu niets anders dan substitutie van deze schatting (10.13) in (10.14) hetgeen resulteert in:

$$\eta^*_{2|1} = \mu_{\eta_2} - \frac{\sigma_{\eta_1\eta_2}}{\sigma^2_{\eta_1}}\,\mu_{\eta_1} + \frac{\sigma_{\eta_1\eta_2}}{\sigma^2_{\eta_1}}\,\eta^*_1,$$

ofwel

$$\eta^*_{2|1} = \mu_{\eta_2} - \frac{\sigma_{\eta_1\eta_2}}{\sigma^2_{\eta_1}}\,\frac{\sigma_{y_1\eta_1}}{\sigma^2_{y_1}}\,(y_1 - \mu_{y_1}). \tag{10.15}$$

De berekening van de variantie van (10.15) gaat recht toe recht aan en levert op:

$$P_{2|1} = \sigma^2_{\eta_2} - \frac{\sigma^2_{\eta_1\eta_2}\sigma^2_{y_1\eta_1}}{\sigma^4_{\eta_1}\sigma^2_{y_1}}. \tag{10.16}$$

Deze voorspelling en schattingsfoutvariantie zijn in wezen niets anders dan de a priori informatie met betrekking tot η_2 gegeven de waarneming y_1 op tijdstip 1. Merk op dat deze a priori informatie in feite een a priori verdeling voor η_2 is met gemiddelde $\eta^*_{2|1}$ en variantie $P_{2|1}$, die in ons voorbeeld normaal verdeeld is. Als we op tijdstip 2 deze a priori informatie in het dynamische geval vergelijken met de a priori informatie bij de statische benadering, dan valt op dat het gemiddelde μ_{η_2} in het dynamische geval gecorrigeerd wordt (vergelijking 10.15) en dat de variantie $\sigma^2_{\eta_2}$ verkleind wordt (zie 10.16). Met andere woorden, onze a priori informatie op tijdstip 2 wordt meer specifiek voor een individu, daar we immers rekening houden met de geobserveerde score y_1 van dit individu. Bovendien is de hoeveelheid informatie groter, zodat de onzekerheid over iemands positie in de populatie afneemt.

Als we het meetresultaat op tijdstip 2, y_2, willen combineren met de a priori kennis op tijdstip 2, dan kan dat beschreven worden als het combineren van twee schatters (zie ook paragraaf 10.2.2 voor de combinatie van een a priori schatter en de geobserveerde-score-schatter) of, analoog aan hierboven, door het bepalen van de conditionele verwachting $\mathscr{E}(\eta_2 \mid Y_1, Y_2)$. Beide resulteren in de volgende schatting voor η_2:

$$\eta^*_2 = \eta^*_{2|1} + K_2(y_2 - \eta^*_{2|1}),$$

waarbij K_2 gegeven is door:

$$K_2 = P_{2|1}(P_{2|1} + \sigma^2_\varepsilon)^{-1}.$$

De bijbehorende schattingsfoutvariantie, P_2, wordt gegeven door:

$$P_2 = P_{2|1} - K_2 P_{2|1}.$$

De bepaling van een schatter voor η_3 gaat analoog aan de procedure voor η_2. Voorspel η_3 met behulp van η_2, vul de lopende schatting van η_2 in deze vergelijking in en combineer deze predictie met de observatie y_3 op het derde tijdstip. Uiteraard kunnen we zo doorgaan voor de volgende tijdstippen. Merk op dat we voor de voorspelling van η_3 alleen η_2 gebruiken en niet η_1. Met andere woorden, we gaan ervan uit dat, gegeven η_2, η_1 ons niets meer kan leren over η_3. Of anders gezegd, de partiële correlatie tussen η_1 en η_3 veronder- stellen we gelijk aan nul. Dat geldt ook op de andere tijdstippen, dus alle partiële correlaties tussen de latente variabelen zijn 0, behalve voor aanliggende tijdstippen. Dit impliceert dat de covariantiematrix van η een bepaalde structuur heeft, die in de literatuur aangeduid wordt met 'autoregressief van de eerste orde'. De hier beschreven recursieve schattingsprocedure staat bekend als het Kalmanfilter, de schattingen als Kalmanfilterschattingen.

Team B is tevreden met het resultaat. Men signaleert echter één minpunt. Men realiseert zich dat de Kalmanfilterschattingen voor de vier tijdstippen niets anders zijn dan de conditionele verwachtingen: $\mathcal{E}(\eta_1 \mid y_1)$, $\mathcal{E}(\eta_2 \mid y_1, y_2)$, $\mathcal{E}(\eta_3 \mid y_1, y_2, y_3)$ en $\mathcal{E}(\eta_4 \mid y_1, y_2, y_3, y_4)$. Bezien we deze reeks, dan kan geconstateerd worden dat het aantal waarnemingen waarop deze conditionele verwachtingen gebaseerd zijn in de tijd toeneemt. Op het eerste tijdstip gebruiken we slechts één waarneming, terwijl op het vierde tijdstip gebruik gemaakt is van alle meetresultaten. Beschikken we over vier waarnemingen, dan geldt alleen voor de Kalmanfilterschatting op het vierde tijdstip dat alle informatie uit de data verwerkt is in de schatter. Voor de Kalmanfilterschatting op tijdstip 3, bijvoorbeeld, hebben we geen gebruik gemaakt van de laatste waarneming. Het ligt dus voor de hand die informatie alsnog toe te voegen, dat is, door $\mathcal{E}(\eta_3 \mid y_1, y_2, y_3, y_4)$ te bepalen. Voor de Kalmanfilterschattingen op tijd- stip 2 en 1, berekenen we dan respectievelijk $\mathcal{E}(\eta_2 \mid y_1, y_2, y_3, y_4)$ en $\mathcal{E}(\eta_1 \mid y_1, y_2, y_3, y_4)$. De conditionele verwachting van η, op een tijdstip gegeven alle data duidt men aan met de naam gladgestreken Kalmanfilterschatting. Het bepalen van de gladgestreken schattingen kan eenvoudig geïllustreerd worden aan het kleinste-kwadratenprobleem in het begin van deze paragraaf. Daar zochten we de conditionele verwachting van η gegeven Y. Maar dit is in wezen niets anders dan de univariate regressie van η op Y. Stel dat we de multivariate lineaire regressie bepalen van de vector η op de vector $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)'$, dat is

$$\hat{\eta} = \mathcal{E}(\eta \mid \boldsymbol{Y}) = \boldsymbol{\mu}_\eta + \Sigma_{\boldsymbol{Y}\eta}\, \Sigma_{\boldsymbol{Y}}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{Y}}), \tag{10.17}$$

dan beschikken we in een keer over de gladgestreken schattingen in de vector $\hat{\eta}$. De covariantiematrix van de gladgestreken schattingen is

$$\boldsymbol{P} = \Sigma_{\eta} - \Sigma_{\boldsymbol{Y}\eta}\,\Sigma_{\boldsymbol{Y}}^{-1}\,\Sigma_{\eta}. \qquad (10.18)$$

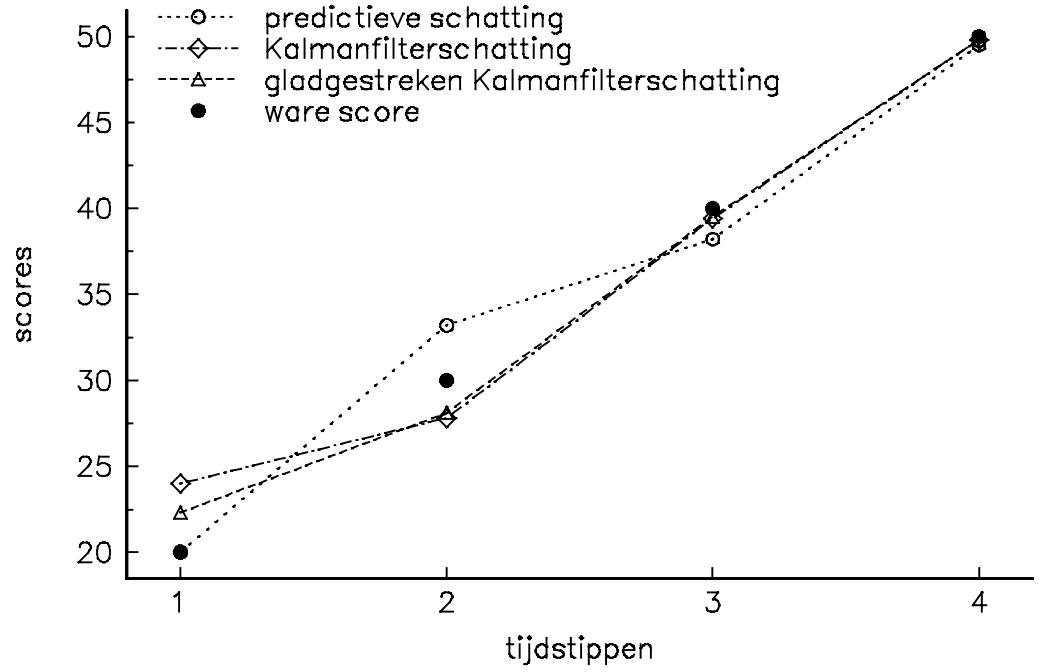
Merk op dat voor de klassieke testtheorie geldt dat de covariantiematrix tussen de vectoren $\boldsymbol{Y}$ en η, $\Sigma_{\boldsymbol{Y}\eta}$, gelijk is aan de variantie-covariantiematrix van de vector η, dat wil zeggen $\Sigma_{\boldsymbol{Y}\eta} = \Sigma_{\eta}$.

Tabel 10.2
Schatters en schattingsfoutvarianties voor de vier tijdstippen

	predictie		Kalmanfilter		gladgestreken Kalmanfilter	
tijd-stip	$\eta^*_{t\mid t-1}$	$P_{t\mid t-1}$	η^*_t	P_t	$\hat{\eta}_t$	P_t
1	20	25	$4 + .8y_1$	5	$\eta^*_1 + .33(\hat{\eta}_2 - 14 - .8\eta^*_1)$	4.06
2	$14 + .8\eta^*_1$	12.20	$\eta^*_{2\mid 1} + .66(y_2 - \eta^*_{2\mid 1})$	4.13	$\eta^*_2 + .28(\hat{\eta}_3 - 16 - .8\eta^*_2)$	3.47
3	$16 + .8\eta^*_2$	11.65	$\eta^*_{3\mid 2} + .65(y_3 - \eta^*_{3\mid 2})$	4.07	$\eta^*_3 + .28(\hat{\eta}_4 - 18 - .8\eta^*_3)$	3.47
4	$18 + .8\eta^*_3$	11.60	$\eta^*_{4\mid 3} + .65(y_4 - \eta^*_{4\mid 3})$	4.06	η^*_4	4.06

Een recursieve procedure (nu achterwaarts) voor het berekenen van de gladgestreken schat- tingen, waarin alleen gebruik gemaakt wordt van de predictieve filterschattingen en Kalman- filterschattingen met bijbehorende covarianties, staat vermeld in Jazwinski (1970).

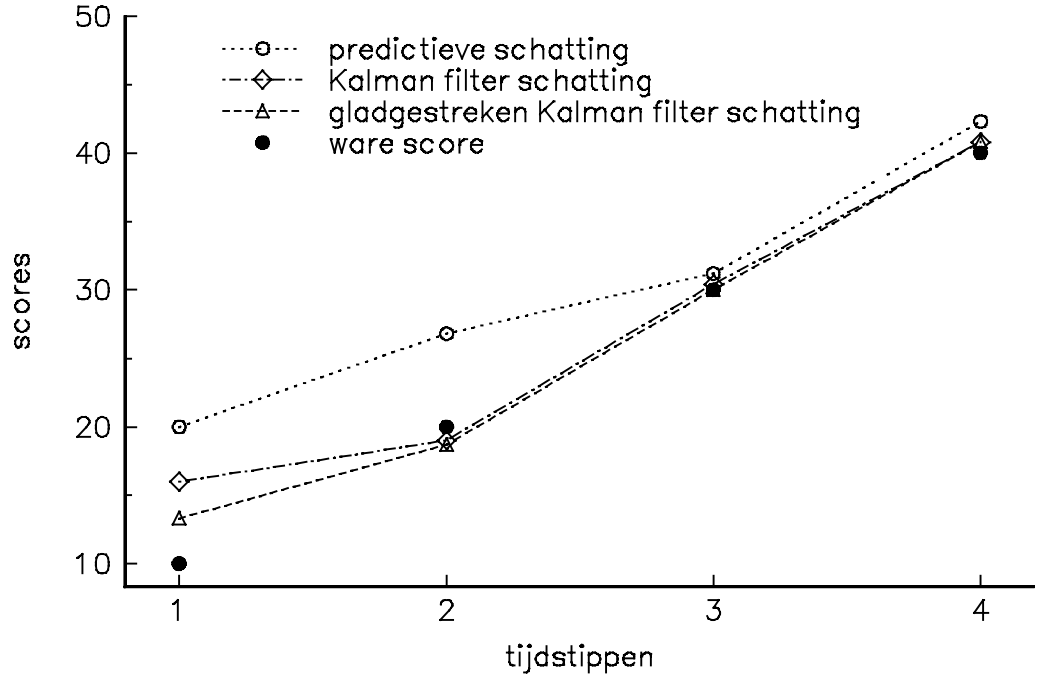
Ook team B gaat de ware scores uitrekenen voor de 1000 leerlingen in de steekproef. In tabel 10.2 zijn zijn de resultaten voor de predictie-, de Kalmanfilter- en de gladgestreken Kalmanfilterschattingen op de vier tijdstippen vermeld. Tenslotte kijkt team B naar het functioneren van de door hen geconstrueerde schatters.



Figuur 10.4 Schattingen van de ware scores voor de 'gemiddelde' leerling A

Om het gedrag van de schatters te onderzoeken maakt men, net als bij team A, gebruik van de scores van leerling A en leerling B (zie paragraaf 10.2.2). In figuur 10.4 zijn de resultaten voor leerling A weergegeven. Op het eerste tijdstip is de a priori kennis bij de statische en dynamische aanpak even groot, met uitzondering van de gladgestreken Kalmanfilterschatting. De reden hiervoor is dat de a priori schatting en de predictieve schatting samenvallen, dus ook de Kelley-schatting en de Kalmanfilterschatting. Merk op dat de gladgestreken schatting op het eerste meetmoment en in mindere mate op het tweede tijdstip het dichtst komt bij de ware score. Voor deze leerling kan de informatie uit de latere tijdstippen dus de schattingen op de eerste twee tijdstippen tot op zekere hoogte in de goede richting corrigeren. Kijken we naar de schattingen voor leerling B (zie figuur 10.5), dan valt op dat de predictieve schattingen op de laatste drie meetmomenten dichter bij de ware scores liggen dan de a priori schattingen in het statische geval.

Dit heeft tot gevolg dat de Kalmanfilterschattingen op deze momenten de ware scores beter benaderen dan de Kelley-schattingen bij de statische benadering. Het plaatje is wederom het fraaist voor de gladgestreken schattingen. Deze schattingen, komen over de vier tijdstippen bezien, immers het dichtst bij de ware scores.



Figuur 10.5

Schattingen van de ware scores voor leerling B

10.2.4 Evaluatie statische en dynamische benadering

Het wordt tijd om de door team A en team B voorgestelde schatters te evalueren. Beide teams hebben voor de 1000 leerlingen in de steekproef op alle vier de tijdstippen schattingen en bijbehorende schattingsfoutvarianties uitgerekend en ter evaluatie aan de heer Knikker aan- geboden. Om de schatters te kunnen evalueren, zullen we eerst enige criteria moeten aan- nemen waarop de evaluatie van de schatters kan plaatsvinden. De heer Knikker besteedt deze klus uit aan een statisticus, aan wie hij alle materiaal, inclusief de ware scores, beschikbaar stelt. Deze statisticus ziet twee mogelijke manieren om de zaak te evalueren. In de eerste plaats kan hij de schatters beoordelen op hun statistische eigenschappen. Omdat alle gegevens beschikbaar zijn, kan hij ook de schattingen en de ware scores van alle 1000 leerlingen vergelijken; dit is de tweede manier.

We bekijken eerst de statistische eigenschappen. In de eerste plaats valt op dat alle voorgestelde schatters, zowel die van team A als die van team B, kleinste-kwadraten-schatters zijn, die alleen verschillen in de mate waarin ze de beschikbare informatie gebruiken. De volgende tabel 10.3 vat de bron en de hoeveelheid informatie voor de diverse schatters samen. De bron van de informatie refereert aan het meetmodel en het groeimodel, terwijl de hoeveelheid informatie het aantal tijdstippen aanduidt.

Tabel 10.3

Hoeveelheid informatie van de diverse schatters uitgesplitst naar bron

	bron informatie	
	groeimodel	meetmodel
schatter op tijdstip t	η_t	y_t
a priori	t	geen
geobserveerde score	geen	t
Kelley	t	t
predictieve	1 t/m t	1 t/m $t-1$
Kalmanfilter	1 t/m t	1 t/m t
gladgestreken Kalmanfilter	alle	alle

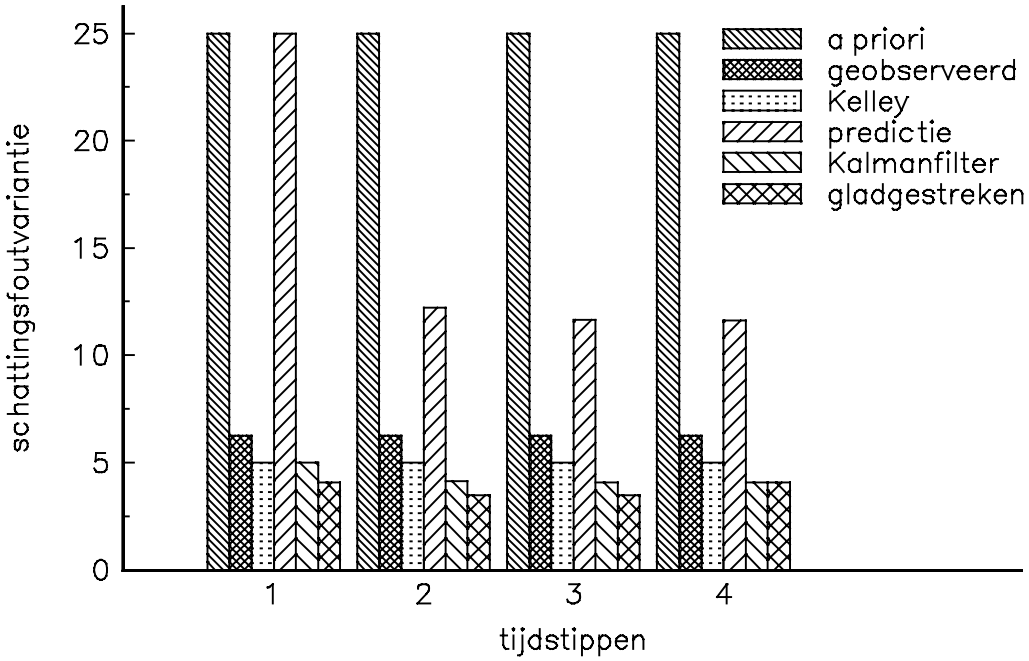
Naarmate een schatter meer informatie gebruikt is de schattingsfoutvariantie kleiner, zoals uit de statistiek bekend is. De schattingsfoutvariantie, als indicatie voor de zekerheid van de schatting, is dan ook het eerste criterium om de schatters te vergelijken. Merk op dat, met de klassieke testtheorie als meetmodel, alle schattingsfoutvarianties op voorhand bekend zijn zonder ook maar een observatie gedaan te hebben, dat is als men de relatie kent tussen de toevalsvariabelen η_t en Y_t. In figuur 10.6 zijn met behulp van staafdiagrammen op de vier tijdstippen de schattingsfoutvarianties van de zes besproken schatters grafisch weergegeven.

We vergelijken eerst de schattingsfoutvarianties van de drie cross-sectionele schatters. De schattingsfoutvarianties van de afzonderlijke schatters zijn over de vier tijdstippen gelijk (gelijke betrouwbaarheid voor elk tijdstip). De kleinste schattingsfoutvariantie heeft de Kelley- schatter (5), gevolgd door de geobserveerde-score-schatter (6.25) en de a priori schatter (25). In het algemeen kan men zeggen dat van de cross-sectionele schatters de Kelley-schatter altijd de kleinste variantie heeft.

De Kelley-schatter gebruikt immers alle cross-sectionele informatie. De betrouwbaarheid van

de toets bepaalt de volgorde van de andere twee cross-sectionele schatters. Is de betrouwbaarheid groter dan .5, dan heeft de geobserveerde-score-schatter een kleinere variantie dan de a priori schatter; het omgekeerde geldt als de betrouwbaarheid kleiner is dan .5. Kijken we vervolgens naar de dynamische benadering, dan zien we dat de gladgestreken Kalmanfilterschatter op alle tijdstippen de kleinste schattingsfoutvariantie heeft, behoudens op het laatste tijdstip waarop deze schatter gelijk is aan de

Kalmanfilterschatter. Ook zien we dat de Kalmanfilterschatters het beter doen dan de predictieve schatters. Dit is logisch, daar de eerstgenoemde schatter in vergelijking met de predictieve schatter een extra waarneming, dat wil zeggen, extra informatie gebruikt. De orde van grootte van de schattingsfoutvariantie van de predictieve schatter hangt natuurlijk af van de mate waarin we in staat zijn de latente vaardigheid te voorspellen op een volgend tijdstip. Een maat hiervoor is de gekwadrateerde correlatie tussen de latente vaardigheden op twee tijdstippen. Een vergelijking van de schattingsfoutvarianties van de statische en dynamische schatters leert ons dat de statische equivalenten van de dynamische schatters een beduidend grotere schattingsfoutvariantie hebben. Hoe groot de verschillen zijn, hangt in het algemeen af van de toetsbetrouwbaarheid en van de mate waarin de latente vaardigheid voorspeld kan worden.

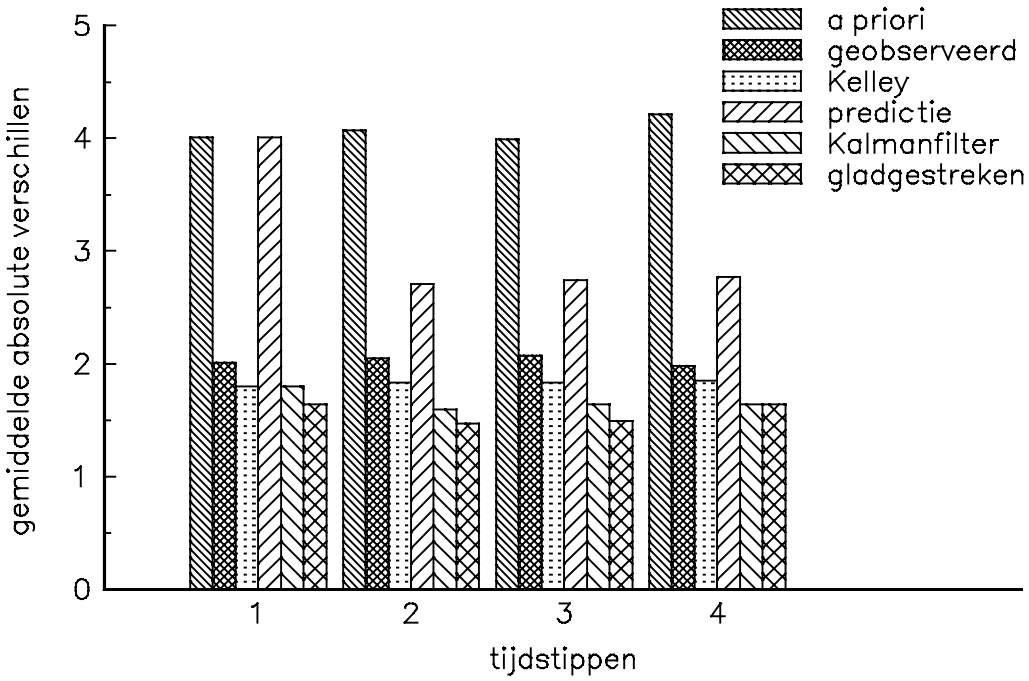


Figuur 10.6
Histogram voor de diverse schattingsfoutvarianties voor de vier tijdstippen

De tweede statistische eigenschap om schatters te beoordelen is de zuiverheid van schatters. Alle besproken schatters zijn zuiver in de populatie terwijl de geobserveerde-score-schatter bovendien zuiver is voor een individu. Aan deze laatstgenoemde vorm van zuiverheid hebben we echter niet zoveel, aangezien we op een tijdstip voor een individu meestal niet over replicaties beschikken. Wel kan deze eigenschap van de geobserveerde-score-schatter handig zijn voor het berekenen van groepsgemiddelden. Denk hierbij bijvoorbeeld aan een apart gemiddelde voor jongens en meisjes.

De statisticus concludeert dat op het criterium zuiverheid de schatters elkaar in wezen niet ontlopen en besluit daarom, het criterium zuiverheid niet te laten meewegen en zich alleen te beperken tot de schattingsfoutvariantie.

Een tweede evaluatiemogelijkheid behelst het vergelijken van de schattingen en de ware scores in de steekproef. Twee criteria om de schatters te beoordelen, acht de statisticus zinvol



Figuur 10.7

Histogram gemiddelde absolute verschil ware scores en diverse schattingen voor de vier tijdstippen

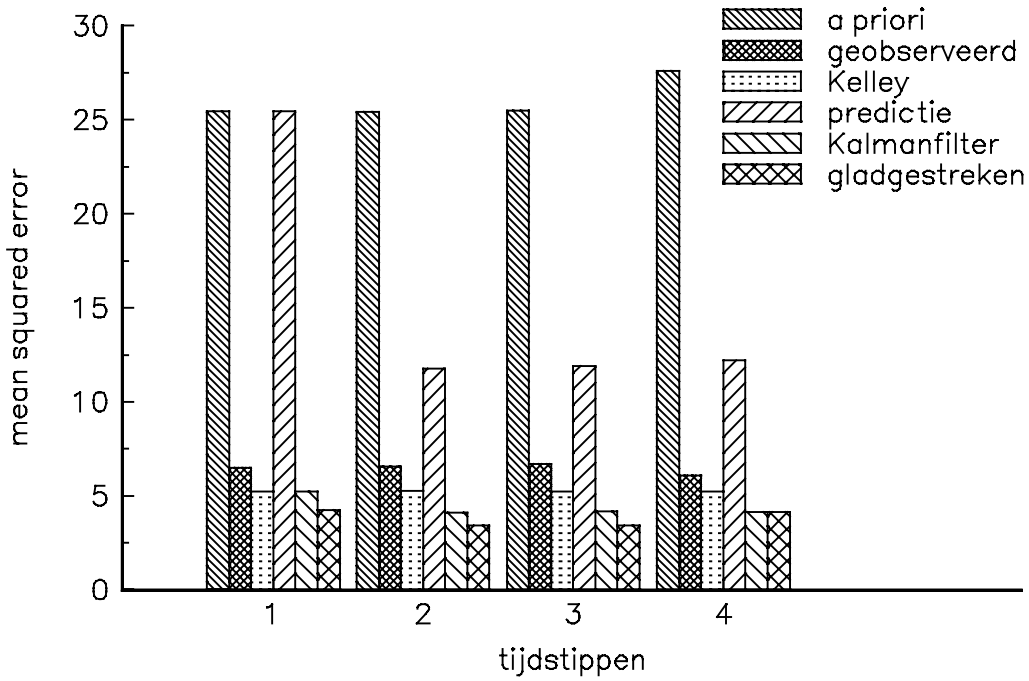
in dit verband: het gemiddelde absolute verschil en het gemiddelde gekwadrateerde verschil (’Mean Squared Errors’). In figuur 10.7 is voor elk tijdstip het gemiddelde absolute verschil tussen ware score en schatting voor de diverse schatters weergegeven, en in figuur 10.8 het gemiddelde gekwadrateerde verschil.

De conclusies aangaande de rangorde van de schatters is niet anders dan bij de bespreking van de schattingsfoutvarianties. Dit is niet zo verwonderlijk als men zich realiseert dat voor grote steekproeven de MSE gelijk zal zijn aan de schattingsfoutvariantie. Bovendien hebben absolute verschillen en gekwadrateerde verschillen een hoop gemeen.

De statisticus komt tot de volgende conclusies aangaande de analyses van de psychometrici. Als men kiest voor momentopnamen, dat is de statische benadering, dan is de Kelley-schatter aan te bevelen. Kiest men een dynamische aanpak terwijl men

bovendien over de data van alle tijdstippen beschikt, dan is de gladgestreken Kalmanfilterschatter de aangewezen keus. Wil men echter tussentijds al over schattingen kunnen beschikken, de meest voorkomende situatie, dan is de Kalmanfilterschatting te prefereren. Heeft men longitudinale data, kies dan ook voor een dynamische aanpak. De winst die een dynamische benadering oplevert, kan erg groot zijn.

Knikker vindt de resultaten redelijk. Toch constateert hij dat de psychometrici er soms behoorlijk naast zitten. Afhankelijk van de gekozen schatter zitten zij er gemiddeld gezien ongeveer 1.5 tot 4 punten naast op de knikkervaardigheidsschaal. Ook verbaast het Knikker, dat de schattingsfoutvarianties van de diverse schatters, hoewel van verschillende grootte, voor elke leerling gelijk zijn. Knikker verwachtte namelijk dat het vaardigheidsniveau van sommige leerlingen nauwkeuriger geschat zou kunnen worden dan dat van andere leerlingen.



Figuur 10.8
Histogram 'Mean Squared Errors' (MSE) voor de vier tijdstippen

Tenslotte vraagt Knikker zich af of de resultaten anders geweest zou zijn als hij niet alle informatie ter beschikking had gesteld. Hij had de psychometrici bijvoorbeeld alleen de geobserveerde toetsscores kunnen verschaffen en niet de informatie over de populatie. Aangaande dit laatste punt kunnen de psychometrici Knikker gerust stellen. Onder zekere assumpties en restricties is het mogelijk de gegevens van de populatie te

achterhalen. Een methode om de populatieparameters te schatten staat beschreven in de volgende paragraaf.

10.2.5 Schattingen van structurele parameters

In het voorbeeld van de knikkervaardigheid was het uitgangspunt dat alle parameters behalve de ware scores bekend waren. In de praktijk zal dat niet zo zijn en zullen de parameters uit de observaties geschat moeten worden. Dit is mogelijk door de individuele tijdreeksen te beschouwen als replicaties van een onderliggende tijdreeks op populatieniveau. Hoe het een en ander zijn beslag krijgt, kan het beste geïllustreerd worden aan de hand van het zogenaamde simplexmodel. Het simplexmodel is een model met een bepaalde covariantiestructuur die vaak van toepassing is op longitudinale data. Hierbij wordt dezelfde variabele bij dezelfde individuen op verschillende tijdstippen gemeten, of in een situatie waarbij de variabelen niet geordend zijn in de tijd maar bijvoorbeeld naar toenemende complexiteit. Een voorbeeld van laatstgenoemde situatie kan men vinden bij Guttman (1954) voor spreekvaardigheid. De typische structuur van simplexmodellen, in de correlatiematrix nemen de correlaties van de diagonaal af gezien af, wordt gegenereerd door een onderliggend eerste-orde-autoregressief proces. Voor een uitvoerige introductie van deze modellen verwijzen we naar Guttman (1954), Jöreskog (1970) en Imbos (1989).

De schattings- en identificatieproblematiek van de parameters van het simplexmodel bespreken we in het kort. Omwille van de eenvoud beperken we ons tot gestandaardiseerde metingen op vier tijdstippen, y_t $(t = 1, 2, 3, 4)$. Het meetmodel op de vier tijdstippen kan wederom beschreven worden met de meetvergelijking uit de klassieke testtheorie

$$y_t = \eta_t + \varepsilon_t \qquad t = 1, 2, 3, 4.$$

Het groeimodel heeft een autoregressieve structuur die met de volgende drie vergelijkingen beschreven kan worden

$$\eta_t = \beta_t \eta_{t-1} + \zeta_t \qquad t = 2, 3, 4. \tag{10.19}$$

In (10.19) kan β_t geïnterpreteerd worden als de regressiecoëfficiënt van η_t op η_{t-1} en ζ_t als de meetfout met bijbehorende variantie Ψ_t (het onverklaarde deel van de variantie van η_t). Merk op dat de latente variabelen η_t en de geobserveerde variabelen y_t op dezelfde schaal liggen, zodat bij gestandaardiseerde metingen geldt dat, voor alle t,

$\mathcal{E}(\eta_t) = \mathcal{E}(y_t) = 0$. De correlatiematrix Σ_y van de geobserveerde variabelen heeft de volgende vorm:

$$\sum_y = \begin{pmatrix} \sigma^2_{\eta_1} + \sigma^2_{\varepsilon_1} & & & \\ \beta_2 \sigma^2_{\eta_1} & \sigma^2_{\eta_2} + \sigma^2_{\varepsilon_2} & & \\ \beta_2 \beta_3 \sigma^2_{\eta_1} & \beta_3 \sigma^2_{\eta_2} & \sigma^2_{\eta_3} + \sigma^2_{\varepsilon_3} & \\ \beta_2 \beta_3 \beta_4 \sigma^2_{\eta_1} & \beta_3 \beta_4 \sigma^2_{\eta_2} & \beta_4 \sigma^2_{\eta_3} & \sigma^2_{\eta_4} + \sigma^2_{\varepsilon_4} \end{pmatrix},$$

waarbij $\sigma^2_{\eta_t} = \beta^2_t \sigma^2_{\eta_{t-1}} + \Psi_t \quad (t = 2, 3, 4)$. Het blijkt dat niet alle parameters geïdentificeerd zijn (Jöreskog en Sörbom, 1989). Het kan aangetoond worden dat er identificatieproblemen zijn bij de verzamelingen parameters $\{\beta_2, \sigma^2_{\eta_1}, \sigma^2_{\varepsilon_1}\}$ en $\{\sigma^2_{\varepsilon_4}, \sigma^2_{\eta_4}\}$. Hoe dat precies in zijn werk gaat, is hier niet van belang. In het geval dat de metingen op dezelfde schaal zijn uitgevoerd, is de meest natuurlijke en gangbare manier om deze onbepaaldheden te elimineren door het introduceren van de restricties $\sigma^2_{\varepsilon_1} = \sigma^2_{\varepsilon_2}$ en $\sigma^2_{\varepsilon_3} = \sigma^2_{\varepsilon_4}$. Bij de analyse van de correlatiematrix impliceert de eerste restrictie dat de betrouwbaarheden van de eerste twee toetsen gelijk zijn, de tweede restrictie impliceert dat de betrouwbaarheden van de laatste twee toetsen gelijk zijn. Het schatten van de parameters en de modeltoetsing kan plaatsvinden met behulp van standaardprogrammatuur voor lineaire structurele modellen zoals LISREL (Jöreskog & Sörbom, 1989) en EQS (Bentler, 1985). De waarde van het formuleren, schatten en toetsen van het model ligt voornamelijk in het feit van de beschikbaarheid van de programmatuur en de impliciete mogelijkheden om het model te toetsen. Daarnaast is er een zekere flexibiliteit om het model uit te breiden naar meer indicatoren voor een latente vaardigheid alsook het simultaan analyseren van verschillende latente vaardigheden.

Uiteraard zijn er naast de LISREL-benadering meer mogelijkheden om de onbekende structurele parameters te schatten. Een manier, die soelaas kan bieden in een situatie met ontbrekende waarnemingen staat beschreven in Shumway en Stoffer (1982).

10.3 Itemresponstheorie en groeiscores

In deze paragraaf werken we de bepaling van groeiscores nader uit, waarbij we een itemresponsmodel als meetmodel hanteren. Aan de hand van een concreet voorbeeld, de Schaal Vorderingen in Spellingvaardigheid (SVS) (Van den Bosch, Gillijns, Krom & Moelands, 1991), zullen we het traject voor de bepaling van groeiscores doorlopen. In tegenstelling tot bij het klassieke meetmodel, is bij itemresponsmodellen de relatie tussen de ware score of latente vaardigheid en het toetsresultaat of observaties niet lineair. Zoals zal blijken, is deze complicatie niet wezenlijk voor het bepalen van groeiscores.

10.3.1 Schaal Vorderingen in Spellingvaardigheid

Met de SVS kan men vaststellen hoe goed een leerling kan spellen in de aanvangsfase van het basisonderwijs, of anders gezegd: kan men spellingvaardigheid meten op het niveau van groep 3 en 4 van de basisschool. In deze paragraaf schetsen we summier op welke wijze dit instrument tot stand is gekomen. Bij spellen gaat het erom woorden om te zetten in schriftbeelden. Daarbij kan onderscheid gemaakt worden tussen klankzuivere en niet-klankzuivere woorden. De eerste fase van het spellingonderwijs richt zich op het correct leren schrijven van de klankzuivere woorden: je schrijft op wat je hoort. Al snel daarna komen de niet-klankzuivere woorden, de woorden waarbij er geen eenduidige relatie is tussen klank en letter, zoals bij bomen, trein, begin. Om die goed te schrijven moeten de leerlingen regels kunnen toepassen, of een woord naar analogie van een ander woord kunnen schrijven. De SVS beperkt zich tot eenvoudige klankzuivere en niet-klankzuivere woorden van een of twee lettergrepen (zie Van den Bosch e.a., 1991). De afname is klassikaal: de leerkracht leest een woord hardop voor en de leerlingen schrijven het op. De scoring is dichotoom: een correct geschreven woord levert 1 punt op en een fout geschreven woord 0 punten. In totaal bestaat het aantal opgaven van de SVS uit 173 woorden. Uit deze woorden zijn toetsen samengesteld, in totaal negen verschillende modules van elk ongeveer 20 items. In wisselende combinaties zijn deze modules op vier tijdstippen, medio en eind groep 3 en medio en eind groep 4, afgenomen bij dezelfde landelijke gestratificeerde steekproef (circa 1800 leerlingen). Het afnamedesign is al aan de orde geweest in hoofdstuk 8 en is daar weergegeven in figuur 8.5. Elke afnamegroep maakt op een tijdstip twee modules; bovendien is er voor gezorgd dat geen enkele leerling twee maal dezelfde module maakt. Dit resulteert in een design dat onvolledig is zowel op als over tijdstippen. In

equivaleerterminologie hebben we op tijdstippen met horizontaal equivaleren en over tijdstippen met verticaal equivaleren te maken. Zoals gesteld in hoofdstuk 8 komt het equivaleren neer op het calibreren van dit structurele onvolledige design met een itemresponsmodel. Bij de calibratie, dat is het schatten en toetsen van de modelparameters, is voor de SVS gebruik gemaakt van het 'One Parameter Logistic Model' (OPLM; Verhelst & Eggen, 1989). De basisvergelijking van dit model is gegeven door:

$$P(X_{vi} = x_{vi} \mid \theta_v, a_i, \beta_i) = \frac{\exp\left[a_i(\theta_v - \beta_i)x_{vi}\right]}{1 + \exp\left[a_i(\theta_v - \beta_i)\right]}.$$

In het geval van de SVS is in deze vergelijking X_{vi} een dichotome stochast bevattende de score van leerling v op item i met mogelijke waarden 0 (woord fout geschreven) en 1 (woord correct geschreven). Verder duidt θ_v de latente vaardigheid aan voor leerling v en zijn β_i en a_i respectievelijk de moeilijkheidsparameter en de discriminatie-index van item i. Voor een gedetailleerde beschrijving van dit model alsmede schattings- en modeltoets- procedures wordt verwezen naar de hoofdstukken 4 en 5. Met behulp van het OPLM bleek het mogelijk, een goede beschrijving van de SVS-data te geven. Dit resulteerde in discriminatie-indices en schattingen van de moeilijkheidsparameters voor de SVS-items. Het model werd expliciet getoetst op twee vormen van itemonzuiverheid (zie hoofdstuk 9), te weten: ethniciteit en tijdstip. Items bleken hetzelfde te functioneren voor allochtonen en autochtonen en op verschillende tijdstippen.

Nu we de items van de SVS op een schaal hebben afgebeeld, gaan we op zoek naar de nog onbekende latente vaardigheden voor de individuele leerlingen, θ_v. De itemparameters veronderstellen we in het vervolg bekend, geen onredelijke aanname gezien de omvang van de steekproef.

10.3.2 Het schatten van de latente vaardigheid

Nu de calibratie van de SVS-items met succes is afgerond, kunnen alle items in een itembank worden opgeslagen. Merk op dat er geen aanname gemaakt is over een populatieverdeling van de latente vaardigheid; de calibratie is immers uitgevoerd met CML en niet met MML (zie ook paragraaf 8.3.3). De volgende stap is het plaatsen van de individuele vaardigheden op dezelfde schaal als de items. Als vaardigheidsparameters en itemparameters op dezelfde schaal geplaatst zijn, is het meten van veranderingen in principe zonder meer mogelijk. Vaardigheden van leerlingen kunnen vergeleken worden op en over tijdstippen, en ook een terugkoppeling naar beheerste leerstof is

mogelijk door interpretatie van de itemparameters. Hoe de individuele vaardigheid geschat kan worden met een itemresponsmodel als meetmodel zullen we nu demonstreren. Wederom vergelijken we de statische en de dynamische aanpak.

Statische aanpak

Analoog aan paragraaf 10.2.2 bekijken we de tijdstippen afzonderlijk. Ook negeren we vooralsnog alle a priori kennis omtrent de populatie waartoe een leerling behoort. Op een tijdstip beschikken we voor een leerling v dus alleen over zijn toetsresultaat. In het geval dat we OPLM als meetmodel hanteren, is het toetsresultaat de som over de gemaakte items van de responsvariabele gewogen met de discriminatie-index: $s = \sum_i a_i x_{vi}$. Merk op dat het toets- resultaat s een voldoende statistiek is voor de vaardigheidsparameter θ. De vraag is nu of we de latente vaardigheid van een leerling op een tijdstip kunnen schatten uit de itemparameters en het toetsresultaat. Stel dat we de vaardigheid van een leerling opvatten als een onbekende constante, dat wil zeggen een statistische parameter die geschat moet worden. In het OPLM is het toetsresultaat een voldoende statistiek voor de vaardigheidsparameter. Een goede schatter voor de vaardigheidsparameter is de gewogen-grootste-aannemelijkheidsschatter (WML), geïntroduceerd door Warm (1989). In paragraaf 4.5 is deze schatter al besproken; hier volstaan we met het geven van de schattingsvergelijking, die wordt gegeven door het maximaliseren van de aannemelijkheidsfunctie gewogen met de toetsinformatie

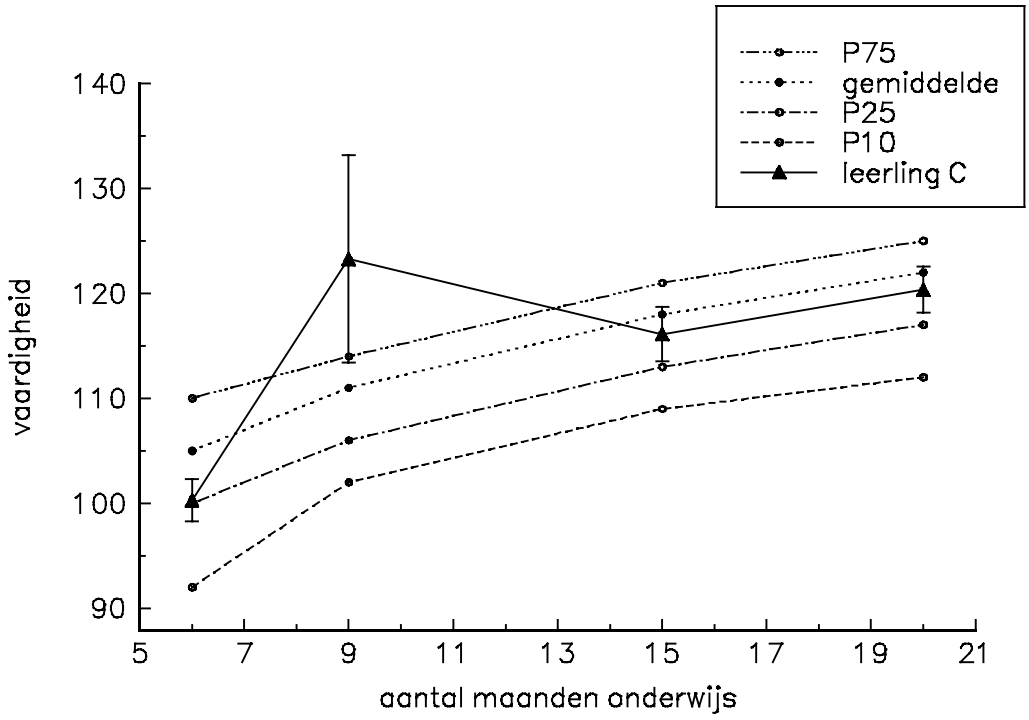
$$\max_{\theta} P(s \mid \theta) \sqrt{I(\theta)} .$$

De WML-schatter is onzes inziens de aangewezen schatter als we iemands vaardigheid opvatten als een onbekende constante. Deze schatter is immers nagenoeg zuiver op het individuele niveau en bestaat ook voor leerlingen die alles fout dan wel goed hebben, dit in tegenstelling tot de gewone grootste-aannemelijkheidsschatter. De WML-schatter voor de latente vaardigheid in een itemresponsmodel is het equivalent van de geobserveerde-score-schatter van de ware score in het klassieke meetmodel. In tegenstelling tot de geobserveerde-score-schatter uit het klassieke meetmodel is de WML-schatter een niet-lineaire transformatie van het toetsresultaat s. Uiteraard hoort bij de WML-schatter een schattingsfoutvariantie. De schattingsfoutvariantie van de geobserveerde-score-schatter in het klassieke meetmodel is gelijk aan de meetfoutvariantie en onafhankelijk van de ware score van een leerling, en is voor elke geobserveerde score even groot. Daarentegen is de schattingsfoutvariantie van de WML- schatter

afhankelijk van de latente vaardigheid en dus voor leerlingen met een ongelijk toetsresultaat verschillend.

Vanwege de eigenschap van zuiverheid van de WML-schatter is het mogelijk, populatie- karakteristieken te achterhalen als percentielen en gemiddelden. Deze populatiekarakteristieken kunnen dan vervolgens dienen als referentiegegevens voor individuele resultaten. Stel dat we voor de SVS referentiegegevens zoals gemiddelden en percentielen willen bepalen voor de Nederlandse populatie leerlingen per tijdstip, dan kan dat simpel door bijvoorbeeld de WML-schattingen in de steekproef te middelen, of bij het bepalen van percentielen de WML-schattingen in de steekproef te sorteren naar oplopende grootte en die waarden te kiezen die corresponderen met de percentages. Daar de steekproef in het voorbeeld van de SVS gestratificeerd was naar schoolgewicht (zie ook paragraaf 7.1), diende uiteraard een weging plaats te vinden naar de Nederlandse populatie. In figuur 10.9 zijn voor de Nederlandse populatie leerlingen per tijdstip het gemiddelde en de percentielen 10, 25 en 75 weergegeven. Tevens zijn in figuur 10.9 voor leerling C de WML-schatting op de vier momenten weergegeven.

Met behulp van de referentiegegevens kunnen we nu bepalen hoe goed een leerling het doet ten opzichte van de groep op de vier meetmomenten. Kijken we naar de WML-schattingen van leerling C, dan kunnen we constateren dat na zes maanden onderwijs de vaardigheid van deze leerling rond percentiel 25 ligt, na negen maanden onderwijs ver boven percentiel 75 en terugvalt onder het gemiddelde na vijftien en twintig maanden onderwijs. Rond de schat-tingen voor leerling C is een betrouwbaar-heidsinterval aangegeven, plus en min een standaardafwijking van de schattingsfout, de verticale lijntjes in figuur 10.9. De orde van grootte van de betrouwbaarheidsintervallen is ongeveer 5 punten op de schaal voor de SVS, met uitzondering voor tijdstip 2, dat is na 9 maanden onderwijs; daar omvat het interval circa



Figuur 10.9 Gemiddelden en percentielen (P10, P25 en P75) voor de Nederlandse populatie in groep 3 en 4 van de basisschool voor de SVS en de WML-schattingen voor leerling C

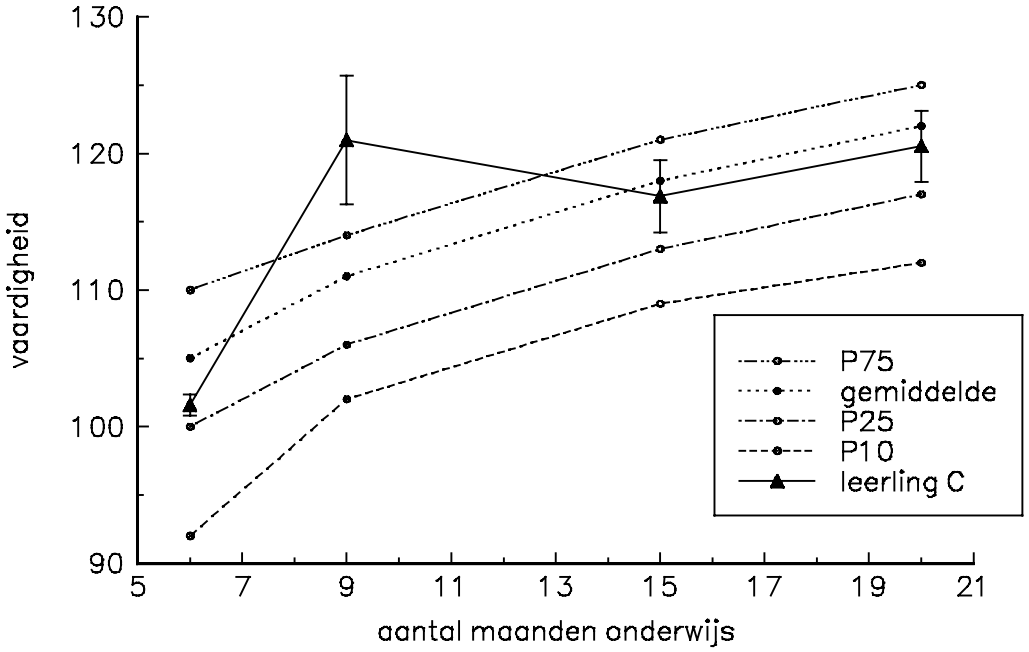
20 punten. Op tijdstip 2 hebben we de vaardigheid van leerling C dus zeer onnauwkeurig gemeten. Dit is problematisch als men resultaten wil interpreteren of conclusies verbinden aan de ontwikkeling van leerling C met betrekking tot spellingvaardigheid. In de praktijk van het onderwijs is het beeld als geschetst voor leerling C, eerder regel dan uitzondering. Deze fluctuaties van de vaardigheid in de tijd voor een leerling is voor het leeuwedeel te wijten aan de vaak zeer onbetrouwbare metingen.

In het kader van de itemresponstheorie zijn er diverse mogelijkheden om de nauwkeurigheid van de metingen te vergroten. Te denken valt aan vormen van adaptief toetsen. We komen hier straks op terug. Een andere mogelijkheid is de schatting van de latente vaardigheid van een leerling niet alleen te laten afhangen van zijn eigen toetsresultaat, maar ook van informatie over de groep waartoe deze leerling behoort. Merk de analogie met de Kelley schatter in paragraaf 10.2.2 op. Het equivalent van de Kelley-schatter uit het klassieke meetmodel in de itemresponstheorie is de 'expected a posteriori' of EAP-schatter. De EAP-schatter is al besproken in hoofdstuk 4; hier volstaan we alleen met de schattingsvergelijking:

$$\mathscr{E}(\theta \mid s) = \frac{\int \theta P(s \mid \theta)\, g(\theta)\, d\theta}{\int P(s \mid \theta)\, g(\theta)\, d(\theta)}, \qquad (10.20)$$

waarbij $g(\theta)$, de kansdichtheidsfunctie van θ is in de populatie, dus de populatie-informatie met betrekking tot θ.

Om de EAP-schatter uit te kunnen rekenen moeten we over populatie-informatie $g(\theta)$ beschikken. Daartoe zullen we $g(\theta)$ moeten specificeren. Gebruikelijk is, hiervoor de normale verdeling te kiezen. Gemiddelde en variantie van deze a priori verdeling zullen we moeten schatten. Schattingen kunnen we onder andere verkrijgen met behulp van de MML- methode, besproken in hoofdstuk 4, of door statistiek te bedrijven met de WML-schattingen (Verhelst & Kamphuis, 1989; Hoijtink & Boomsma, 1991). Hier volstaan we met het geven van schattingen van deze verdelingen op de vier tijdstippen. Deze zijn voor het gemiddelde respectievelijk 105.2, 111.3, 117.3 en 121 en voor de varianties respectievelijk 101.6, 53.6, 51.1 en 56.7. In wezen zijn dit de a priori schattingen uit paragraaf 10.2.2, waarbij men het



Figuur 10.10
EAP-schattingen voor leerling C

gemiddelde kan opvatten als schatter en de variantie als schattingsfoutvariantie. In figuur 10.10 zijn voor leerling C de EAP-schattingen en de betrouwbaarheidsintervallen (plus en min één standaardafwijkingen van de schattingsfout) weergegeven.

Men kan constateren dat op alle tijdstippen de WML-schattingen in de richting van het populatiegemiddelde zijn opgeschoven. De verschuiving is het grootst op tijdstip 2 waar de WML-schatting het meest onbetrouwbaar was. Ook kan geconstateerd worden dat

in dit geval de schattingsfout bij de EAP-schattingen kleiner is dan bij de WML-schattingen. Dit hoeft niet altijd het geval te zijn.

Resumerend kunnen we stellen dat bij de statische benadering van groeiscores de schatters uit de klassieke testtheorie equivalenten hebben in de itemresponstheorie.

Dynamische benadering

Ook de drie besproken schatters bij de dynamische benadering in paragraaf 10.2.3, de predictieve, Kalmanfilter- en gladgestreken Kalmanfilterschatters, hebben hun equivalenten in de itemresponstheorie. Merk op dat met betrekking tot het groeimodel, op populatieniveau geformuleerd, er niets verandert als we in plaats van de klassieke testtheorie de itemresponstheorie als meetmodel hanteren. Het groeimodel beschrijft immers niets anders dan de ontwikkeling van de latente vaardigheid in de tijd ongeacht de wijze waarop we die vaardigheid ook trachten te meten. Dit houdt in dat de predictieve schatter voor beide modellen dezelfde vorm heeft, alleen de schatting die we invullen in bijvoorbeeld (10.14) is anders en wordt nu bepaald door het gebruikte meetmodel. Uitgaande van hetzelfde autoregressieve groeimodel als besproken in paragraaf 10.2.3, kan de procedure voor het verkrijgen van de dynamische schatters in de volgende stappen uiteengelegd worden:

(1) Bepaal op het eerste tijdstip $\mathscr{E}(\theta_1 \mid s_1, \mu_{\theta_1}, \sigma^2_{\theta_1})$, dat is de EAP-schatter gegeven het toetsresultaat s_1 en de marginale verdeling van θ op tijdstip 1 met gemiddelde μ_{θ_1} en variantie $\sigma^2_{\theta_1}$, en bijbehorende schattingsfoutvariantie (Kalmanfilterschatter).

(2) Deze conditionele verwachting en schattingsfoutvariantie substitueren we in de pre- dictievergelijking 10.14. Nu beschikken we over de predictieve schatter en schattingsfoutvariantie op meetmoment 2.

(3) Bepaal de Kalmanfilterschatting op tijdstip 2, dat is de EAP-schatter gegeven toetsresultaat, s_2, en de predictieve schatter en schattingsfoutvariantie uit stap 2.

(4) Herhaal stap 2 en 3 tot alle meetmomenten verwerkt zijn.

(5) Bepaal met behulp van de nu beschikbare Kalmanfilterschattingen en schattings-foutvarianties de gladgestreken schattingen en bijbehorende schattingsfoutvarian-ties.

In de klassieke testtheorie kwam de combinatie van populatieinformatie en toetsresultaat in essentie neer op het combineren van twee onafhankelijke schatters, de geobserveerde-score-schatter en de predictieve schatter tot de Kelley-schatter. In de itemresponstheorie vervult de EAP-schatter de rol van de Kelley-schatter.

De vraag resteert hoe we de gemiddelden en de covariantiematrix van de latente vaardigheid op populatieniveau kunnen schatten. Het voert te ver hier op in te gaan; we volstaan met een verwijzing naar Kamphuis en Engelen (in voorbereiding). In het voorbeeld van de SVS is een autoregressief model van de eerste orde geschat voor de vier meetmomenten:

$$\theta_t = a_t + b_t \theta_{t-1} + \zeta_t \qquad t = 2, 3, 4,$$

waarbij t de tijdstipindex, a en b de regressiecoëfficiënten en ζ_t een storingsvariabele met verwachting 0 en variantie Ψ_t (onverklaarde variantie op een tijdstip t) is. Schattingen voor de parameters in deze vergelijkingen staan vermeld in tabel 10.4.

Gemiddeld groeit de populatie circa 6 punten tussen tijdstippen, uitgezonderd voor het laatste tijdstip. De voorspellingen van tijdstip naar tijdstip verklaren respectievelijk 62%, 70% en 81% van de variantie op de desbetreffende tijdstippen. Laten we eens zien wat de consequenties zijn als we dit groeimodel toepassen op leerling C. In figuur 10.11 zijn de Kalmanfilterschattingen voor leerling C weergegeven en in figuur 10.12 de gladgestreken Kalmanfilterschattingen. Als we kijken naar tijdstip 2, dan kunnen we constateren dat de Kalmanfilterschatter nog meer dan de EAP-schatter de schaalscore heeft verminderd, respectievelijk 116.31 en 120.98.

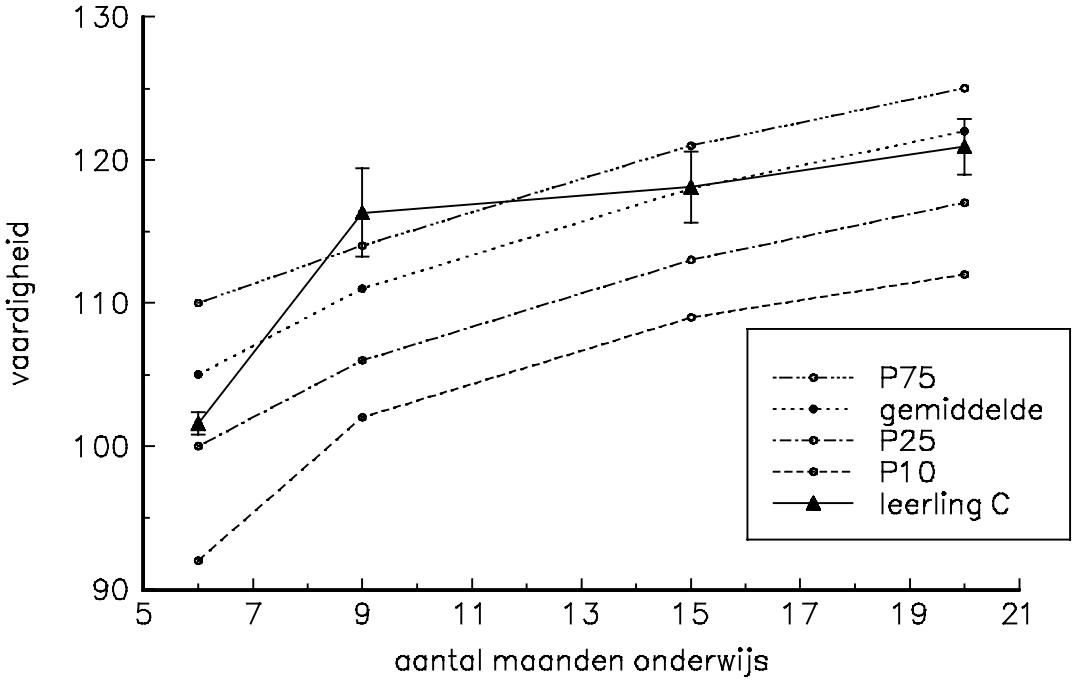
Tabel 10.4

Schattingen van de parameters van het SVS groeimodel met tussen haakjes het aantal maanden onderwijs

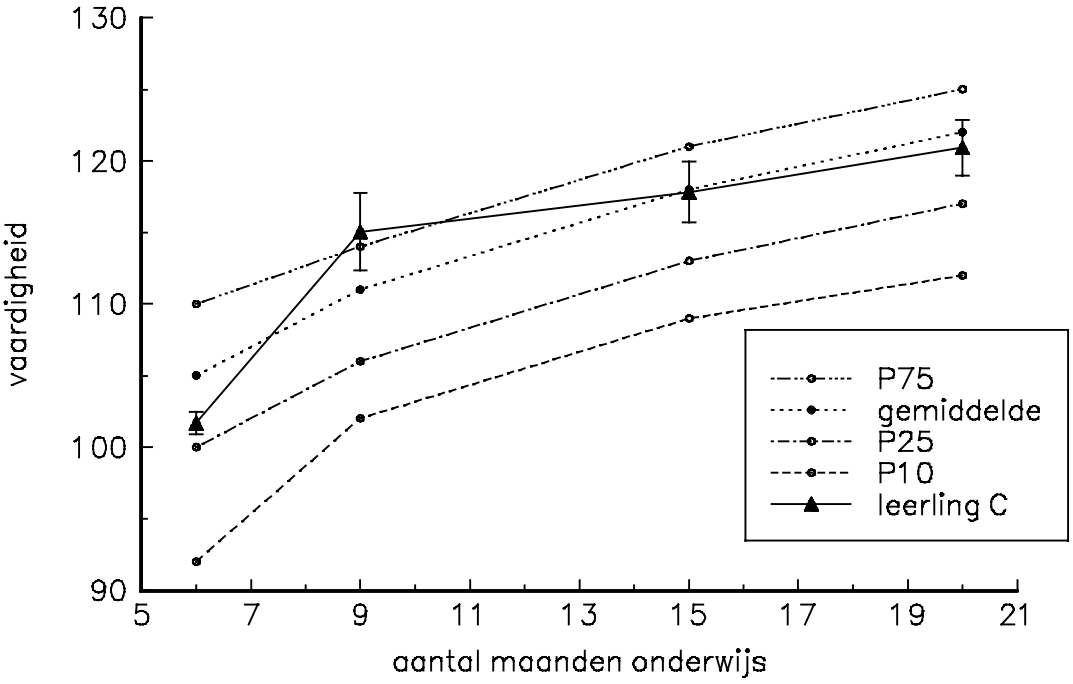
	tijdstip			
parame-ter	1(6)	2(9)	3(15)	4(20)
μ_θ	105.15	111.32	117.34	120.95
σ^2_θ	101.60	53.58	51.10	56.74
Ψ		20.18	15.52	10.53
a		51.02	26.62	9.38
b		.57	.81	.95

De predictieve schatting en schattingsfout, de a priori kennis op tijdstip 2, bedroeg 109.27 en 4.51 (niet weergegeven in figuur 10.11). Bij de EAP-schatter daarentegen was de a priori kennis gebaseerd op een gemiddelde 111.32 en een standaarddeviatie van 7.32. Ook constateren we weer dat toevoegen van informatie uit het groeimodel de schattingsfouten reduceert. De gladgestreken schatting op tijdstip 2 voor leerling C ligt in vergelijking met de Kalmanfilterschatting meer in lijn met de andere schattingen.

Ook constateren we weer dat de standaardschattingsfouten van de gladgestreken Kalmanfilterschattingen iets kleiner uitvallen dan die van de Kalmanfilterschattingen.



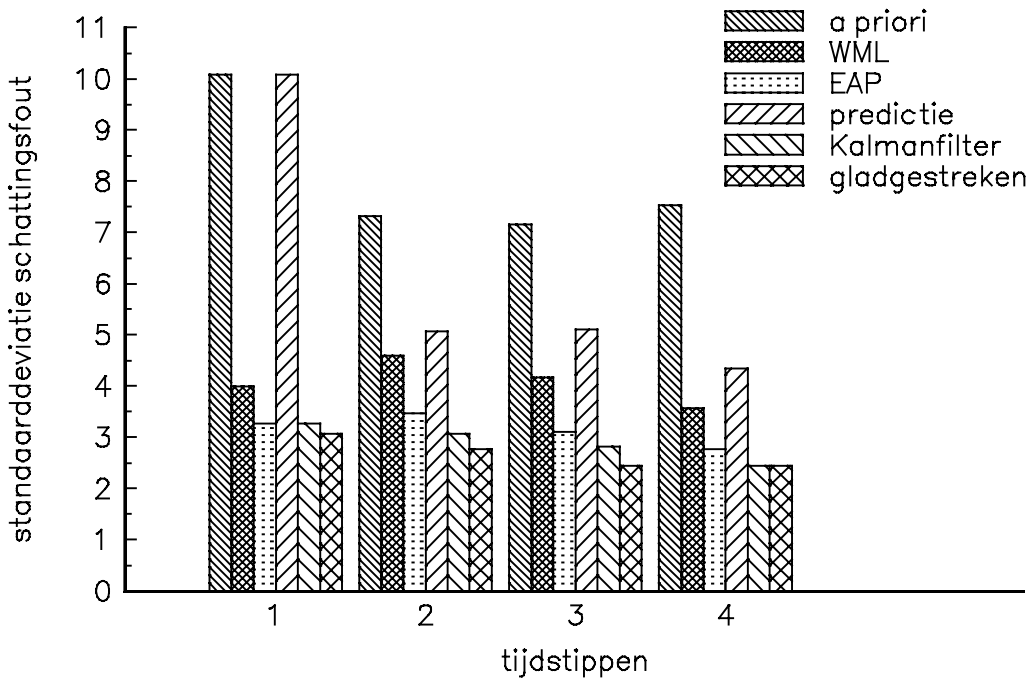
Figuur 10.11
Kalmanfilterschattingen voor leerling C



Figuur 10.12 Gladgestreken Kalmanfilterschattingen voor leerling C

Evaluatie statische en dynamische benadering

De conclusies die getrokken zijn in de evaluatie van de statische en dynamische benadering bij het bepalen van individuele vaardigheden in paragraaf 10.2.4 gelden natuurlijk onverkort in de situatie waarin een itemresponsmodel wordt gebruikt als meetmodel. In het voorbeeld van de SVS beschikken we echter niet over de ware vaardigheden van de individuele leerling zoals in het voorbeeld van knikkervaardigheid. Dus, om de in deze paragraaf besproken statische en dynamische schatters te evalueren kunnen we alleen terugvallen op de statistische eigenschappen van deze schatters. Daar alle besproken schatters wederom zuiver zijn in de populatie, beperken we ons ook deze keer tot een vergelijking van een maat voor de spreiding van de schattingsfout van de diverse schatters. In figuur 10.13 is de gemiddelde standaardafwijking van de schattingsfout voor de diverse schatters op de verschillende tijdstippen weergegeven. We vergelijken eerst de standaardafwijkingen van de schattingsfout van de cross-sectionele schatters. De EAP-schatter heeft op alle tijdstippen de kleinste standaardafwijking, gevolgd door de WML-schatter en de a priori schatter. Verder valt op dat de stan-



Figuur 10.13 Gemiddelde standaarddeviatie van de schattingsfout voor de diverse schatters op de vier tijdstippen voor de leerlingen in de SVS-steekproef (n = 1800)

daardafwijking van de a priori schatter op het eerste tijdstip groter is dan op de volgende tijdstippen (circa 10 versus circa 7). Aanvankelijke verschillen in spellingvaardigheid in de populatie worden blijkbaar door het effect van het onderwijs deels geneutraliseerd. Ook constateren we dat de standaardafwijking van de WML-schatter op het tweede tijdstip in vergelijking met de andere tijdstippen het grootst is. De toetsmodules die zijn afgenomen op het tweede tijdstip leverden de minste informatie op over de spellingvaardigheid. Met andere woorden: deze modules zijn niet op maat gesneden voor de populatie op dat tijdstip. Bezien we de dynamische schatters, dan is het beeld niet anders dan beschreven in paragraaf 10.2.4: de gladgestreken Kalmanfilterschattingen zijn het meest nauwkeurig, gevolgd door de Kalmanfilterschattingen en op afstand de predictieve schattingen.

Ook hier constateren we dat de dynamische schatters hun statische equivalenten overtreffen als het gaat om de meetnauwkeurigheid. De mate waarin, wordt bepaald door de precisie van de meetresultaten en de mate van nauwkeurigheid van de predicties.

10.4 Epiloog

In dit hoofdstuk is het meten van veranderingen en het bepalen groeiscores behandeld. De kern van het verhaal ligt besloten in de vraag: Hoe combineren we informatie uit twee bronnen, groei- en meetmodel, tot één vaardigheidsschatting? We zagen dat het mogelijk was om met behulp van een groeimodel iemands vaardigheid te voorspellen op een bepaald tijdstip. Bovendien konden we op dat tijdstip de actuele meting met behulp van een meetmodel omzetten in een schatting van de vaardigheid. Groei- en meetmodel leverden dus beiden een indicatie op over iemands vaardigheid, welke gecombineerd konden worden tot één schatting. Afhankelijk van het gekozen meet- en/of groeimodel en de keuze hoe men de vaardigheid beziet, als een onbekende parameter of als een toevalsvariabele, ziet de schatter er anders uit. Welke schatter men prefereert, is vaak een persoonlijke keuze. De meest informatieve schatter is de gladgestreken Kalmanfilterschatter. De minst informatieve schatter is in de klassieke testtheorie de geobserveerde score en in de itemresponstheorie de WML-schatter. De keuze voor de minst informatie schatter wordt vaak gemotiveerd door te stellen dat men geen a priori informatie wil meenemen in de schatting van de vaardigheid omwille van de eerlijkheid. Met andere woorden, de schatting van de vaardigheid mag alleen berusten op het meetresultaat en niet mede bepaald worden door eerdere meetresultaten of door de groep waartoe iemand behoort. Dit lijkt een nobel standpunt. Statistisch bezien is dit standpunt echter onrealistisch daar alle ingrediënten van deze schatters populatie afhankelijk zijn. In de klassieke testtheorie zijn de indexen voor de betrouwbaarheid zonder de definitie van een populatie betekenisloos. In de itemresponstheorie hebben de itemparameters altijd betrekking op een populatie, ook al bestaan er fraaie schattingsprocedures voor de itemparameters die steekproefonafhankelijk zijn. In de onderwijspraktijk levert dit standpunt dan ook problemen op: Hoe moeten we onbetrouwbare schattingen van de vaardigheid voor een leerling, die excessief fluctueren in de tijd, interpreteren? Dit excessief fluctueren van de vaardigheid in de tijd op het individuele niveau, door Rubin (1980) in een ander kader het "bouncing beta problem" genoemd, kan onderdrukt worden door populatie-informatie (groeimodel) te gebruiken bij de schattingen van iemands vaardigheid. Tevens reduceert dit deels de onbetrouwbaarheid van de schattingen. Een andere mogelijkheid om de onbetrouwbaarheid van de schattingen te reduceren, kan gevonden worden in de toepassing van betere meetprocedures. Met als uitgangspunt een schatter die informatie uit groei- en meetmodel combineert, bezien we welke mogelijkheden er zijn om de nauwkeurigheid van de schattingen te verhogen.

Eerst kijken we op het niveau van de populatie naar het groeimodel. In de voorbeelden die gebruikt zijn in dit hoofdstuk werd groei voor één vaardigheid gemoduleerd middels een simpel autoregressief model van de eerste orde waarbij één populatie werd

verondersteld. In de praktijk zal een dergelijke aanname waarschijnlijk een te grove benadering van de werkelijkheid zijn. Realistischer is het te veronderstellen dat er subpopulaties of groepen zijn te onderscheiden waarbij de groei verschillend verloopt. Denkbaar is ook dat we niet kunnen volstaan met een eerste orde autoregressief groeimodel, maar dat er andere modellen te vinden zijn die een betere beschrijving van de data opleveren. In de praktijk zullen we dus moeten onderzoeken, welk groeimodel we kiezen voor wie. Naast modelselectie dienen de modellen uiteraard naar behoren getoetst te worden. Om groepen op te sporen waarvoor groei verschillend verloopt zijn er een aantal procedures denkbaar. Een eerste procedure zou kunnen starten met een opdeling van de populatie naar achtergrondkenmerken. Men zou bijvoorbeeld na kunnen gaan of groei anders gemodelleerd dient te worden voor meisjes en jongens. Een andere mogelijkheid zou kunnen zijn een latente klasse analyse uit te voeren. Bij deze benadering vormen personen die hetzelfde groeipatroon hebben één (latente) klasse. De problemen bij deze laatste benadering zijn echter legio; vooralsnog is deze benadering dan ook toekomstmuziek.

De crux van het modelleren van groei is de voorspellingen zo nauwkeurig mogelijk te krijgen. Daarom is ook additionele informatie, bijvoorbeeld informatie met betrekking tot andere vaardigheden, bruikbaar om de predicties te verbeteren. Oud en Mommers (1988) gebruiken een longitudinaal verklaringsmodel voor de samenhang tussen de vaardigheden technisch lezen, begrijpend lezen en spellen. In dit model kan bij de predictie van spellingvaardigheid op een zeker tijdstip, informatie van de vaardigheden technisch lezen en begrijpend lezen worden verbeterd.

De mogelijkheden om de onbetrouwbaarheid van de schattingen van de vaardigheid te reduceren met behulp van het meetmodel zijn sterk afhankelijk van het gebruikte meetmodel. Merk ook op dat reductie van de schattingsfouten alleen kan plaatsvinden bij een nieuwe afname, reeds afgenomen toetsen kunnen niet meer bijgesteld worden. Laten we eens aannemen dat er aan de hand van een longitudinale gegevensverzameling een groeimodel voor een bepaalde populatie geschat hebben. Het is nu in principe mogelijk de meetprocedure voor toekomstige afnames te verfijnen op basis van de reeds beschikbare gegevens. Wel moeten we dan bedenken dat we bepaalde assumpties moeten maken, bijvoorbeeld dat de leerlingen bij de toekomstige afname beschouwd kunnen worden een steekproef uit oorspronkelijke populatie of dat de itemparameters in een itemresponsmodel constant blijven in de tijd. Zeker in een longitudinale context, waarbij de tijdspannes vaak groot zijn, is het wenselijk deze assumpties te controleren. Het is bijvoorbeeld denkbaar dat itemparameters als gevolg van onderwijskundige ontwikkelingen, door de loop der jaren veranderen. Stel dat er voor een leerling een vaardigheidsschatting beschikbaar is op een bepaald tijdstip. Met behulp van het

groeimodel is het mogelijk te voorspellen hoe vaardig de leerling op een volgend tijdstip zal zijn. Gegeven deze voorspelling, kunnen we dan voor deze leerling een toets op 'maat' kiezen, dat wil zeggen een toets kiezen die de meetfout minimaliseert. Hoe we toetsen op maat kunnen samenstellen wordt besproken in hoofdstuk 11. Ook kunnen predicties van de vaardigheid gebruikt worden als startwaarden in adaptieve toetsprocedures, dat is biedt opgaven aan met een moeilijkheid in de buurt van de lopende schatting van de vaardigheid. Merk op dat de mogelijkheden van toetsen op maat sterk bepaald zijn door het gebruikte meetmodel. Al met al bieden itemresponsmodellen in zijn algemeenheid meer mogelijkheden voor verfijnde toetsprocedures dan het klassieke meetmodel.

Het belang van de keuze van een geschikt meet- en groeimodel bij het meten van veranderingen kan niet genoeg benadrukt worden. Zowel het meetmodel als het groeimodel kunnen in belangrijke mate bijdragen aan de reductie van de onbetrouwbaarheid van de vaardigheidsschattingen voor individuele leerlingen. Als we de vaardigheid van de leerlingen in de tijd nauwkeurig kunnen bepalen, kunnen we ook het probleem van een verfijnd referentiekader (zie paragraaf 10.1.3) aanpakken. We kunnen dan de individuele groei nauwkeurig afzetten tegen relevante andere individuen, groepen en populaties maar ook tegen onderwijsinhoudelijke criteria. Maar dan moet het ook mogelijk zijn om ongewenste ontwikkelingen of problemen te signaleren, bijvoorbeeld achterstand. Tenslotte nog een laatste opmerking. De signalering van problemen alleen is niet voldoende; diagnostisering van problemen en de ontwikkeling van hulpprogramma's voor achterstanden verdienen de nodige zorg en aandacht. Hopelijk biedt het hier geschetste kader voor het meten van veranderingen, waarbij meet- en groeimodel gekoppeld zijn, voldoende aanknopingspunten voor de gerichte ontwikkeling van diagnose- en hulpmateriaal.

11

Het samenstellen van toetsen

Bij het samenstellen van toetsen kunnen we te maken krijgen met drie soorten eisen: psychometrische, inhoudelijke en praktische eisen. De psychometrische eisen zullen veelal betrekking hebben op de gewenste meetnauwkeurigheid van de samen te stellen toetsen. Met inhoudelijke eisen worden de vakinhoudelijke en onderwijskundige eisen bedoeld: de verdeling van de vragen over de leerstofcategorieën, de gewenste moeilijkheidsgraad van de toets en dergelijke. Ook relaties op itemniveau kunnen een rol spelen bij het samenstellen van toetsen. Als bijvoorbeeld het antwoord op item 4 een aanwijzing bevat voor de antwoorden op item 16 en item 400, dan kan de toetsconstructeur eisen dat als item 4 in de toets wordt opgenomen, item 16 en item 400 niet meer worden opgenomen. Onder praktische eisen worden die aspecten van toetsconstructie verstaan die psychometrische noch inhoudelijke betekenis hebben, maar bij het samenstellen van toetsen wel degelijk een rol spelen. Een voorbeeld is de tijd die voor het afnemen van een toets beschikbaar is. Aangezien die tijd niet onbeperkt is, zal men hiermee bij het samenstellen van een toets rekening moeten houden. Een ander voorbeeld betreft het budget dat beschikbaar is om een toets te kunnen afnemen. Een bepaald budget zou kunnen betekenen dat niet meer dan drie beoordelaars ingeschakeld kunnen worden.

In dit hoofdstuk laten we zien hoe met behulp van wiskundige modellen toetsen samengesteld kunnen worden die voldoen aan de psychometrische, inhoudelijke en praktische specificaties van toetsconstructeurs. De modellen zijn ontleend aan een tak van de wiskunde, aangeduid met operationele research of mathematische programmering, die als doel heeft het ontwikkelen van modellen ter ondersteuning van besluitvorming. De eerste paragraaf van dit hoofdstuk bevat een beknopte bespreking van mathematische programmering. De drie volgende paragrafen bevatten toepassingen van mathematische programmering binnen de itemresponstheorie, de klassieke testtheorie en de generaliseerbaarheidstheorie.

11.1 Mathematisch programmeren

Stel, iemand is op expeditie in Groenland. De bagage wordt vervoerd op een hondenslede waar nog genoeg ruimte over blijft om een paar extra dingen mee te nemen om onderweg in de handelspost te verkopen. De reiziger heeft nog een doos met tien literblikken ananas, een doos met twintig literblikken hondevoer en een jerrycan met twintig liter benzine. In de handelspost is men bereid tweehonderd Groenlandse kronen te betalen voor de ananas, honderd voor het hondevoer en honderd voor de benzine. De doos ananas weegt dertig kilo, het hondevoer veertig kilo en de benzine twintig kilo. Op de hondenslede is nog plaats voor veertig liter extra bagage. De honden mogen echter niet meer trekken dan zestig kilo. Het probleem van onze reiziger is nu, te beslissen welke dingen hij moet meenemen zodat hij de meeste opbrengst in de handelspost heeft. We zullen laten zien hoe modellen voor dit soort problemen geformuleerd worden binnen de mathematische programmering en hoe deze problemen vervolgens opgelost worden.

Het besluit om een bepaald produkt mee te nemen kunnen we voorstellen door een zogenaamde beslisvariabele. Deze variabele neemt de waarde 1 aan als het desbetreffende produkt wordt meegenomen en de waarde 0 als het produkt niet wordt meegenomen. Variabelen die alleen waarden 0 en 1 kunnen aannemen, worden binaire variabelen genoemd. Noemen we de beslisvariabele die betrekking heeft op het meenemen van de benzine x_1, het meenemen van de ananas x_2 en van het hondevoer x_3, dan kunnen we de opbrengst uitdrukken als $100\,x_1 + 200\,x_2 + 100\,x_3$. Deze functie wordt de doelfunctie genoemd. Het totale volume van de mee te nemen produkten wordt uitgedrukt als $20\,x_1 + 10\,x_2 + 20\,x_3$ en het totale gewicht als $20\,x_1 + 30\,x_2 + 40\,x_3$. Het doel van de reiziger is een zo hoog mogelijke opbrengst te realiseren, terwijl de beperkingen ten aanzien van volume en gewicht niet worden overschreden. Deze beperkingen worden de restricties genoemd. De verzameling van alle beslissingen die toegelaten zijn, dat wil zeggen beantwoorden aan de restricties, heet de oplossingsruimte. Het model voor het probleem van de reiziger kunnen we nu formuleren als:

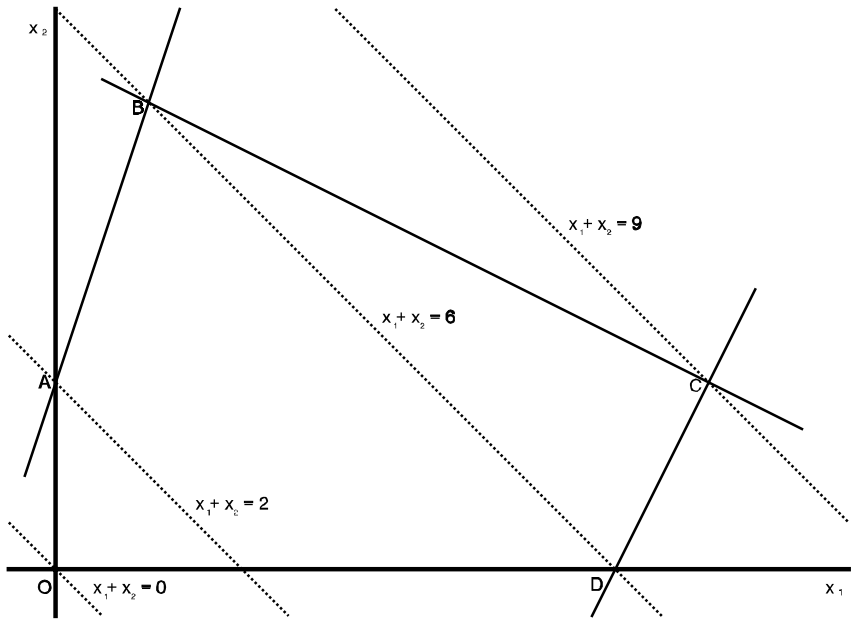
maximaliseer	$100\,x_1 + 200\,x_2 + 100\,x_3$	(opbrengst)
onder voorwaarde dat	$20\,x_1 + 10\,x_2 + 20\,x_3 \leq 40$	(volume)
	$20\,x_1 + 30\,x_2 + 40\,x_3 \leq 60$	(gewicht)
	$x_1,\ x_2,\ x_3\ \varepsilon\ \{0,1\}$.	(binaire variabelen)

Modellen waarvan de doelfunctie en alle restricties lineair zijn en alle beslisvariabelen continu, noemen we lineaire programmeringsmodellen. Wanneer de beslisvariabelen geen continue maar binaire variabelen zijn, zoals in ons reizigersprobleem, dan spreken we van binaire programmeringsmodellen.

Een populaire oplosmethode voor lineaire programmeringsmodellen is de simplexmethode. Om een grafische illustratie van de methode mogelijk te maken, nemen we een voorbeeld met twee variabelen. Het model voor het voorbeeld luidt:

maximaliseer $x_1 + x_2$ (11.1)

onder voorwaarde dat
$2x_1 - x_2 \leq 12$ (11.2)
$x_1 + 2x_2 \leq 11$ (11.3)
$-3x_1 + x_2 \leq 2$ (11.4)
$x_1, x_2 \geq 0.$ (11.5)



Figuur 11.1
Voorbeeld van de simplexmethode

De oplossingsruimte wordt hier gegeven door ongelijkheden (11.2)-(11.5) en wordt voorgesteld door de veelhoek OABCD in figuur 11.1. Zo wordt restrictie (11.2) weergegeven door het gebied links van de lijn CD, restrictie (11.3) door het gebied onder de lijn BC, restrictie (11.4) door het gebied rechts van de lijn AB en restrictie (11.5) door het gebied rechtsboven het assenkruis in figuur 11.1. De hoekpunten van

de oplossingsruimte (hier O, A, B, C en D) worden ook wel extreme punten genoemd. De oplossing die correspondeert met een extreem punt wordt een basisoplossing genoemd. Lineaire programmeringsproblemen hebben de eigenschap dat er altijd een optimale oplossing kan worden gevonden in de groep van basisoplossingen. Van deze eigenschap wordt door de simplexmethode gebruik gemaakt door op een systematische manier de groep van basisoplossingen af te zoeken. In ieder extreem punt zijn slechts twee restricties actief, dat wil zeggen dat aan twee ongelijkheden met een strikte gelijkheid wordt voldaan. Uitgaande van een extreem punt zoekt de simplexmethode steeds een naburig extreem punt op met een hogere doelfunctiewaarde. Dit gebeurt in de grafiek in figuur 11.1 door de doelfunctie (11.1) evenwijdig aan zichzelf naar rechts te verschuiven. In figuur 11.1 start de simplex in punt O. Hier zijn de beide niet-negativiteitseisen (11.5) actief. Punt O heeft twee naburige extreme punten: A en D. Zij hebben beide een hogere doelfunctiewaarde dan punt O. De simplex kiest het gunstigste naburige extreme punt, en verwisselt daarmee steeds één actieve restrictie door een andere. De simplex gaat steeds door naar een naburig punt totdat punt C, het optimum, bereikt wordt. De simplexmethode stopt zodra een extreem punt is gevonden met alleen naburige extreme punten die een lagere doelfunctiewaarde hebben. Zolang de simplexmethode van ieder extreem punt alleen naar een hoger gelegen extreem punt kan gaan, zorgt het feit dat een oplossingsruimte slechts een eindig aantal extreme punten heeft ervoor dat het optimum ook daadwerkelijk wordt bereikt.

Problemen in de praktijk zijn vaak complexer dan het probleem in dit voorbeeld, maar de simplexmethode zoekt nog steeds op ongeveer dezelfde manier de basisoplossingen af. Uitbreiding naar meer dan twee beslisvariabelen en daarmee samenhangend uitbreiding naar meer dan twee dimensies is niet eenvoudig in te zien. Er zijn nu geen twee actieve restricties, maar evenveel als er dimensies zijn. Tevens neemt het aantal basisoplossingen sterk toe bij toenemende dimensionaliteit. In bijvoorbeeld Dirickx, Baas en Dorhout (1987) vindt men een uitgebreide beschrijving van de simplex voor problemen met meer variabelen, alsmede de andere technieken die in dit hoofdstuk aan de orde komen.

Branch-and-bound methode

De oplosmethode voor binaire programmeringsmodellen is eveneens gebaseerd op de simplexmethode. De geheeltalligheidseisen ($x_j \in \{0,1\}$) worden gerelaxeerd, dat wil zeggen dat ze vervangen worden door de restricties $0 \leq x_j \leq 1$. Het zo ontstane continue probleem wordt vervolgens opgelost met behulp van de simplexmethode. Is

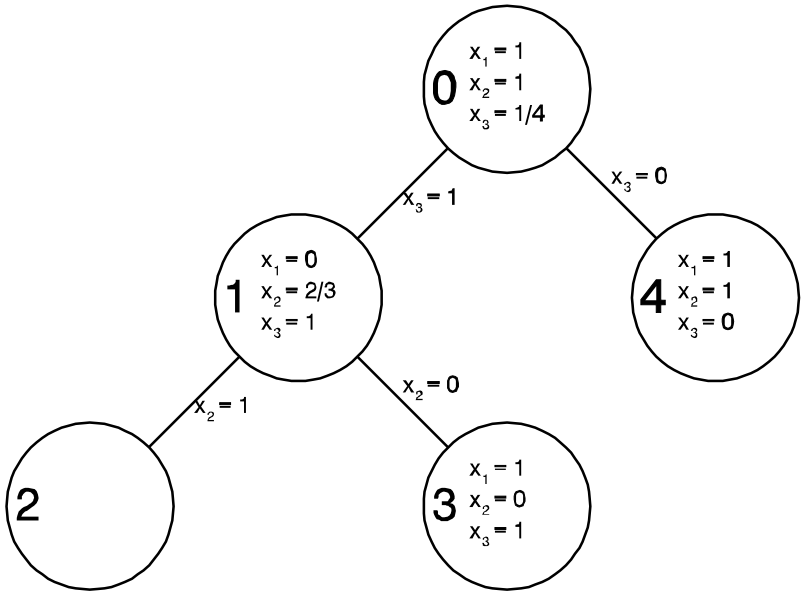
de optimale oplossing geheeltallig, dan is de optimale oplossing voor het continue probleem tevens de oplossing voor het binaire probleem. In het algemeen is de gevonden oplossing niet geheeltallig. De optimale oplossing van het continue probleem is nu niet meer een toegelaten oplossing voor het binaire probleem, maar het geeft wel een bovengrens voor de optimale doelfunctiewaarde voor het geheeltallige probleem. De extra geheeltalligheidseis legt een beperking op waardoor geen enkele geheeltallige oplossing een doelfunctiewaarde kan hebben die beter is dan de reeds gevonden oplossing. Van dit gegeven wordt handig gebruik gemaakt in de zogenaamde branch-and-boundmethode. Wanneer geen geheeltallige oplossing gevonden wordt, wordt het probleem gesplitst in twee subproblemen (branching). Er wordt een beslisvariabele gekozen die in de continue oplossing een niet-gehele waarde heeft. Vervolgens worden aan de hand van deze variabele twee kleinere problemen opgelost. Eén waarbij de beslisvariabele verplicht de waarde 1 krijgt en één waarbij de beslisvariabele de waarde 0 krijgt. Deze problemen worden subproblemen of knopen genoemd. Voor beide subproblemen wordt de procedure herhaald. Is er binnen een subprobleem nog geen geheeltallige oplossing gevonden, dan wordt er weer een variabele gekozen waarop de knoop wordt vertakt. Men gaat net zo lang door met vertakken tot er of een geheeltallige oplossing gevonden is of dat de gerelaxeerde oplossing van het beschouwde subprobleem een lagere doelfunctiewaarde heeft dan een eerder gevonden geheeltallige oplossing (bound). Wordt er een geheeltallige oplossing gevonden die beter is dan de tot dan toe beste oplossing, dan wordt deze oplossing vastgehouden als kandidaat voor de optimale oplossing. Is de optimale doelfunctiewaarde van het beschouwde subprobleem lager dan de kandidaatoplossing, dan heeft verder vertakken geen zin meer. De gevonden oplossing is immers een bovengrens voor de oplossing van alle subproblemen van het beschouwde probleem. Hiermee kan worden bewezen dat het subprobleem geen oplossingen kan geven die beter zijn dan de kandidaatoplossing. Ook kan het zijn dat de oplossingsruimte voor het subprobleem leeg is. Aangezien verdere subproblemen ook geen toegelaten oplossingen meer kunnen bevatten, heeft vertakken geen zin meer. De branch-and-boundmethode stopt als alle knopen beschouwd zijn. De gevonden kandidaat blijkt daadwerkelijk de optimale oplossing voor het oorspronkelijke probleem. De volgorde van vertakken is niet van wezenlijk belang voor de werking van de branch-and-boundmethode. In de praktijk wordt eerst de knoop waaraan men werkt verder vertakt, en pas als alle subproblemen van deze knoop zijn onderzocht wordt de tweede knoop onderzocht. De branch-and-boundmethode lijkt weliswaar omslachtig, maar als er een oplossing bestaat voor een probleem dan vindt de branch-and-bound altijd de optimale oplossing.

De branch-and-boundmethode zullen we toelichten aan de hand van het model voor het reizigersprobleem:

maximaliseer $100\,x_1 + 200\,x_2 + 100\,x_3$,

onder voorwaarde dat $20\,x_1 + 10\,x_2 + 20\,x_3 \leq 40$,
$20\,x_1 + 30\,x_2 + 40\,x_3 \leq 60$,
$x_1, x_2, x_3 \in \{0,1\}$. (geheeltalligheidseis)

De branch-and-boundmethode begint met de geheeltalligheidseis te vervangen door $0 \leq x_1, x_2, x_3 \leq 1$. Dit probleem duiden we aan met **0**. De simplex geeft voor **0** als optimum $x_1 = 1$, $x_2 = 1$, $x_3 = ¼$, met als opbrengst een bedrag van 325 kronen. Dit is geen geheeltallige oplossing en dus moet er worden gesplitst. In figuur 11.2 wordt in een zogenaamde beslisboom weergegeven hoe de problemen worden gesplitst en welke oplossing zij hebben.



Figuur 11.2
De beslisboom van de branch-and-bound procedure voor het reizigersprobleem

Eerst wordt subprobleem **1**, met als substitutie $x_3 = 1$, opgelost:

maximaliseer $100\,x_1 + 200\,x_2 + 100$,

onder voorwaarde dat $20\,x_1 + 10\,x_2 \leq 20,$
$20\,x_1 + 30\,x_2 \leq 20,$
$0 \leq x_1, x_2 \leq 1.$

Voor dit subprobleem wordt het optimum bereikt bij $x_1 = 0, x_2 = \frac{2}{3}, x_3 = 1$, met als opbrengst een bedrag van 233 kronen. Aangezien er weer geen geheeltallig optimum is gevonden, wordt er weer gesplitst. Let wel dat de subproblemen van **1** opgelost worden voordat er een nog openstaand probleem, namelijk probleem **4**, opgelost wordt. Het nieuwe subprobleem, probleem **2** genoemd, en ontstaan na substitutie van $x_2 = 1$, luidt:

maximaliseer $100\,x_1 + 300,$

onder voorwaarde dat $20\,x_1 \leq 10,$
$20\,x_1 \leq -10,$
$0 \leq x_1 \leq 1.$

Dit probleem heeft echter geen toegelaten oplossingen. Er wordt nu niet verder gegaan met splitsen, maar wordt het eerstvolgende nog openstaande probleem beschouwd. Dit is het subprobleem van **1**, probleem **3** genoemd, ontstaan na substitutie van $x_2 = 0$ en dit probleem luidt:

maximaliseer $100\,x_1 + 100,$

onder voorwaarde dat $20\,x_1 \leq 20,$
$20\,x_1 \leq 20,$
$0 \leq x_1 \leq 1..$

Nu wordt er wel een geheeltallig optimum bereikt bij $x_1 = 1, x_2 = 0, x_3 = 1$, met als opbrengst 200 kronen. Dit is de opbrengst die de reiziger krijgt als hij benzine en hondevoer meeneemt. We noemen deze oplossing nu de kandidaatoplossing, gaan niet verder met splitsen maar beschouwen het eerstvolgende nog openstaande probleem **4**. Merk op dat voor ieder volgend subprobleem de optimale doelfunctiewaarde is gedaald. Het nu nog openstaande probleem is het subprobleem van **0**, probleem **4**, ontstaan door substitutie van $x_3 = 0$, dat luidt:

maximaliseer $100\,x_1 + 200\,x_2,$

onder voorwaarde dat $20\,x_1 + 10\,x_2 \leq 40,$
$20\,x_1 + 30\,x_2 \leq 60,$
$0 \leq x_1, x_2 \leq 1.$

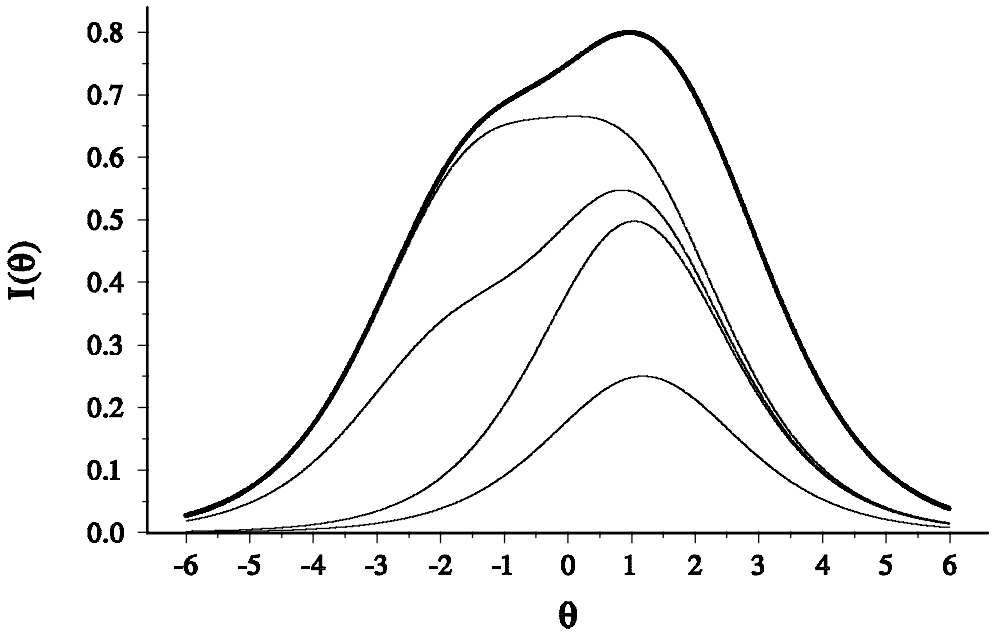
Hier wordt het optimum bereikt bij $x_1 = 1, x_2 = 1, x_3 = 0$, wat betekent dat de reiziger benzine en ananas moet meenemen, met als opbrengst een bedrag van 300 kronen. Er wordt dus weer een geheeltallig optimum gevonden. De opbrengst is nu echter hoger dan de opbrengst van de kandidaatoplossing, zodat de kandidaatoplossing wordt vervangen door de nu gevonden oplossing. Aangezien er geen openstaande subproblemen meer zijn is dit tevens de optimale oplossing voor het oorspronkelijke probleem.

11.2 Het samenstellen van toetsen in de itemresponstheorie

In de inleiding merkten we op dat psychometrische eisen betrekking hebben op de meetnauwkeurigheid van de toets. Binnen de itemresponstheorie worden voor het specificeren van de meetnauwkeurigheid continue functies gebruikt. De belangrijkste zijn de iteminformatie- en toetsinformatiefunctie. Zo is de standaarddeviatie van de grootste aannemelijkheidsschatter van de vaardigheid θ een functie van θ en gelijk aan $\mathrm{SE}(\hat{\theta}) = I(\theta)^{-½}$, waarbij $I(\theta)$ de toetsinformatiefunctie in het punt θ is (zie paragraaf 4.5.1). De informatie van een toets met lengte k is gelijk aan de som van de iteminformaties en wordt gegeven door

$$I(\theta) = \sum_{i=1}^{k} I_i(\theta).$$

Voor het Raschmodel is de iteminformatie gegeven door $I_i(\theta) = e^{(\theta-\beta_i)} \big/ \{1+e^{(\theta-\beta_i)}\}^2$, en deze functie is maximaal als $\theta = \beta_i$. Belangrijk voor toetsconstructie is het feit dat deze functies lokale meetnauwkeurigheid aangeven, dat wil zeggen dat de informatie afhankelijk is van het vaardigheidsniveau. Items die niet te moeilijk en niet te gemakkelijk zijn geven een hogere meetnauwkeurigheid dan zeer moeilijke en zeer gemakkelijke items. Figuur 11.3 laat zien hoe de toetsinformatie toeneemt wanneer we items aan een toets toevoegen. Telkens stijgt de toetsinformatiefunctie met de iteminformatiefunctie van het toegevoegde item.



Figuur 11.3 Toetsinformatiefunctie bij toenemende toetslengte

In Birnbaum (1968) en Lord (1980) vindt men een beschrijving van een trial-and-error heuristiek die van deze eigenschap gebruik maakt om toetsen met een bepaalde doelfunctie te construeren: de gewenste toetsinformatiefunctie, die afhankelijk is van het toetsdoel, wordt één voor één opgebouwd met de informatiefunctie van de gekozen items. Een belangrijke overweging is dat men vaak slechts in een beperkt gebied van de vaardigheidsschaal geïnteresseerd is, bijvoorbeeld in het cesuurgebied bij zakslaagbeslissingen. Men kan dan eisen stellen aan de meetnauwkeurigheid in het cesuurpunt en op twee punten daar net iets onder en boven. Men legt dan de toetsinformatiefunctie vast op een aantal punten maar blijft toch gebruik maken van het gegeven dat de toetsinformatie in ieder punt op de vaardigheidsschaal de som is van de iteminformaties. In het algemeen is het zo, dat continue functies voor bepaalde doeleinden goed gerepresenteerd kunnen worden door functies die uitsluitend zijn gedefinieerd op een aantal met zorg gekozen discrete punten.

Samenvattend: zeer belangrijk voor het probleem van het samenstellen van toetsen binnen de itemresponstheorie zijn de noties dat we voor een aantal punten op de vaardigheidsschaal de toetsinformatie specificeren en dat in elk punt de iteminformaties gesommeerd kunnen worden tot toetsinformatie. Op deze overwegingen is het gebruik van mathematische programmering bij toetsconstructie binnen itemresponstheorie gebaseerd. Al naar gelang de omstandigheden kan men eisen voor de toets met betrekking tot toetsinformatie formuleren als doel of als restrictie. Van beide zullen

later voorbeelden gegeven worden, zie ook Theunissen (1985, 1986), Van der Linden en Boekkooi-Timminga (1989).

11.2.1 Lineaire programmeringsproblemen

Voor de psychometrische en praktische eisen geldt dat ze als doelfunctie of als restrictie geformuleerd kunnen worden. Voor de inhoudelijke eisen geldt dat zij normaliter als restrictie geformuleerd worden. In de doelfunctie formuleert de toetsconstructeur wat er moet worden geoptimaliseerd, waarbij zowel minimaliserings- als maximaliseringsproblemen voor kunnen komen. Zoals we hebben gezien in paragraaf 11.1 zijn zowel doelfuncties als restricties te formuleren als eenvoudige lineaire expressies, waarbij men zich moet blijven realiseren dat de items in de expressies gerepresenteerd worden door binaire beslisvariabelen. Lineaire programmeringsmodellen worden algemeen geformuleerd als:

maximaliseer $$\sum_{i=1}^{K} c_i x_i,$$

onder voorwaarde dat $$\sum_{i=1}^{K} A_{ji} x_i \leq b_j, \quad j = 1,...,M$$

$$x_i \geq 0, \qquad i = 1,...,K.$$

Hierin zijn A_{ji}, b_j en c_i constanten, K het totaal aantal items in de itembank en M het aantal restricties.

We concentreren ons voorlopig op de doelfunctie. De variabelen x_i kunnen de waarden 1 en 0 aannemen. Ongeacht de betekenis van c_i is het duidelijk dat als $x_i = 0$, de daarbij behorende waarde van c_i niet zal bijdragen aan de waarde van de doelfunctie. De doelfunctie betreft een maximalisering: dat wil zeggen dat we proberen zoveel mogelijk van 'iets' te krijgen en dat 'iets' moet gunstig zijn in de ogen van de toetsconstructeur. Stel nu dat c_i de iteminformatie van item i is op een bepaald vaardigheidspunt, dan zegt bovenstaande doelfunctie niets anders als: 'maak een toets met een zo hoog mogelijke toetsinformatie (som van iteminformaties)'. Uiteraard dienen restricties toegevoegd te worden aan deze doelfunctie omdat anders alle beschikbare items in de toets zouden worden opgenomen. Stel nu dat de doelfunctie als volgt geformuleerd was:

minimaliseer $\sum_{i=1}^{K} c_i x_i$.

Ook hier nemen de x_i de waarden 1 en 0 aan, aangevend of item *i* al dan niet in de toets komt. Stel, dat de constructeur een bepaald doel voor ogen staat en we geven in deze doelfunctie aan alle c_i de waarde 1, dan houdt bovengenoemde doelfunctie niets meer in dan 'probeer aan alle (nog verderop te formuleren) voorwaarden te voldoen met een zo klein mogelijk aantal items', ofwel maak een toets van minimale omvang die nog aan eventuele andere voorwaarden beantwoordt. Een ander voorbeeld: stel dat -om herkenning te voor-komen- de toetsconstructeur vooral items in de toets op wil nemen die nog niet vaak gebruikt zijn en dat de gebruiksfrequentie voor alle items bekend is. We noemen de gebruiksfrequentie over een bepaalde periode voor item *i* hier dan c_i. Dus als item *i* bijvoorbeeld de afgelopen vier jaar twaalf maal gebruikt is, dan geldt $c_i = 12$. Omdat de doelfunctie als een minimalisering geformuleerd is, zullen items met een hoge bijbehorende waarde van c_i alleen in de toets worden opgenomen als er geen items in de bank beschikbaar zijn met een lagere waarde van c_i. Ook hier geldt uiteraard weer dat de gebruiksfrequentie van een item meetelt in de doelfunctie als de beslisvariabele x_i voor item *i* de waarde 1 aanneemt.

Behalve een doelfunctie zijn er ook randvoorwaarden in het probleem. Deze restricties zouden kunnen luiden:

$$\sum_{i=1}^{K} A_i x_i = b, \qquad \text{ofwel} \tag{11.6}$$

$$\sum_{i=1}^{K} A_i x_i \leq b, \qquad \text{ofwel} \tag{11.7}$$

$$\sum_{i=1}^{K} A_i x_i \geq b. \tag{11.8}$$

In de *b*'s in (11.6) - (11.8) kunnen de *b*'s van probleem tot probleem telkens iets anders betekenen en hoeven niet in dezelfde eenheden te zijn uitgedrukt. Hetzelfde geldt voor de A_i's. De flexibiliteit van deze eenvoudige modellen blijkt uit de zeer uiteenlopende interpretaties die men aan (11.6) - (11.8) kan toekennen. Zo kan men de eis dat de te maken toets van een specifieke lengte moet zijn formuleren als restrictie (11.6). Vaak is een vaste lengte de gewoonte, zoals bijvoorbeeld enkele meerkeuze examens van het voortgezet onderwijs die altijd vijftig items bevatten. Een restrictie als (11.6) wordt dan ingevuld door $A_i = 1$ te stellen voor alle items en uiteraard geldt $b = 50$. De restrictie zegt dan dat de te maken toets uit precies vijftig items moet bestaan, ongeacht doelfunctie of eventuele andere voorwaarden. Zou aan de eis dat er van alle items die

betrekking hebben op een bepaald hoofdstuk uit een leerboek precies twintig in de toets voorkomen moeten worden voldaan, dan geldt b = 20. Verder geldt dat de A_i's van alle items die horen bij dit hoofdstuk de waarde 1 krijgen, terwijl de A_i's voor de andere items de waarde 0 krijgen. Het is duidelijk, dat het geven van een waarde 1 of 0 aan de A_i's aanduidt of een item al dan niet 'meedoet' in de restrictie. Verderop zullen we zien dat aan de A_i's wel degelijk ook fractionele waarden toegekend kunnen worden of waarden groter dan 1.

Restricties als in (11.7) komen voor wanneer men in de toets bepaalde aspecten van die toets aan een grens wil verbinden die niet overschreden mag worden. Stel dat voor b de maximale afnametijd voor de gehele toets (zeg 50 minuten) wordt gekozen en voor A_i de benodigde tijd voor item i. Dan geeft restrictie (11.7) de eis weer dat de maximale toetsafnametijd vijftig minuten is. Het moge duidelijk zijn dat restricties als in (11.8) voorkomen als bepaalde zaken in een toets aan een ondergrens verbonden worden. Stel dat de toetsconstructeur eist dat op één bepaald vaardigheidspunt de toetsinformatie minimaal gelijk moet zijn aan 12.5. De waarde voor b wordt nu 12.5. Vervolgens berekent men voor alle items de iteminformatie voor dat specifieke vaardigheidspunt. Voor het Raschmodel zullen deze waarden liggen tussen 0 en het maximum 0.25, aannemend dat genormeerd is op een logistische schaal met gemiddelde 0 en discriminatieparameter gelijk aan 1. Dit zijn dan de waarden die aan de A_i's in restrictie (11.8) worden toegekend en in het optimaliserings-model worden opgenomen.

Het is niet mogelijk om een continue toetsinformatiefunctie te specificeren. Wel is het mogelijk om niet één vaardigheidspunt te definiëren maar meer. Zo worden de continue informatiefuncties gediscretiseerd. In alle zogenaamde discretisatiepunten worden de iteminformatiefuncties berekend en wordt een gewenste toetsinformatie opgegeven.

Hier ziet men trouwens hoe een zo belangrijke zaak als toetsinformatie in het optimale toetsconstructieproces kan verschijnen in ofwel de doelfunctie, ofwel in een restrictie. In het algemeen geldt dat dit voor verschillende aspecten van het toetsconstructieproces het geval kan zijn, zie het andere voorbeeld hierboven betreffende de toets van minimale lengte (doelfunctie) of vaste lengte (restrictie). Een combinatie van (11.7) en (11.8) zou kunnen zijn een voorwaarde waarin de onder- en bovengrenzen van aantallen items uit de onderscheiden leerstofcategorieën worden gespecificeerd:

$$A^{\ell} \leq \sum_{i=1}^{K} A_i x_i \leq A^{u}. \qquad (11.9)$$

Stel dat het aantal kennisvragen in de toets tussen een bepaald minimum en maximum moet liggen, zeg tussen vijftien en twintig. In dat geval wordt $A^{\ell} = 15$ en $A^{u} = 20$.

Definiëren we $A_i = 1$ voor alle kennisitems en $A_i = 0$ voor alle andere items, dan geeft (11.9) de eis weer dat er tussen vijftien en twintig kennisitems in de toets moeten worden opgenomen.

11.2.2 Praktijkvoorbeelden

Hoewel uit enkele combinaties van doelfuncties met restricties eenvoudige voorbeelden van toetsconstructie kunnen worden geformuleerd, zal er in de praktijk meestal sprake zijn van één doelfunctie en nagenoeg altijd van verschillende restricties. Het moge duidelijk zijn dat het probleem van het construeren van een toets van minimale lengte met een gespecificeerde ondergrens voor toetsinformatie op één discretisatiepunt zonder verdere restricties triviaal is vanuit zowel psychometrisch standpunt als optimaliseringsstandpunt. Daar in nagenoeg alle gevallen van toetsconstructie in het kader van itemresponstheorie gebruik wordt gemaakt van specificaties van toetsinformatie, zal eerst een aantal gevallen worden behandeld die in de toetspraktijk zullen voorkomen, waarbij we ons concentreren op deze toetsinformatie. Uitgewerkte voorbeelden worden om praktische redenen tot beperkte omvang gehouden. Bij de voorbeelden hierna zal voor de vaardigheidsschaal de logistische θ-schaal gebruikt worden in het praktische bereik van $\theta = -3$ tot $\theta = 3$.

Bij de specificatie van de toetsinformatie wordt de toetsconstructeur gedwongen goed voor ogen te houden wat het gebruiksdoel van de toets is. Daar er in de praktijk altijd met een eindig aantal items gewerkt wordt, is het mogelijk dat er geen enkele toets is te vinden die aan alle te bedenken gebruiksdoelen op gelijkwaardige wijze voldoet. Stel dat een toets-constructeur vooral geïnteresseerd is in zak-slaagbeslissingen. Een eis die aan de te maken toets gesteld moet worden is dat deze het meest nauwkeurig meet op het zak-slaagpunt, aangezien er voor kandidaten met een geschatte vaardigheid in dit gebied belangrijke beslissingsfouten gemaakt kunnen worden. Kandidaten met hoge of lage vaardigheid zullen door meetonnauwkeurigheid in het cesuurgebied niet benadeeld of bevoordeeld worden. Stel dat het cesuurpunt ligt op die vaardigheid, zodat vijftig procent van de groep studenten zakt en vijftig procent slaagt. De gewenste ondergrens voor de toetsinformatie in dit vaardigheids-punt wordt gesteld op 10. Voor het 25e en 75e percentiel wordt een toetsinformatie van 8 geëist. Dit heeft als gevolg dat het verloop rondom de piek van de toetsinformatie iets vlakker wordt. Het volgende schema kan dan gepresenteerd worden (zie tabel 11.1).

Tabel 11.1

Het eerste programmeringsprobleem

Specificatie θ-niveau bij percentiel	Iteminformatie voor item 1, item 2,..., item K	Ondergrens toetsinformatie bij θ-niveau
25	$I_1(\theta_1),...,I_K(\theta_1)$	$I(\theta) = 8$
50	$I_1(\theta_2),...,I_K(\theta_2)$	$I(\theta) = 10$
75	$I_1(\theta_3),...,I_K(\theta_3)$	$I(\theta) = 8$

Stel dat het de wens van de toetsconstructeur is deze specificatie met een zo gering mogelijk aantal items te bereiken, dan zal voor bovengenoemd voorbeeld de mathematische formulering van het optimaliseringsprobleem als volgt luiden:

minimaliseer $x_1 + x_2 + \cdots + x_K$

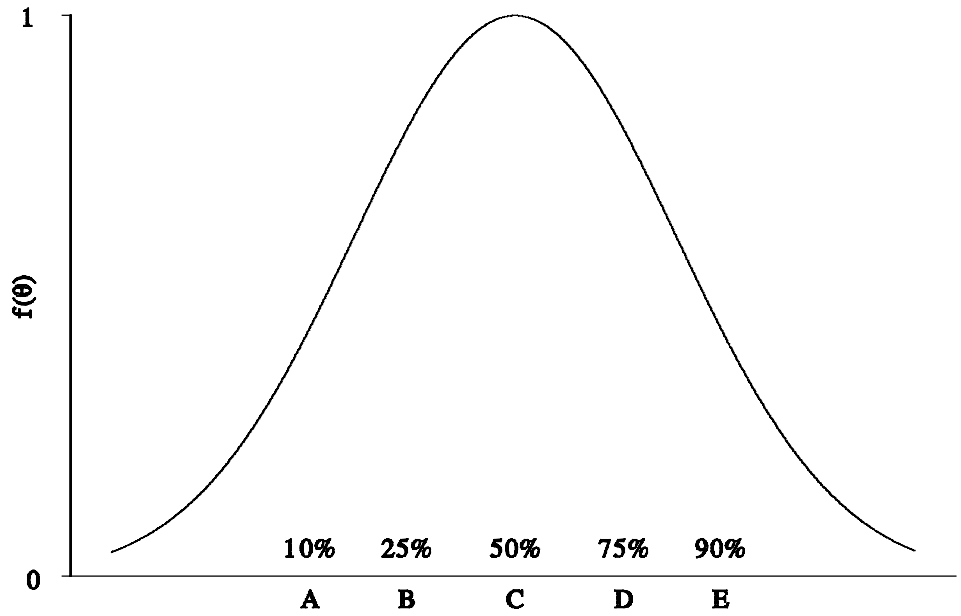
onder voorwaarde dat $I_1(\theta_1)x_1 + I_2(\theta_1)x_2 + \cdots + I_K(\theta_1)x_K \geq 8,$

$I_1(\theta_2)x_1 + I_2(\theta_2)x_2 + \cdots + I_K(\theta_2)x_K \geq 10,$

$I_1(\theta_3)x_1 + I_2(\theta_3)x_2 + \cdots + I_K(\theta_3)x_K \geq 8,$

$x_i = \varepsilon \{0,1\}, \quad i = 1,...,K.$

Uitgaande van een itembank van vijfhonderd gecalibreerde rekenitems kunnen we de praktijk van toetsconstructie verduidelijken. Als discretisatiepunten kiezen we hier de vaardigheden die overeenkomen met het 25e, 50e en 75e percentiel in de doelgroep. Alleen op deze discretisatiepunten worden de iteminformatie-functies, de te bereiken toetsinformatiefunctie en de bereikte toetsinformatie beschouwd. Deze vaardigheidsniveaus zijn in figuur 11.4 aangegeven met B, C en D.

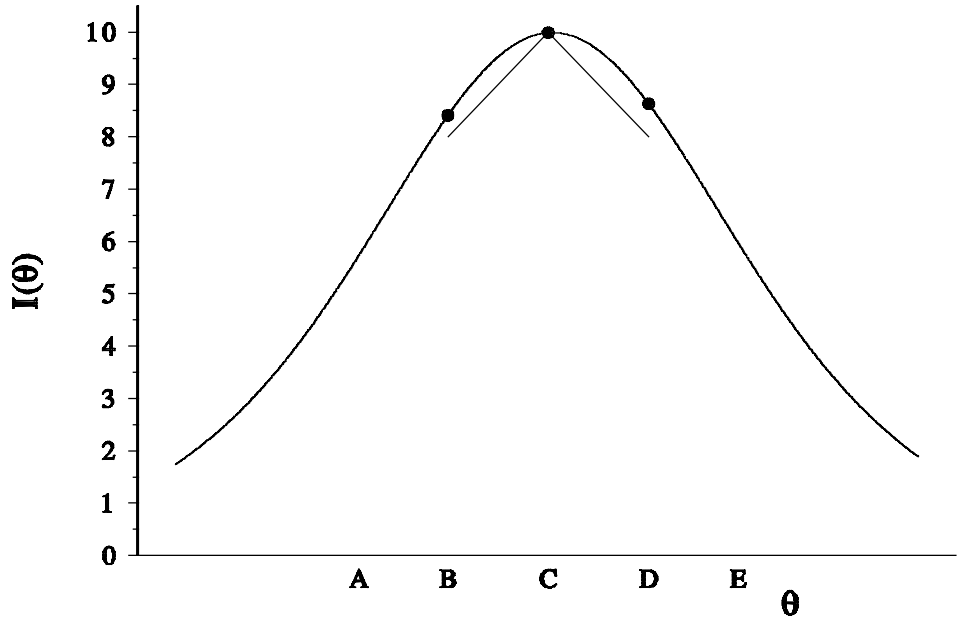


Figuur 11.4 Discretisatiepunten voor de toetsconstructie

Het voert te ver om in detail te laten zien hoe de branch-and-bound een oplossing vindt voor dit probleem. Het resultaat van de oplosmethode kunnen we echter wel laten zien. De geëiste toetsinformatie en de bereikte toetsinformatie met 40 items staan weergegeven in figuur 11.5

Punten A en E zijn in figuur 11.4 en figuur 11.5 toegevoegd om een vergelijking met het probleem van figuur 11.6 te vereenvoudigen.

Stel echter dat de constructeur een geheel ander doel voor ogen staat, namelijk een toets voor zeer algemeen gebruik voor het meten van vaardigheid en hij of zij vindt, dat - uit hoofde van sociale rechtvaardigheid - iedere leerling er recht op heeft met ongeveer dezelfde nauwkeurigheid gemeten te worden. Dit impliceert dat de gewenste toetsinformatie over het relevante gedeelte van de vaardigheidsschaal zoveel mogelijk uniform moet zijn.



Figuur 11.5 Geëiste en bereikte toetsinformatie voor het eerste probleem

Als de toetslengte niet onbeperkt toe mag nemen, impliceert dit tevens dat de gespecificeerde (uniforme) toetsinformatie beduidend lager moet zijn dan in het eerste voorbeeld. Stel de toetsspecificatie is het maken van een toets van minimale omvang en met toetsinformatie 6.0 op de θ-niveaus die behoren bij het 10e, 25e, 50e, 75e en 90e percentiel. Een schema van de formulering van dit probleem wordt weergegeven in tabel 11.2.

Tabel 11.2
Schema van het tweede probleem

Specificatie toetsinformatie θ-niveau bij percentiel	Iteminformatie voor item 1, item 2, ..., item K	Ondergrens bij θ-niveau
10	$I_1(\theta_1),...,I_K(\theta_1)$	$I(\theta_1) = 6$
25	$I_1(\theta_2),...,I_K(\theta_2)$	$I(\theta_2) = 6$
50	$I_1(\theta_3),...,I_K(\theta_3)$	$I(\theta_3) = 6$
75	$I_1(\theta_4),...,I_K(\theta_4)$	$I(\theta_4) = 6$
90	$I_1(\theta_5),...,I_K(\theta_5)$	$I(\theta_5) = 6$

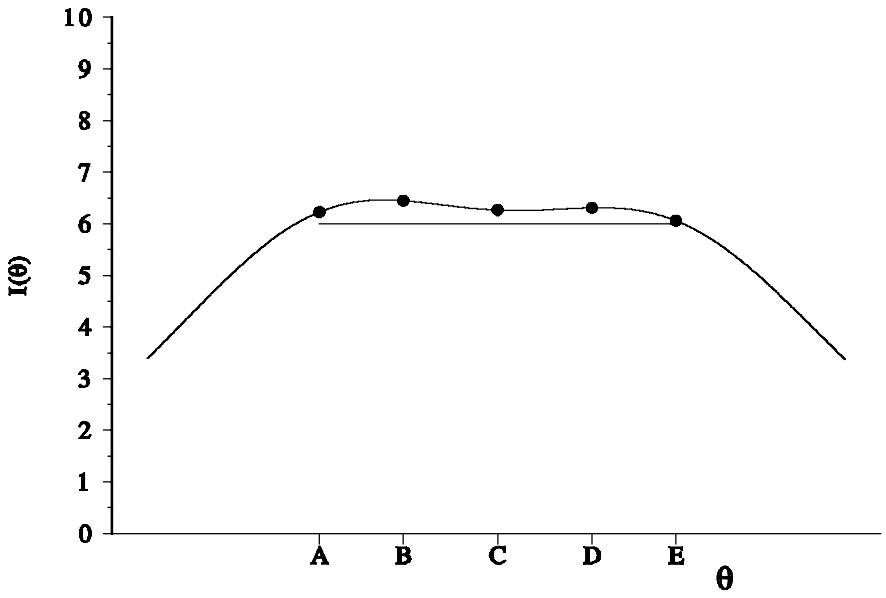
Als ook hier de toets uit een zo gering aantal mogelijk aantal items moet bestaan, dan zal de mathematische formulering luiden:

minimaliseer $\sum_{i=1}^{K} x_i$

onder voorwaarde dat $\sum_{i=1}^{K} I_i(\theta_m)\, x_i \geq I(\theta_m), \qquad m = 1,...,5$

$x_i \; \varepsilon \; \{0,1\}, \qquad i = 1,...,K.$

Figuur 11.6 laat de informatiefunctie van de nu geconstrueerde toets zien. Deze toets bestaat uit 40 items, net als de toets die geconstrueerd is voor het eerste probleem. Merk op dat om een meer gelijkmatige meetnauwkeurigheid te bereiken de toetsinformatie in het punt C lager is dan in het eerste voorbeeld.



Figuur 11.6
De toetsinformatie voor het tweede probleem

In voorgaande voorbeelden werd de gewenste toetsinformatie geformuleerd als een restrictie in het optimaliseringsprobleem. We geven nu een voorbeeld van toetsinformatie in de doelfunctie, waarbij een gewenste vorm van de toetsinformatiefunctie wordt gespecificeerd in plaats van de hoogte. Dit is nuttig als de toetsconstructeur slechts globaal kan aangeven hoe de verhouding van de toetsinformatie voor de verschillende vaardigheidsgebieden moet zijn. Deze situatie zal

bij voorbeeld ontstaan als de toetsconstructeur wel weet waarvoor de informatiefunctie dient, maar geen ervaring heeft in het omgaan met deze functie of met de getalsmatige aspecten ervan. De constructeur zou dan op de gewenste M specificatiepunten op de vaardigheidsschaal fiches kunnen plaatsen, zodanig dat de aantallen r_m (m = 1,...,M) de gewenste verhouding weerspiegelen. Vervolgens moeten de items zo gekozen worden dat de toetsinformatie gemaximaliseerd wordt met behoud van de vorm. Dit houdt in dat de toetsinformatie voor het θ_m-punt waarvoor de verhouding tussen toetsinformatie en r_m het laagst is, wordt gemaximaliseerd. Dit wordt in de volgende doelfunctie geformuleerd:

$$\text{maximaliseer} \left\{ \text{minimum} \quad \frac{I(\theta_m)}{r_m} \right\} = \left\{ \text{minimum} \quad \frac{\sum_{i=1}^{K} I_i(\theta_m)\, x_i}{r_m} \right\}.$$

Hierbij geldt $x_i \varepsilon \{0,1\}$. Daar de simplexmethode lineariteit van de doelfunctie vereist, dus geen 'knik' in het functieverloop of discontinuïteit toestaat, moet er een extra maatregel genomen worden. Dit is de introductie van een dummyvariabele y die de doelfunctie lineair maakt. Dummyvariabelen worden gebruikt om een probleem te kunnen formuleren maar spelen zelf geen rol in de oplossing van het eigenlijke probleem. Dit leidt dan tot het volgende optimaliseringsprobleem:

maximaliseer $\quad y$

onder voorwaarde dat $\quad y \leq \dfrac{\sum_{i=1}^{K} I_i(\theta_m)\, x_i}{r_m}, \qquad m = 1,...,M$

ofwel, na herschrijving, $\quad \sum_{i=1}^{K} I_i(\theta_m)\, x_i - r_m y \geq 0 \qquad m = 1,...,M.$

In deze restrictie worden ondergrenzen $r_m\, y$ aan de toetsinformatie geformuleerd voor elk van de θ_m-punten. De maximalisatie van y, en daarmee van de grootheden $r_m\, y$, 'duwt' de toetsinformatie omhoog. Zoals eerder vermeld leidt deze formulering tot opname van alle beschikbare items. Dus wordt de volgende restrictie toegevoegd:

$$\sum_{i=1}^{K} x_i = k,$$

waar k de gewenste lengte van de toets is. Voorts uiteraard weer $x_i \in \{0,1\}$ en y niet-negatief (waarom?). Deze modellen staan bekend onder de naam maximinmodellen, vanwege het feit dat het minimum over een aantal functies wordt gemaximaliseerd.

Ook hier geven we een voorbeeld uit de eerder genoemde itembank van vijfhonderd rekenitems. Naast calibratiegegevens zijn echter ook vakinhoudelijke gegevens beschikbaar: ieder item is gecategoriseerd als een optelling, een aftrekking, een vermenigvuldiging of een deling. Deze categorieën zijn hieronder vermeld als categorie 10, 11, 12 en 13. Stel dat de toetsconstructeur een toets wil samenstellen van veertig items, met tien optellingen, tien aftrekkingen, tien vermenigvuldigingen en tien delingen. Deze eis kan worden geformuleerd zoals in (11.9). Voor m kiezen we 10, 11, 12 en 13. Verder definiëren we $A_{10,i} = 1$ voor alle optellingen, en $A_{10,i} = 0$ voor de andere items. De andere A_{mi}'s worden op dezelfde wijze gedefinieerd. Nu geldt:

$$\sum_{i=1}^{K} A_{mi} x_i = 10, \qquad m = 10,...,13.$$

Daarnaast moet de toets nauwkeuriger meten in het vaardigheidsgebied van de iets zwakkere leerlingen: de toetsinformatie voor het 10e tot het 50e percentiel moet twee keer zo hoog zijn als de toetsinformatie voor het 90e percentiel. Dit komt tot uitdrukking in figuur 11.7. Hier geldt $r_1 = 10$, $r_2 = 10$, $r_3 = 5$.

Het gehele model wordt geformuleerd als:

maximaliseer y

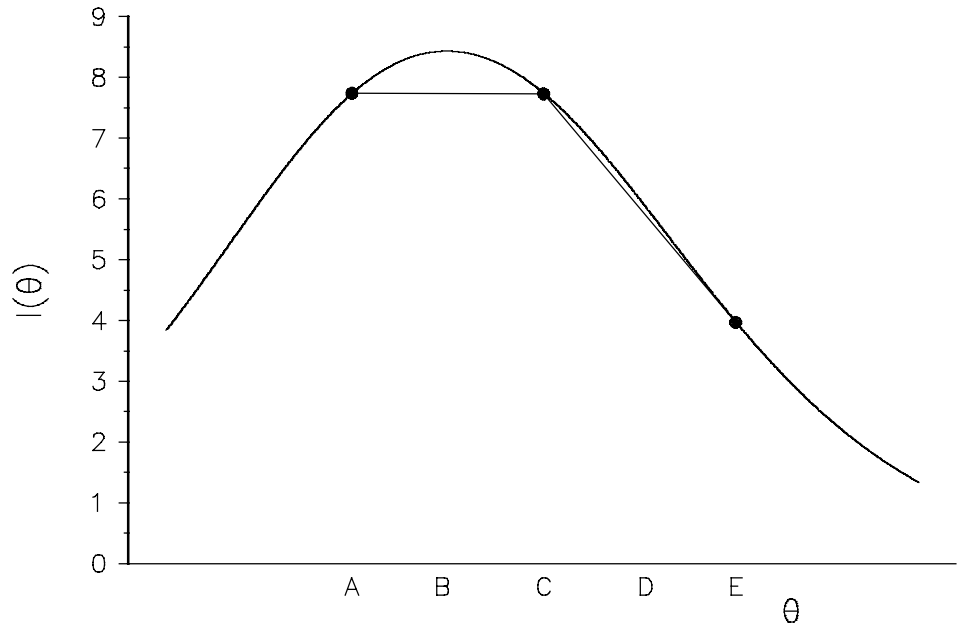
onder voorwaarde dat
$$\sum_{i=1}^{K} I_i(\theta_m) x_i - r_m y \geq 0 \qquad m = 1,...,M$$

$$\sum_{i=1}^{K} x_i = 40$$

$$\sum_{i=1}^{K} A_{mi} x_i = 10 \qquad m = 10,...,13$$

$$x_i \in \{0,1\} \qquad i = 1,...,K.$$

Merk op dat in figuur 11.7 de variabele y de waarde 0.77 heeft in de optimale oplossing. In figuur 11.7 zijn zowel de bereikte toetsinformatie gegeven als de grootheden $r_m y$.



Figuur 11.7
De toetsinformatie behorend bij het derde probleem

In de praktijk ontstaan vaak situaties waarbinnen behoefte is aan toetsen die dezelfde lokale meetnauwkeurigheid hebben. In het kader van de itemresponstheorie worden toetsen zwak parallel genoemd als ze identieke toetsinformatiefuncties hebben (Samejima, 1977). Behoefte aan parallelle toetsen ontstaat in die situaties waarin het gewenst is dat toetsen uitwisselbaar zijn, bijvoorbeeld bij kort opeenvolgende herhaalde metingen van dezelfde personen. Parallelle toetsconstructie kan zowel sequentieel als simultaan plaatsvinden. Voor een uit-gebreid overzicht zie Boekkooi-Timminga (1990). Bij sequentiële constructie is er sprake van een opeenvolging van toetsconstructies, waarbij men steeds rekening moet houden met hetgeen voorafging. Bij simultane constructie probeert men gelijktijdig een verzameling items te verdelen over een aantal toetsen. Voor een itembank die goed gevuld is met items die relevant zijn voor het toetsconstructieprobleem dat aan de orde is, blijkt het in de praktijk vaak al voldoende om eerst één toets volgens de specificatie te laten maken. Vervolgens geeft men bij de aanmaak van de tweede toets die items die reeds in de eerste toets zijn opgenomen een gewicht van 2 in plaats van 1 in de doelfunctie. Hierdoor is het vrijwel uitgesloten dat

deze items in een volgende toets worden opgenomen. Dit geldt uiteraard alleen als de doel-functie de minimalisering van het aantal items betreft.

Gebruikt men het eerder beschreven maximinmodel dan kan door het toevoegen van de drie volgende restricties een tweede parallelle toets geconstrueerd worden:

$$\sum_{i=1}^{K} I_i(\theta_m)\, x_{it} \geq (1 - p)\, I(\theta_m)$$

$$\sum_{i=1}^{K} I_i\,(\theta_m)\, x_{it} \leq (1 + p) I(\theta_m)$$

$$\sum_{t=1}^{T} x_{it} \leq 1 \qquad i = 1,...,K.$$

De eerste twee restricties geven een ondergrens en een bovengrens voor afwijking tussen de gewenste toetsinformatie en de bereikte informatie. De derde restrictie stipuleert dat geen enkel item in meer dan een toets aanwezig mag zijn. In het model is $I(\theta_m)$ de reeds bestaande toetsinformatie op punt θ_m, en p is het maximaal toegestane verschil in toetsinformatie tussen de reeds bestaande en nog te realiseren toets. Ook hier geldt dat het raadzaam is goed gevulde banken te gebruiken.

11.2.3 Specificeren van restricties en relaties

In het voorgaande lag de nadruk op specificaties van aantallen items en toetsinformatie in zowel doelfuncties als restricties. Vele andere specificaties kunnen eveneens gemodelleerd worden als restrictie of doelfunctie (Van der Linden & Boekkooi-Timminga, 1989). Zo werd reeds gewezen op de mogelijkheid de verdeling van items over inhoudelijke categorieën te modelleren. Hetzelfde geldt voor de afnametijd van de toets, door ofwel een bovengrens als restrictie op te nemen, ofwel door opname in de doelfunctie, ofwel door het zogenaamde multi-stage programming, waarin specificaties afwisselend in doelfunctie en restrictie terecht komen. Voorts blijkt het mogelijk om te werken met clusters van items, een situatie die zich voordoet bij tekstbegriptoetsen (Theunissen, 1987). Hier horen bij één tekst in de regel verschillende items en kunnen de teksten alleen met alle bijbehorende items tegelijk geselecteerd worden.

Nu zullen we zien hoe het constructieproces desnoods zeer gedetailleerd op verschillende niveaus en zeer specifiek gestuurd kan worden, zodat bijvoorbeeld ook aan detailwensen gehoor kan worden gegeven. We sluiten daarbij weer aan bij het volgende basismodel:

minimaliseer $\sum_{i=1}^{K} c_i x_i$

onder voorwaarde dat $\sum_{i=1}^{K} A_i x_i \geq b,$

$x_i \in \{0,1\}.$

Eén mogelijkheid behelst het introduceren van dummyvariabelen. Dit kan bijvoorbeeld nodig zijn voor bijsturing van het toetsconstructieproces op het niveau van de restricties. Stel we hebben in de specificatie van het constructieproces opgenomen de restrictie:

$$\sum_{i=1}^{K} A_i x_i \leq b. \tag{11.10}$$

Laten we nu aannemen dat deze restrictie niet altijd van kracht hoeft te zijn, maar pas geldt als een bepaald item, of een bepaalde groep van items, in de toets wordt opgenomen. Bovenstaande restrictie kan bijvoorbeeld betrekking hebben op de gemiddelde tijd die nodig is voor het maken van een item, waarbij de coëfficiënten A_i voor de antwoordtijd per item staan en b de maximale toetstijd is. De restrictie wordt geacht mee te gaan spelen bij opname van items met een lange antwoordtijd. Als dit gebeurt wordt een dummyvariabele δ gelijk gesteld aan 1 en vervolgens geldt

$$\delta = 1 \rightarrow \sum_{i=1}^{K} A_i x_i \leq b, \tag{11.11}$$

waar $\rightarrow$ betekent 'impliceert'. We stellen nu het getal G als een bovengrens voor de uitdrukking $\sum A_i x_i - b$. Als $\delta = 1$ (ofwel, als $1 - \delta = 0$), wensen we dat $\sum A_i x_i - b \leq 0$, hetgeen volgt uit (11.11). Als G voldoende groot wordt gekozen, zal dit het geval zijn als $\sum A_i x_i - b \leq G(1 - \delta)$. Na enige herordening krijgen we dan uit de conditie (11.11) de volgende restrictie:

$$\sum_{i=1}^{K} A_i x_i + G\delta \leq G + b. \tag{11.12}$$

Uit (11.12) volgt, dat als $\delta = 0$ er geen sprake is van een restrictie, terwijl bij $\delta = 1$ de restrictie (11.10) van kracht is. Het verband tussen het 'optreden' van item *i* en de dummyvariabele δ wordt gelegd door de volgende restrictie te introduceren: $x_j - \delta \leq 0$. Dit houdt in dat δ de waarde 1 aanneemt als x_i groter is dan 0, dat wil zeggen, gelijk is aan 1.

Na formuleringen besproken te hebben die betrekking hebben op het niveau van de restricties van het toetsconstructieprobleem, zijn we nu aangekomen op het punt waar formuleringen worden gebruikt op het niveau van de items en hun onderlinge relaties. De variabelen zijn hier weer de beslisvariabelen x_i, die aangeven of de desbetreffende items gekozen worden. Uitspraken over een item of over de relaties tussen items worden geformuleerd via de volgende verzameling van operatoren:

- $\vee$ betekent of x of y of allebei,
- $\wedge$ betekent x en y tegelijk,
- $\neg$ betekent niet x,
- $\rightarrow$ betekent als...dan (implicatie),
- $\leftrightarrow$ betekent dan en slechts dan.

We kunnen bovenstaande operatoren met enige eenvoudige voorbeelden demonstreren. We stellen ons voor dat uit een itembank toetsen samengesteld moeten worden waarbij steeds de items 1 en 2 een rol spelen. Door verschillen in de toetsspecificatie kunnen onder andere de volgende verschillende eisen aan items 1 en 2 gesteld worden.
De eis, dat ofwel item 1 ofwel item 2 ofwel beide moet worden opgenomen, wordt geformuleerd als $x_1 \vee x_2$ en in de vorm van restrictie in het optimaliseringsprobleem als $x_1 + x_2 \geq 1$. De eis, dat zowel item 1 als item 2 moeten worden opgenomen, wordt geformuleerd als $x_1 \wedge x_2$ en in de vorm van restrictie als $x_1 + x_2 \geq 2$. De eis dat item 1 niet op-genomen mag worden wordt uitgedrukt als $\neg x_1$ en in de vorm van restrictie als $x_1 = 0$. De eis dat als item 1 wordt opgenomen ook item 2 moet worden opgenomen, wordt $x_1 \rightarrow x_2$ en in de vorm van restrictie $x_1 - x_2 \leq 0$. De eis dat item 1 en item 2 alleen tezamen mogen worden opgenomen, wordt geformuleerd als $x_1 \leftrightarrow x_2$ en in de vorm van restrictie als $x_1 - x_2 = 0$. Het verschil tussen beide laatste formuleringen ligt in het feit dat in het laatste geval item 2 alleen kan optreden samen met item 1, terwijl in het voorlaatste geval item 2 ook los van item 1 kan optreden, vandaar in het voorlaatste geval het '≤' teken. Vanuit deze elementaire uitdrukkingen kunnen verdere expressies geformuleerd worden van iedere noodzakelijke graad van complexiteit.

Tot besluit een voorbeeld: stel we formuleren als eis dat, als item 1 of item 2 of beide worden opgenomen, dan minstens één van de items 3, 4 of 5 moet worden opgenomen. Dit wordt geformuleerd als:

$$(x_1 \vee x_2) \rightarrow (x_3 \vee x_4 \vee x_5). \qquad (11.13)$$

Het linker lid van (11.13) wordt als restrictie $x_1 + x_2 \geq 1$, en het rechter lid $x_3 + x_4 + x_5 \geq 1$. Vervolgens introduceren we een nieuwe indicatorvariabele δ en stellen dat moet gelden $x_1 + x_2 \geq 1 \rightarrow \delta = 1$, en tevens dat $\delta = 1 \rightarrow x_3 + x_4 + x_5 \geq 1$.
Eis (11.13) wordt dan geformuleerd als de volgende twee restricties: $x_1 + x_2 - 2\delta \leq 0$ en $-x_3 - x_4 - x_5 + \delta \leq 0$. Met gebruik van dit soort formuleringen kan het proces van samen-stellen van toetsen minutieus gestuurd worden. Er kan echter ook een nadeel aan kleven. Als er teveel restricties toegevoegd worden aan het optimaliseringsprobleem, kan er een situatie ontstaan waarbij de algoritmen die gebruikt worden om de oplossing te vinden minder effectief worden.

Binnen het korte bestek van deze paragraaf kon niet alles wat er te zeggen valt over de optimale samenstelling van toetsen binnen de itemresponstheorie aan de orde komen. Zo werd niet ingegaan op de mogelijkheid om verscheidene doelfuncties te samen te optimaliseren, het zogenaamde 'multi-objective' programmeren. Ook is grotendeels onbesproken gelaten de ont-wikkeling van heuristische methoden die gebruikt kunnen worden als exacte algoritmen voor de oplossing van optimaliseringsproblemen teveel computertijd zouden vergen. Ook is weinig aandacht besteed aan de beschikbaarheid van computerprogrammatuur voor de optimale samenstelling van toetsen. Voor dit laatste verwijzen we naar de handleiding van het computerprogramma Optimal Test Design (Verschoor, 1991).

11.3 Het samenstellen van toetsen in de klassieke testtheorie

In zijn boek over klassieke testtheorie opent Gulliksen (1950) het hoofdstuk over itemselectie als volgt: 'Basically, item analysis is concerned with the problem of selecting items for a test, so that the resulting test will have certain specified characteristics' (p. 363). In hoofdstuk 3 zagen we dat in de klassieke testtheorie de betrouwbaarheid een belangrijk kenmerk van een toets is. Gulliksen beschrijft een grafische procedure voor de selectie van items die de betrouwbaarheid van de toets maximaliseert wanneer de samen te stellen toets uit een vooraf bepaald aantal items bestaat. Welke items de betrouwbaarheid meer doen toenemen dan andere items, kan toegelicht worden aan de hand van Cronbachs coëfficiënt alpha, die gedefinieerd is als

$$\alpha = k\,(k-1)^{-1}\left[1 - \left(\sum_{i=1}^{k}\sigma_i^2\right) \Big/ \left(\sum_{i=1}^{k}\sigma_i\,\rho_{it}\right)^2\right], \tag{11.14}$$

waarbij k het aantal items in de toets, σ_i^2 de variantie van item i, en ρ_{it} de correlatie tussen de score op item i en de score op de toets is. Uit formule (11.14) kan afgeleid worden dat wanneer het aantal items in de toets gefixeerd is, coëfficiënt alpha gemaximaliseerd wordt door het minimaliseren van de ratio

$$\left(\sum_{i=1}^{k}\sigma_i^2\right) \Big/ \left(\sum_{i=1}^{k}\sigma_i\,\rho_{it}\right)^2. \tag{11.15}$$

De ratio (11.15) laat zien dat minimalisatie kan worden bereikt door verkleining van de teller, de som van de varianties van de items, of door vergroting van de noemer, de gekwadrateerde som van de betrouwbaarheidsindices van de items. Merk op dat de variantie van de items zowel in de teller als in de noemer van de ratio voorkomt. In hoofdstuk 3 zagen we dat aanzienlijke verschillen in moeilijkheidsgraad slechts aanleiding geven tot kleine verschillen in itemvarianties. Het onderzoek van Ebel (1967) laat dan ook zien dat de betrouwbaarheid minder afhangt van de teller dan van de noemer van (11.15). Dit betekent dat voor het maximaliseren van de betrouwbaarheid met name items met een hoge item-testcorrelatie geselecteerd moeten worden. Het laatste gegeven betekent dat we de niet-lineaire doelfunctie (11.15) kunnen vervangen door een lineaire doelfunctie. Het oplossen van problemen met lineaire doelfuncties veel eenvoudiger is dan het oplossen van problemen met niet-lineaire doelfuncties. Adema en Van der Linden (1989) formuleerden het volgende lineaire programmeringsmodel voor het samenstellen van toetsen:

maximaliseer
$$\sum_{i=1}^{K}\rho_{it}\,x_i \tag{11.16}$$

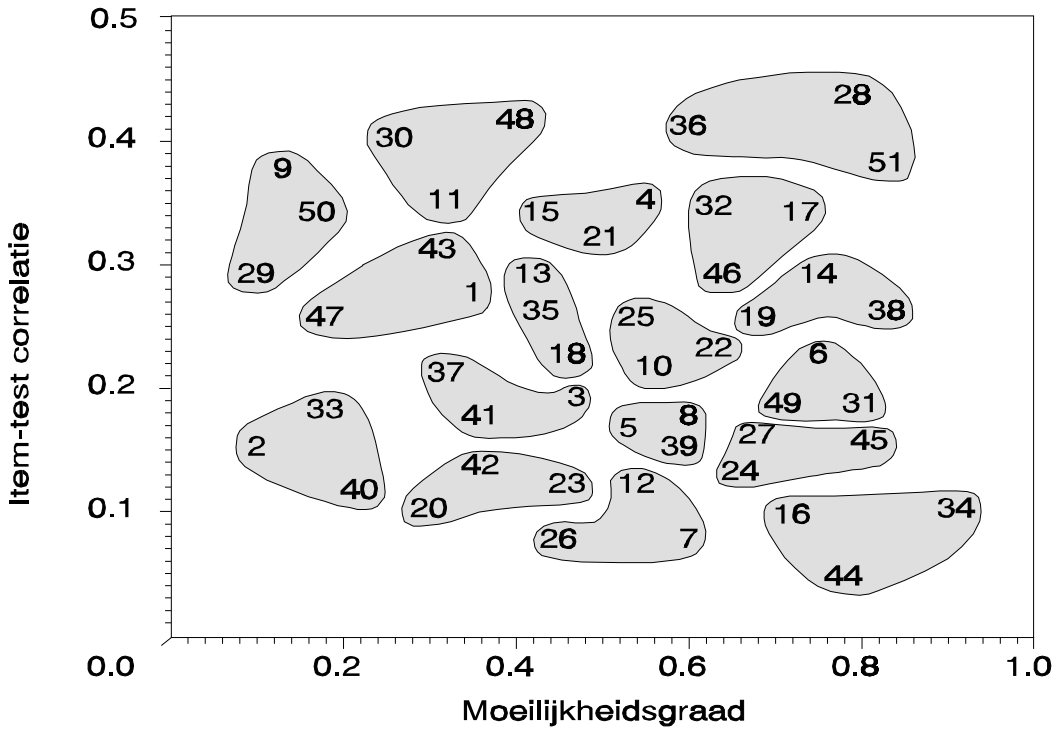
onder voorwaarde dat
$$\sum_{i=1}^{K}x_i = k, \tag{11.17}$$

$$\sum_{i=1}^{K}t_i\,x_i \le 35\,k, \tag{11.18}$$

$$x_i \in \{0,1\}, \quad i = 1,\ldots,K. \tag{11.19}$$

In het bovenstaande model wordt de betrouwbaarheid gemaximaliseerd door middel van een doelfunctie (11.16) die de voorkeur verwoordt voor items met hoge item-testcorrelaties. In dit model worden verder nog twee voorwaarden geformuleerd. Dat de toets uit *k* items moet bestaan wordt in voorwaarde (11.17) geformuleerd. De opname van deze voorwaarde in het model is noodzakelijk om de lengte van de toets te beperken omdat elk item met een positieve item-testcorrelatie de betrouwbaarheid van de toets verhoogt. In voorwaarde (11.18) staat dat er t_i seconden nodig zijn voor de beantwoording van item *i*. In de voorwaarden wordt echter ook gesteld dat de totale toets binnen 35 *k* seconden afgenomen moet zijn, wat de selectie van items met een relatief korte antwoordtijd impliceert.

Voor de samenstelling van parallelle toetsen ontwikkelde Gulliksen de 'matched random subtests method' (1950, p. 207 ev.). Hierbij wordt elk item afgebeeld als een punt in een grafiek met als abscis de moeilijkheidsgraad en als ordinaat de item-testcorrelatie. Op basis van deze itemparameters worden de items dan eerst simultaan gekoppeld en daarna wordt ieder item van elk gekoppeld paar of drietal random toegewezen aan een toets.



Figuur 11.8

De constructie van drie parallelle tests door simultane koppeling van item op basis van moeilijkheidsgraad en item-testcorrelatie

Figuur 11.8 laat voor 51 items het resultaat zien van de eerste stap van deze twee-staps procedure, namelijk 17 gekoppelde drietallen. De tweede stap is dat item 2 aan bijvoorbeeld de eerste toets, item 33 aan de tweede toets, item 40 aan de derde toets, item 20 aan de tweede toets, enz. wordt toegewezen. Het resultaat van de procedure is drie parallelle toetsen die elk uit 17 items bestaan.

Van der Linden en Boekkooi-Timminga (1988) ontwikkelden een binair programmerings-model voor de 'matched random subtests method' van Gulliksen. Voor de constructie van twee parallelle toetsen luidt het model:

$$\text{minimaliseer} \quad \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} [(\pi_i - \pi_j)^2 + (\rho_{it} - \rho_{jt})^2]^{½} \; x_{ij} \tag{11.20}$$

$$\text{onder voorwaarde dat} \quad \sum_{i=1}^{j-1} x_{ij} + \sum_{i=j+1}^{K} x_{ji} = 1 \quad (j = 1,...,K) \tag{11.21}$$

$$x_{ij} \varepsilon \{0,1\} \qquad (i = 1,...,K-1; j = i+1,...,K). \tag{11.22}$$

De eerste stap van Gulliksen's grafische methode vervangen Van der Linden en Boekkooi-Timminga door doelfunctie (11.20), minimalisatie van de som van de binnen-paar Euclidische afstanden, en door de voorwaarden in (11.21) die garanderen dat elk item tot niet meer dan één paar items behoort. De binaire beslisvariabele x_{ij} geeft aan of i en j wel een paar zijn $(x_{ij} = 1)$ of geen paar zijn $(x_{ij} = 0)$. De eerste stap resulteert in $K/2$ paren items. Ook voor de tweede stap, het random toewijzen van items aan toetsen, formuleren zij binaire programmeringsmodellen met als doelfuncties gelijke gemiddelden en/of varianties.

Van der Linden en Boekkooi-Timminga geven de voorkeur aan een model voor parallelle toetsconstructie waarbij de items eerst in paren, drietallen enzovoort gekoppeld worden en niet direct aan toetsen toegewezen worden. Dit laatste model zou tot minder wenselijke toetsen kunnen leiden omdat de ene toets dan bijvoorbeeld uit items met nagenoeg dezelfde moeilijkheidsgraad bestaat terwijl de andere toets uit items van nogal verschillende moeilijkheidsgraad bestaat. Toetsen met dezelfde itemparameters - en daardoor ook dezelfde toetsparameters - voor corresponderende items, worden sterk-parallelle toetsen genoemd, terwijl toetsen met alleen dezelfde toetsparameters, zwak-parallelle toetsen genoemd worden. Het probleem van het

construeren van een toets die identiek is aan een reeds bestaande toets, hier aangeduid met referentietoets, is een variant van het probleem van het samenstellen van sterk-parallelle toetsen. Een oplossing voor dit probleem met behulp van technieken uit de mathematische programmering wordt beschreven in Armstrong, Jones en Wu (1992). Hun oplossing bestaat hieruit dat eerst getracht wordt de items in de itembank zo goed mogelijk te koppelen aan de items uit de referentietoets. Daarna worden parallelle toetsen samengesteld die zo weinig mogelijk afwijken van de referentietoets.

Het samenstellen van parallelle toetsen vormt ook het uitgangspunt van de twee modellen beschreven door Verschoor en Sanders (1993). Het samenstellen van een enkele toets wordt opgevat als een speciaal geval van parallelle toetsconstructie, namelijk een toets die parallel is met zichzelf. Het doel van model 1 van Verschoor en Sanders is om onder bepaalde voorwaarden het aantal items van de samen te stellen parallelle toetsen te minimaliseren. Het doel van model 2 is om onder bepaalde voorwaarden de betrouwbaarheid van parallelle toetsen te maximaliseren. De twee modellen gaan uit van klassieke itemparameters, dat wil zeggen van een verzameling items die met het klassieke testmodel gecalibreerd zijn of waarvan de klassieke itemparameters afgeleid zijn van de itemparameters van een item-responstheorie model. Deze laatste mogelijkheid kan nuttig zijn voor personen die onvoldoende bekend zijn met itemresponstheorie maar toch gebruik willen maken van nieuwe technieken voor het samenstellen van toetsen.

Model 1 beoogt om met zo weinig mogelijk items parallelle toetsen samen te stellen die een gespecificeerde betrouwbaarheid, gemiddelde toetsscore en standaarddeviatie hebben. De formulering van model 1 luidt:

minimaliseer
$$\sum_{i=1}^{K} x_{i1} \tag{11.23}$$

onder voorwaarde dat
$$\sum_{i=1}^{K} x_{i1} = \sum_{i=1}^{K} x_{it}, \qquad t = 2,...,T \tag{11.24}$$

$$\alpha^{\ell} \leq \alpha_t \leq \alpha^{u}, \qquad t = 1,...,T \tag{11.25}$$

$$\pi^{\ell} \sum_{i=1}^{K} x_{it} \leq \sum_{i=1}^{K} \pi_i \, x_{it} \leq \pi^{u} \sum_{i=1}^{K} x_{it}, \qquad t = 1,...,T \tag{11.26}$$

$$\sigma_x^{l} \leq \sigma_x \leq \sigma_x^{u} \tag{11.27}$$

$$\sum_{t=1}^{T} x_{it} \leq 1 \qquad i = 1,...,K. \tag{11.28}$$

De doelfunctie (11.23) beoogt het minimaliseren van het aantal items van de parallelle toetsen. In voorwaarde (11.24) staat dat voor alle T toetsen geldt dat ze uit evenveel items als de eerste toets dienen te bestaan. Voor elk item i is beslisvariabele x_{it} gedefinieerd als 1 indien item i in toets t is opgenomen en als 0 indien item i niet in toets t is opgenomen. In voorwaarde (11.25) worden de ondergrens, α^{ℓ}, en de bovengrens, α^{u}, van coëfficiënt alpha gespecificeerd. In voorwaarde (11.26) worden een ondergrens en een bovengrens van de moeilijkheidsgraad van de toetsen gespecificeerd. In voorwaarde (11.27) worden de onder- en bovengrens van de standaarddeviatie van de toetsen gespecificeerd. In voorwaarde (11.28) staat dat de toetsen niet dezelfde items mogen bevatten.

Het model 2 van Verschoor en Sanders beoogt parallelle toetsen samen te stellen met een zo hoog mogelijke betrouwbaarheid gegeven een bepaald aantal items de gemiddelde toetsscore en de standaarddeviatie. De formulering van model 2 luidt:

maximaliseer $\quad \text{minimum} \quad \alpha_t \qquad (11.29)$

onder voorwaarde dat $\quad \sum_{i=1}^{K} x_{it} = k, \qquad t = 1,...,T \qquad (11.30)$

$$\pi^{\ell} \leq \sum_{i=1}^{K} \pi_i \, x_{it} \leq \pi^{u}, \quad t = 1,...,T \qquad (11.31)$$

$$\sigma_x^{l} \leq \sigma_x \leq \sigma_x^{u} \qquad (11.32)$$

$$\sum_{t=1}^{T} x_{it} \leq 1 \qquad i = 1,...,K. \qquad (11.33)$$

Het maximaliseren van de betrouwbaarheden van parallelle toetsen staat in de doelfunctie (11.29). Dit doel wordt gerealiseerd door een maximinmodel, dat de betrouwbaarheid van de toets met de laagste betrouwbaarheid maximaliseert. In voorwaarde (11.30) wordt gespecificeerd dat de toetsen uit een vooraf bepaald gelijk aantal items dienen te bestaan. De betekenis van de voorwaarden (11.32) en (11.33) is gelijk aan die van de voorwaarden (11.27) en (11.28). Uiteraard is het ook bij de modellen mogelijk nog andere voorwaarden, bijvoorbeeld de verdeling van items over leerstofcategorieën, te specificeren.

Model 2 illustreren we hier voor het samenstellen van twee parallelle toetsen aan de hand van de reeds eerder gebruikte itembank met vijfhonderd rekenitems. Onze wensen specificeren we met het volgende model:

maximaliseer $\quad \text{minimum} \{ \alpha_1, \alpha_2 \}$

onder voorwaarde dat $\sum_{i=1}^{K} x_{i1} = \sum_{i=1}^{K} x_{i2} = 20$

$$10.0 \leq \sum_{i=1}^{K} \pi_i \, x_{it} \leq 11.0, \qquad t = 1, 2$$

$$\sum_{i=1}^{K} A_{mi} \, x_{it} = 5, \qquad t = 1, 2; \; m = 10, \ldots, 13$$

$$\sum_{t=1}^{2} x_{it} \leq 1, \qquad i = 1, \ldots, K.$$

In de doelfunctie van het model staat dat de betrouwbaarheden van de twee toetsen zo hoog mogelijk moeten worden. In de eerste voorwaarde wordt de eis geformuleerd dat de twee toetsen uit precies twintig items moeten bestaan. De tweede voorwaarde geeft de grenzen voor de moeilijkheidsgraad van de toetsen aan. In dit geval wordt gespecificeerd dat de gemiddelde toetsscore tussen de 10 en 11 scorepunten moet komen te liggen. Dat de twee toetsen vijf items uit elke leerstofcategorie dienen te bevatten, staat in de derde voorwaarde. In de vierde voorwaarde wordt geëist dat de twee toetsen niet dezelfde items mogen bevatten. De resultaten staan in tabel 11.3.

Tabel 11.3

Constructie van twee parallelle toetsen met model 2

Toets 1				Toets 2			
Item	p	r_{it}	Cat.	Item	p	r_{it}	Cat.
11	0.50	0.406	10	3	0.40	0.368	10
71	0.67	0.375	10	94	0.69	0.349	10
460	0.46	0.341	10	214	0.14	0.340	10
466	0.58	0.380	10	345	0.83	0.365	10
485	0.50	0.470	10	389	0.26	0.348	10
90	0.69	0.378	11	33	0.51	0.364	11
249	0.49	0.358	11	62	0.58	0.369	11
293	0.82	0.343	11	203	0.75	0.337	11
426	0.74	0.360	11	299	0.56	0.361	11
433	0.67	0.402	11	455	0.45	0.443	11
119	0.40	0.379	12	7	0.36	0.477	12
360	0.19	0.406	12	148	0.47	0.306	12
378	0.20	0.387	12	213	0.50	0.356	12
414	0.42	0.316	12	428	0.64	0.422	12
431	0.49	0.364	12	465	0.49	0.356	12
92	0.58	0.454	13	113	0.70	0.392	13
291	0.58	0.336	13	199	0.20	0.403	13
331	0.76	0.361	13	253	0.60	0.453	13
334	0.57	0.360	13	338	0.55	0.363	13
410	0.24	0.408	13	499	0.64	0.398	13
Gemiddelde score: 10.51				Gemiddelde score: 10.33			
α : 0.769				α : 0.769			
s_x: 4.04				s_x: 4.03			

Tabel 11.3 laat zien dat we er zeer goed in geslaagd zijn om twee parallelle toetsen samen te stellen die aan het model voldoen. De betrouwbaarheden zijn hoog en identiek, terwijl de gemiddelde scores en ook de standaarddeviaties van de toetsen nagenoeg gelijk zijn. Merk op dat er in het model geen voorwaarden voor de standaarddeviaties van de toetsen gespecificeerd werden. Ook wordt aan de voorwaarde voldaan dat er vijf items uit elke leerstofcategorie afkomstig moeten zijn. We zien dat de itemparameters binnen elke leerstofcategorie niet gelijk zijn en dat we dus zwak-parallelle toetsen samengesteld hebben.

11.4 Het samenstellen van toetsen in de generaliseerbaarheidstheorie.

In de bespreking van de generaliseerbaarheidstheorie (Cronbach et al., 1972) in hoofdstuk 3 werd een onderscheid gemaakt tussen een generaliseerbaarheidsstudie (G-studie) en een decisiestudie (D-studie). Hier laten we zien hoe de schattingen van variantiecomponenten uit een G-studie gebruikt kunnen worden in een D-studie om te bepalen hoeveel observaties, meestal items of vragen, er per meetobject, meestal een persoon, nodig zijn om de belangrijkste foutenbronnen te controleren of om een gewenste generaliseerbaarheids-coëfficiënt te realiseren.
Voor designs met één facet kan het minimum aantal observaties per persoon als volgt bepaald worden. In hoofdstuk 3 werd de betrouwbaarheidscoëfficiënt van een één-facet random-model gekruist design, ρ^2, gedefinieerd als:

$$\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{res}^2}{n_v}}, \tag{11.34}$$

waarbij σ_p^2 de variantiecomponent voor personen, σ_{res}^2 de variantiecomponent voor de persoon × facet *v* interactie plus de meetfouten, en n_v het aantal observaties of condities van facet *v* in de D-studie is. Wanneer we (11.34) herschrijven en voor ρ^2 een specifieke betrouwbaarheidscoëfficiënt nemen, dan is het minimum aantal observaties per persoon voor het realiseren van die specifieke coëfficiënt gelijk aan:

$$n_v = \frac{\rho^2\sigma_{res}^2}{\sigma_p^2 - \rho^2\sigma_p^2}. \tag{11.35}$$

Zowel (11.34) als (11.35) illustreren de Spearman-Brown formule uit de klassieke testtheorie: verhoging (verlaging) van het aantal items resulteert in een verhoging (verlaging) van de betrouwbaarheid.
De Spearman-Brown formule kan ook als optimaliseringsprobleem geschreven worden:

minimaliseer n_v (11.36)

onder voorwaarde dat $\rho^2 = g$. (11.37)

In dit optimaliseringsprobleem staat in de doelfunctie (11.36) dat het aantal items, n_v, geminimaliseerd moet worden. In voorwaarde (11.37) staat ρ^2 voor de

betrouwbaarheids-coëfficiënt en g voor de waarde van een specifieke betrouwbaarheidscoëfficiënt.
Aangezien de waarde van de doelfunctie, het aantal items, per definitie geheeltallig is, is het bovenstaande model geformuleerd als:

minimaliseer n_v (11.38)

onder voorwaarde dat $\rho^2 \geq g$, (11.39)

n_v geheeltallig. (11.40)

De opname van drempelvoorwaarde (11.39), een relaxatie van (11.37), en de geheeltalligheidseis (11.40) zorgen voor een oplossing met een geheeltallig aantal items. Vanwege dat laatste kunnen de vergelijkingen (11.38), (11.39) en (11.40) beschouwd worden als een generalisatie van de Spearman-Brown formule voor één-facet designs.

De Spearman-Brown formule, dat wil zeggen de samenhang tussen aantal observaties en betrouwbaarheid, geldt niet voor designs die uit verschillende facetten bestaan. We lichten dit toe aan de hand van het twee-facet random-model gekruist design. De generaliseerbaarheidscoëfficiënt voor dit design is gedefinieerd als:

$$\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pv}^2}{n_v} + \frac{\sigma_{pb}^2}{n_b} + \frac{\sigma_{res}^2}{n_v n_b}}, \qquad (11.41)$$

waarbij σ_p^2 de variantiecomponent voor personen, σ_{pv}^2 de variantiecomponent voor de persoon × facet v interactie, σ_{pb}^2 de variantiecomponent voor de persoon × facet b interactie, σ_{res}^2 de variantiecomponent voor de persoon × facet v × facet b interactie plus de meetfouten is, n_v en n_b de aantallen condities van respectievelijk facet v en facet b in de D-studie zijn. Het totale aantal observaties voor dit design wordt aangegeven met $L = n_v\ n_b$, het produkt van het aantal condities van de twee facetten. Aan formule (11.42) kunnen we zien dat het verhogen van bijvoorbeeld het aantal condities van een facet met een grote foutenvariantie een groter effect zal hebben op de generaliseerbaarheidscoëfficiënt dan het verhogen van het aantal condities van een facet met een kleine foutenvariantie. Met multi-facet designs is het dan ook mogelijk dat de generaliseerbaarheidscoëfficiënt verhoogd wordt terwijl het aantal observaties verlaagd wordt. Vanwege het multi-dimensionale karakter van de foutenvariantie in de generaliseerbaarheidstheorie, is het probleem van het bepalen van het minimum aantal

observaties veel complexer voor multi-facet designs dan voor één-facet designs. Sanders, Theunissen en Baas (1989) laten zien hoe dit probleem met behulp van een branch-and-bound algoritme kan worden opgelost. Hiervoor wordt het probleem eerst in termen van mathematische programmering geformuleerd als:

minimaliseer $L = n_v n_b$ (11.42)

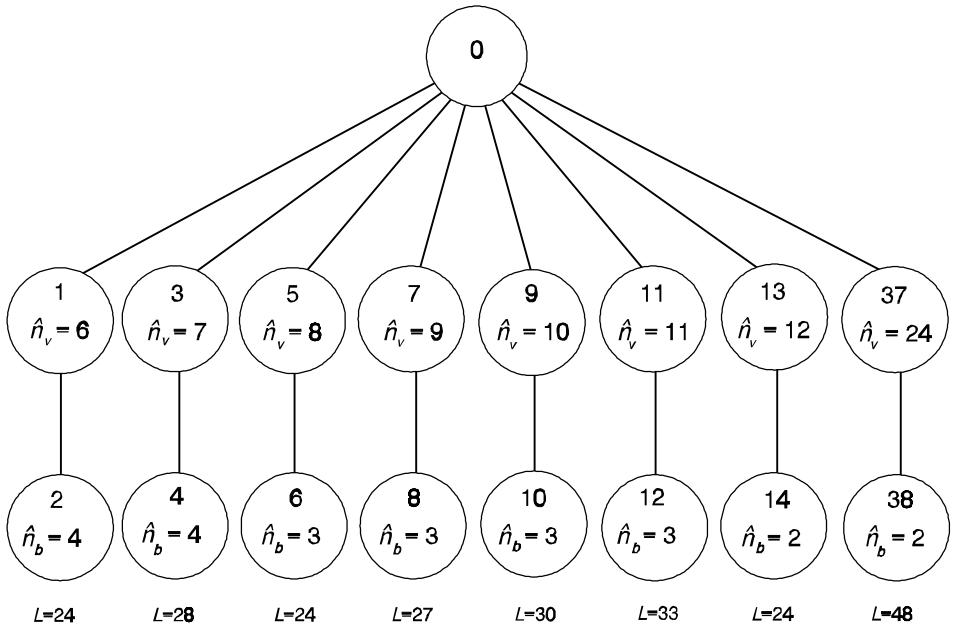
onder voorwaarde dat $\rho^2 \geq g$, (11.43)

$n_v \geq n_b$, (11.44)

n_v en n_b geheeltallig. (11.45)

In de formulering van dit optimaliseringsprobleem is L de waarde van de doelfunctie (11.42) als verschillende aantallen condities, n_v en n_b, voor facet v en facet b gebruikt worden. In de drempelvoorwaarde (11.43) staat ρ^2 voor de generaliseerbaarheidscoëfficiënt van een twee-facet random-model gekruist design en g voor de laagste waarde van de generaliseerbaarheidscoëfficiënt die als acceptabel beschouwd wordt. Voorwaarde (11.44) is een van de vele lineaire ongelijkheidsvoorwaarden die gebruikt kunnen worden. Voor het onderstaande geldt dat een optimale oplossing voor het twee-facet design probleem ook zonder deze voorwaarde verkregen kan worden. Het voordeel van een algoritme met deze voorwaarde is echter dat het irrelevante deel van de oplossingsruimte uitgesloten wordt, waardoor het aantal vertakkingen van de branch-and-bound gereduceerd wordt. In voorwaarde (11.45) staat dat mogelijke waarden voor n_v en n_b geheeltallig moeten zijn.

Nadat het probleem geformuleerd is als een optimaliseringsprobleem, worden grenzen geconstrueerd om het zoekproces te reduceren. In Sanders, Theunissen en Baas (1989) staat hoe die grenzen bepaald worden. Het zoekproces voor een twee-facet random-model gekruist design met $\hat{\sigma}^2_p = 5.435$, $\hat{\sigma}^2_{pv} = 3.421$, $\hat{\sigma}^2_{pb} = 1.140$, $\hat{\sigma}^2_{res} = 11.850$ en $g \geq .80$ staat in de zoekboom in figuur 11.9. De nummering van de knopen geeft aan hoe het zoekproces verloopt. De generaliseerbaarheidscoëfficiënten voor verschillende aantallen condities staan vermeld in tabel 11.4.



Figuur 11.9
Zoekboom van het twee-facet voorbeeld

De startoplossing $(\hat{n}_v = \hat{n}_b = 6)$ met een waarde voor de doelfunctie gelijk aan 36, is in knoop 2 van figuur 11.9 vervangen door een nieuwe 'beste' oplossing met de waarde 24. Oplossingen met dezelfde waarde voor de doelfunctie als knoop 2 zijn $(\hat{n}_v, \hat{n}_b) = (8,3)$ in
knoop 6 en $(\hat{n}_v, \hat{n}_b) = (12,2)$ in knoop 14. Het zoekproces eindigt in knoop 38 met de oplossing $(\hat{n}_v, \hat{n}_b) = (24,2)$ met de waarde 48 voor de doelfunctie die hoger is dan de tot dan toe beste oplossing. Op het eind van het zoekproces blijken er dus drie kandidaten voor een optimale oplossing te zijn: $(\hat{n}_v, \hat{n}_b) = (6,4)$, $(\hat{n}_v, \hat{n}_b) = (8,3)$ en $(\hat{n}_v, \hat{n}_b) = (12,2)$. Volgens tabel 11.4 zou $(\hat{n}_v, \hat{n}_b) = (8,3)$ als de optimale oplossing beschouwd kunnen worden omdat het in een hogere generaliseerbaarheidscoëfficiënt resulteert dan de andere oplossingen. Veelal zullen echter ook andere overwegingen dan het realiseren van een specifieke generaliseerbaarheidscoëfficiënt een rol spelen wanneer een meetinstrument geconstrueerd wordt. Als in het voorbeeld facet *v* items zouden zijn en facet *b* beoordelaars, dan zouden er aanzienlijke verschillen in de kosten per conditie van deze twee facetten bestaan. Omdat beoordelaars waarschijnlijk duurder zijn dan items, zal in het algemeen de voorkeur gegeven worden aan meer items en minder beoordelaars te nemen. Hiervoor dient de voorwaarde $n_v \geq n_b$ vervangen te

worden door een voorwaarde als $n_v \geq 5n_b$. Deze voorwaarde en de specificatie $g \geq 0.80$ geeft de optimale oplossing $(n_v, n_b) = (12, 2)$.

Tabel 11.4.
Waarden voor n_v, n_b, L, variantiecomponenten en ρ^2

n_v	n_b	L	$\hat{\sigma}_p^2$	$\frac{\hat{\sigma}_{pv}^2}{n_v}$	$\frac{\hat{\sigma}_{pb}^2}{n_b}$	$\frac{\hat{\sigma}_{res}^2}{n_v n_b}$	ρ^2
6	4	24	5.4	.57017	.285	.49375	.80166
6	6	36	5.4	.57017	.190	.32917	.83303
7	3	21	5.4	.48871	.380	.56429	.79135
7	4	28	5.4	.48871	.285	.42321	.81952
8	3	24	5.4	.42763	.380	.49375	.80681
9	3	27	5.4	.38011	.380	.43889	.81926
10	3	30	5.4	.34210	.380	.39500	.82951
11	3	33	5.4	.31100	.380	.35909	.83808
12	2	24	5.4	.28508	.570	.49375	.80117
24	2	48	5.4	.14254	.570	.24688	.84996
36	2	72	5.4	.09503	.570	.16458	.86757

Tabel 11.4 laat zien dat hoewel de verschillen tussen aantallen condities soms aanzienlijk zijn, de verschillen tussen de generaliseerbaarheidscoëfficiënten van die designs slechts gering zijn. Dat heeft te maken met de ongevoeligheid van hogere waarden van de coëfficiënt voor zelfs ingrijpende wijzigingen in het design. Let wel dat het verschil van slechts één conditie voor één facet een substantieel verschil kan betekenen voor te maken onderzoekskosten en dergelijke. Met een twee-facet gekruist design kan het verschil van één conditie betekenen dat één beoordelaar minder nodig is om bijvoorbeeld de antwoorden van honderd studenten op tien vragen te beoordelen.

12

Beoordelaarsovereenstemming

Vaak wordt bij het vaststellen van de mate waarin personen of objecten bepaalde kenmerken of eigenschappen bezitten, gebruik gemaakt van twee of meer terzake kundige beoordelaars die onafhankelijk van elkaar te werk gaan. In dergelijke gevallen nemen beoordelaars als het ware de plaats in van items of vragen in een toets of vragenlijst. Denk bijvoorbeeld aan de beoordeling van de kwaliteit van een scriptie, de beoordeling van een sportprestatie, de beoordeling van de geluidskwaliteit van stereo--apparatuur. Per beoordeelde eenheid beschikt men dan over twee of meer beoordelingen of scores. Hoewel te verwachten is dat beoordelaars niet altijd hetzelfde oordeel over een object geven, is bij grote verschillen tussen beoordelaars de bruikbaarheid van de beoordelingsprocedure twijfelachtig.

Wanneer ervaren radiologen aan de hand van röntgenfoto's de kwaadaardigheid van maagzweren beoordelen, blijkt in het algemeen dat ze lang niet altijd tot dezelfde conclusie komen (De Groot, 1966; Hofstee, 1981). Wanneer een patiënt door een arts wordt onder-zocht, is het gewenst dat diens bevindingen (diagnose, geconstateerde symptomen) niet anders luiden dan die van een andere arts die de patiënt onderzoekt. Verschillen tussen artsen impliceren dat in de praktijk sommige patiënten onnodig zullen worden geopereerd, terwijl andere patiënten een noodzakelijke, wellicht levensreddende, ingreep moeten ontberen.

In het onderwijs wordt de objectieve beoordeling van leerlingprestaties nagestreefd. Met objectief wordt bedoeld dat de uitkomst van de beoordeling slechts afhangt van de kwaliteit van de geleverde prestatie en dat ongeacht de beoordelaar hetzelfde beoordelingsresultaat wordt verkregen. Wanneer docenten echter opstellen Nederlands beoordelen, blijken voor één en hetzelfde opstel hun cijfers soms te verschillen van het cijfer 4 tot en met het cijfer 8. Dat betekent dat in examensituaties sommige leerlingen ten onrechte zakken of slagen.

Genoemde voorbeelden illustreren welke consequenties verschillen, of het gebrek aan overeenstemming tussen beoordelaars, kunnen hebben voor personen of objecten die beoordeeld worden. De voorbeelden geven tevens de relevantie aan van onderzoek

waarmee het mogelijk is (het gebrek aan) overeenstemming tussen beoordelaars, of de kwaliteit van beoordelingsprocedures te kwantificeren.

In paragraaf 12.1 van dit hoofdstuk wordt het begrip beoordelaarsovereenstemming gedefinieerd. De keuze van een maat voor beoordelaarsovereenstemming hangt af van het meetniveau van de data. In de paragrafen 12.2, 12.3 en 12.4 worden maten voor beoordelaarsovereenstemming bij data van respectievelijk nominaal, ordinaal en intervalniveau behandeld. In paragraaf 12.5 wordt een overzicht gegeven van mogelijke oorzaken voor lage beoordelaarsovereenstemming en remedies daarvoor. Tenslotte worden in paragraaf 12.6 nog een aantal andere ontwikkelingen aan de orde gesteld.

12.1 Definitie van beoordelaarsovereenstemming

Beoordelaars die oordelen geven, verrichten een beoordelingstaak. Deze taak kan opgevat worden als het classificeren van objecten. Daarmee wordt bedoeld het toewijzen van objecten aan beoordelingscategorieën op basis van een of meer -gepercipieerde- eigenschappen van die objecten. De categorieën in het eerder genoemde voorbeeld van de beoordeling van tumoren zijn bijvoorbeeld 'goedaardig', 'twijfelachtig', 'kwaadaardig'. Bij de beoordeling van prestaties van leerlingen in het onderwijs worden de categorieën gevormd door de bekende cijferschaal 1 tot en met 10. Bij beoordelingen veronderstellen we dus steeds een classificatie-schema dat een verzameling categorieën omvat. Beoordelaarsovereenstemming definiëren we als 'gelijkheid van classificatie' (Popping, 1983). De term gelijkheid in deze omschrijving is van fundamenteel belang. Daarmee wordt bedoeld dat de classificaties die door beoordelaars aan een object gegeven worden identiek zijn. We spreken van volledige overeenstemming tussen twee beoordelaars (ten aanzien van een object), als ze beiden het object toewijzen aan precies dezelfde categorie uit het classificatieschema. Deze (stringente) definitie impliceert dat alle beoordelaars beschikken over hetzelfde classificatieschema en dus niet de vrijheid hebben zelf hun beoordelingsschaal te kiezen.

12.2 Beoordelaarsovereenstemming bij data van nominaal niveau

Beoordelingsdata van nominaal niveau betreffen classificaties van personen of objecten in de zin van naamgeving of het toekennen van labels: 'katholiek', 'protestant', 'democraat', 'republikein', of 'CDA', 'VVD', 'D66'. Er moet gelden dat dergelijke categorieën in een classificatieschema wederzijds uitsluitend zijn: iemand kan dus niet

tegelijk protestant en katholiek zijn. Een ordening van de categorieën wordt niet verondersteld. Er kan niet worden gezegd dat 'protestant' meer of minder van 'iets' is dan 'katholiek'. Voor data van nominaal niveau bespreken we in deze paragraaf twee overeenstemmingsmaten: de proportie overeenstemming en de door Cohen (1960) voorgestelde coëfficiënt kappa.

Proportie overeenstemming

De proportie overeenstemming P_o is gedefinieerd als de verhouding van het aantal overeenstemmende oordelen en het totale aantal oordelen. Het percentage overeenstemming, $P_{\%}$, is gelijk aan $P_o \times 100$. De proportie overeenstemming wordt ook wel genoemd de ruwe (ongewogen) proportie overeenstemming. De proportie overeenstemming tussen twee beoordelaars, P_o, is gedefinieerd als:

$$P_o = \frac{\sum_{i=1}^{n} X_i}{n} \qquad (12.1)$$

waarin:
X_i = 0 als de twee beoordelaars het niet eens zijn over object *i*,
X_i = 1 als de twee beoordelaars het wel eens zijn over object *i*,
n = het aantal objecten dat door de twee beoordelaars wordt beoordeeld.

De proportie overeenstemming geeft dus de proportie van de gevallen aan waarin twee beoordelaars het eens zijn over de categorisering van objecten en deze toewijzen aan dezelfde categorie. Het voordeel van deze index is dat ze eenvoudig te begrijpen is en eenvoudig berekend kan worden. Ofschoon het een van de meest populaire overeenstemmingsmaten is, heeft de proportie overeenstemming helaas ook een belangrijk nadeel. Bij beoordelingen zal meestal, naar we aannemen, het toeval een rol spelen. In welke mate dat het geval is, is onbekend. Een beoordelaar vergist zich wel eens, verliest soms de concentratie, wordt even afgeleid, neemt zijn taak niet serieus, raakt vermoeid of is soms niet consequent. Daardoor zullen niet alle classificaties correct zijn. Het is dan ook aannemelijk dat (twee) beoordelaars soms bij toeval tot eenzelfde oordeel komen. Het nadeel van de proportie overeenstemming is (Bartko & Carpenter, 1976, p. 309) dat ze geen rekening houdt met wat wel toevals-overeenstemming wordt genoemd.

Toevalsovereenstemming is de proportie overeenstemmende oordelen die we op basis van toeval mogen verwachten. We lichten dit toe met twee voorbeelden. In het eerste voorbeeld wordt aan twee beoordelaars gevraagd *n* objecten te beoordelen op een driepuntsschaal. Zij doen dat, onafhankelijk van elkaar, maar nemen hun taak volstrekt niet serieus. Elk van hun scores (categorietoewijzingen) wordt dus geheel door het toeval bepaald en heeft niets met de eigenschap van de beoordeelde objecten te maken. In tabel 12.1 hebben we de classificaties van de twee beoordelaars samengevat. De negen cellen van tabel 12.1 bevatten proporties. De proportie objecten die door de eerste beoordelaar aan categorie 1 en door de tweede beoordelaar aan categorie 2 is toegewezen (.08), staat in de gearceerde cel 1,2. De diagonaal bevat de proportie gevallen waarin identieke oordelen zijn gegeven.

Tabel 12.1
Hypothetische proporties ter illustratie van toevalsovereenstemming

		Beoordelaar 2			
	Categorie	1	2	3	Totaal
Beoordelaar 1	1	.01	.08	.01	.10
	2	.08	.64	.08	.80
	3	.01	.08	.01	.10
	Totaal	.10	.80	.10	1.00

In dit fictieve voorbeeld zien we dat zelfs bij willekeurige toewijzing van objecten, uitsluitend en alleen op basis van toeval, een hoge proportie overeenstemming kan worden verkregen. De proportie ruwe overeenstemming is hier .66, namelijk de som van de proporties op de diagonaal van de tabel. Bij het optreden van toevalsovereenstemming (Popping, 1983, p. 25, Cohen, 1960, p. 38) speelt het aantal beschikbare beoordelingscategorieën een rol, alsmede de situatie waarin beoordelingscategorieën door beoordelaars moeilijk van elkaar zijn te onderscheiden (Schouten, 1985, p. XV).

In het tweede voorbeeld wordt aan twee andere beoordelaars gevraagd *n* objecten te beoordelen op een driepuntsschaal. Zij doen dat uiterst consciëntieus en hun toewijzing van objecten aan categorieën heeft uitsluitend betrekking op de eigenschap van de beoordeelde objecten. In tabel 12.2. vatten we de gegevens samen.

Tabel 12.2
Hypothetische proporties ter illustratie van overeenstemming

		Beoordelaar 4			
	Categorie	1	2	3	Totaal
Beoordelaar 3	1	.24	.13	.03	.40
	2	.05	.20	.05	.30
	3	.01	.07	.22	.30
	Totaal	.30	.40	.30	1.00

Bekijken we de diagonaal van overeenstemmingstabel 12.2, dan stellen we vast dat ook in dit geval de proportie overeenstemming uitkomt op .66, ofschoon we toch een beduidend ander beoordelaarsgedrag veronderstellen. We moeten dan ook concluderen dat de index 'proportie overeenstemming' geen rekening houdt met toevalsovereenstemming. De proportie toevals-overeenstemming wordt bepaald op basis van de marginale proporties. Tabel 12.3 geeft de verwachte celproporties gebaseerd op de marginale proporties in tabel 12.2 bij statistische onafhankelijkheid van beoordelaars. De waarde in de gearceerde cel 1.1 met waarde .12 wordt bijvoorbeeld verkregen als het product van de rij- en kolomtotalen: .40 × .30 = .12.
We zien in tabel 12.3 dat alleen al een proportie overeenstemming van .33, de som van de diagonaalcellen, te verwachten is op basis van de marginale proporties. Dat stelt de eerder gevonden proportie overeenstemming van .66 in tabel 12.2 in een ander licht.

Tabel 12.3
Verwachte celproporties bij onafhankelijkheid van beoordelaars

		Beoordelaar 4			
	Categorie	1	2	3	Totaal
Beoordelaar 3	1	.12	.16	.12	.40
	2	.09	.12	.09	.30
	3	.09	.12	.09	.30
	Totaal	.30	.40	.30	1.00

Resumerend stellen we vast dat de proportie overeenstemming weliswaar eenvoudig te bepalen is, maar als belangrijk bezwaar heeft dat ze geen rekening houdt met toevals-overeenstemming. Cohen (1960) heeft een index voorgesteld die aan dit probleem tegemoet komt.

Coëfficiënt kappa

Coëfficiënt kappa, κ, wordt algemeen aanbevolen als maat voor het bepalen van de overeenstemming tussen twee beoordelaars. Deze overeenstemmingsindex houdt rekening met toevalsovereenstemming en is toepasbaar bij zowel dichotome als polytome data van nominaal meetniveau. Kappa kan ook gegeneraliseerd worden naar situaties met meer dan twee beoordelaars. De berekening van κ veronderstelt dat de categorieën in het classificatieschema functioneel zijn. Daarmee wordt bedoeld dat het niet is toegestaan dat er categorieën in het schema voorkomen die door een beoordelaarspaar in het geheel niet worden gebruikt. Als dat het geval is dient het classificatieschema te worden herzien.

Coëfficiënt κ wordt, net als P_o in formule (12.1), berekend op basis van een zogenaamde overeenstemmingstabel waarin de classificaties van twee beoordelaars tegen elkaar worden afgezet. Een overeenstemmingstabel (zie ook tabel 12.1 en 12.2) bevat evenveel rijen als kolommen, namelijk *c*, het aantal beschikbare categorieën in het classificatieschema. De cellen bevatten proporties. Cel P_{ij} bevat de proportie objecten die door beoordelaar 1 aan categorie *i* en door beoordelaar 2 aan categorie *j* zijn toegewezen. De diagonaal bevat de proportie gevallen waarin identieke oordelen zijn gegeven. De algemene gedaante van een overeenstemmingstabel is gegeven in tabel 12.4.

Tabel 12.4
Overeenstemmingstabel

		Beoordelaar 2					
		1	2	.	*j*	.	*c*	
Beoordelaar 1	1	P_{11}	P_{12}				P_{1c}	P_1 .
	2	P_{21}						P_2 .
	.							.
	i				P_{ij}			P_i .
	.							.
	c	P_{c1}						P_c .
		$P_{\cdot 1}$	$P_{\cdot 2}$	.	$P_{\cdot j}$	.	$P_{\cdot c}$	*n*

De verschillende symbolen in tabel 12.4 hebben de volgende betekenis:

c = het aantal beoordelingscategorieën,
n = totaal aantal beoordeelde objecten (werkstukken, personen),
i = categorie-index voor beoordelaar 1, met $i = 1, ..., c$,
j = categorie-index voor beoordelaar 2, met $j = 1, ..., c$,
P_{ij} = proportie objecten toegewezen aan categorie i en j,
$P_{i.}$ = proportie objecten toegewezen aan categorie i,
$P_{.j}$ = proportie objecten toegewezen aan categorie j.

Om κ te berekenen moet voor de overeenstemmingstabel die men wil gebruiken gelden dat $n \geq 2$ en $c \geq 2$. Er moeten dus twee of meer objecten en twee of meer categorieën zijn. De berekening van κ is niet mogelijk wanneer zowel $P_{i.}$ als $P_{.j} = 0$ (met $i = j$), in welk geval een categorie in het classificatieschema niet wordt benut. Coëfficiënt kappa is gedefinieerd als:

$$\kappa = (P_o - P_e) / (1 - P_e). \qquad (12.2)$$

In (12.2) is de geobserveerde proportie overeenstemming, P_o, gedefinieerd als:

$$P_o = \sum_{i=1}^{c} P_{ii} .$$

Toevalsovereenstemming nulmodel is gedefinieerd als: $P_e = \sum_{i=1}^{c} P_{i.} P_{.i}$.
Coëfficiënt κ is een index voor beoordelaarsovereenstemming die, om Cohen (1960, p. 40) te citeren ..."the proportion of agreement after chance agreement is removed from consideration" weergeeft.

Keren we terug naar de overeenstemmingstabel 12.1 en we berekenen κ, dan vinden we $P_o = .66$ en $P_e = .66$, zodat $\kappa = (P_o - P_e) / (1 - P_e) = (.66 - .66)/(1 - .66) = 0 / .31 = 0$. Met andere woorden: alle waargenomen overeenstemming blijkt toevalsovereenstemming te zijn. Kijken we naar het eerder gegeven tweede voorbeeld, de serieuze beoordelaars in tabel 12.2 (en tabel 12.3) en we berekenen κ, dan vinden we $P_o = .66$ en $P_e = .33$, zodat $\kappa = (P_o - P_e) / (1 - P_e) = (.66 - .33) / (1 - .33) = .33 / .67 = .49$. De proportie overeen-stemming na correctie voor toevalsovereenstemming bedraagt dus .49. Uit de twee voorbeelden blijkt dus nog eens dat de proportie overeenstemming een onjuist beeld van de beoordelaarsovereenstemming kan geven.

De interpretatie van coëfficiënt kappa

Coëfficiënt κ is gelijk aan 1 bij perfecte overeenstemming. Een positieve waarde van κ geeft aan dat beoordelaars vaker met elkaar overeenstemmen dan op basis van toeval mag worden verwacht. Een κ van 0 geeft aan dat de mate van overeenstemming tussen beoordelaars gelijk is aan het kansniveau. Een negatieve waarde van κ geeft aan dat de beoordelaars minder vaak met elkaar overeenstemmen dan op basis van toeval kan worden verwacht, een κ van -1 wijst op een totaal gebrek aan overeenstemming tussen beoordelaars. In de literatuur wordt wel aangegeven dat een κ van .60 als een minimum moet worden beschouwd om van een acceptabele beoordelaarsovereenstemming te kunnen spreken, terwijl een κ waarde van .80 of hoger als 'goed' of 'bevredigend' wordt gekarakteriseerd (Dunn, 1989; Popping, 1983). Muskens (1980, p. 131) noemt deze grenswaarde van .80, een 'convention of the trade'. Landis en Koch (1977, p. 265) stelden het onderstaande, vaak geciteerde, overzicht op voor de interpretatie van κ.

κ			Interpretatie
<.00	<		'poor'
.00	-	.20	'slight'
.21	-	.40	'fair'
.41	-	.60	'moderate'
.61	-	.80	'substantial'
.81	-	1.00	'almost perfect'

Met betrekking tot de hoogte van coëfficiënt kappa moet opgemerkt worden dat het alleen bij gelijke marginale verdelingen in de overeenstemmingstabel mogelijk is dat kappa een maximum van 1.00 bereikt (Bartko & Carpenter, 1976, p. 314). Vandaar dat Dunn (1989, p. 38) voorstelt om bij de interpretatie de gevonden κ coëfficiënt te relateren aan de maximaal bereikbare κ, gegeven de randtotalen van de overeenstemmingstabel. Andere aspecten ten aanzien van de interpretatie van κ worden besproken door Umesh, Peterson en Sauber (1989).

Overeenstemming en associatie

In tabel 12.5 is geteld hoe twee beoordelaars honderd objecten toewijzen aan een van vier beschikbare nominale categorieën in een classificatieschema.

Tabel 12.5

Hypothetische frequenties van honderd objecten

		Beoordelaar 2				
	Categorie	1	2	3	4	Totaal
Beoordelaar 1	1	0	25	0	0	25
	2	0	0	0	25	25
	3	25	0	0	0	25
	4	0	0	25	0	25
	Totaal	25	25	25	25	100

De diagonaal in de tabel bevat alleen maar nullen, wat betekent dat het geen enkele keer voorkomt dat de twee beoordelaars een object aan dezelfde categorie toewijzen. Dit is een geval van perfecte niet-overeenstemming. Nochtans weten we dat als de eerste beoordelaar een object toewijst aan categorie 1, de tweede beoordelaar het object aan categorie 2 toewijst. Er is in dit geval sprake van perfecte samenhang of associatie. Perfecte associatie houdt in dat uit de categorie waaraan de ene beoordelaar het object toewijst, voorspeld kan worden aan welke categorie de andere beoordelaar het object toewijst. Voor één tabel kan dus gelden dat de associatie hoog is en de overeenstemming laag. Het omgekeerde geldt niet: indien er sprake is van overeenstemming geldt er ook associatie. In tabel 12.6 is er sprake van perfecte associatie, maar ook van perfecte overeenstemming.

Tabel 12.6

Hypothetische frequenties van honderd objecten

		Beoordelaar 2				
	Categorie	1	2	3	4	Totaal
Beoordelaar 1	1	25	0	0	0	25
	2	0	25	0	0	25
	3	0	0	25	0	25
	4	0	0	0	25	25
	Totaal	25	25	25	25	100

We zien in tabel 12.6 dat als we weten aan welke categorie de eerste beoordelaar een object toewijst, we ook weten aan welke categorie de tweede beoordelaar het object toewijst. We zien echter ook, dat anders dan in tabel 12.5, alle frequenties op de

diagonaal van de tabel liggen . Dat wil zeggen dat elk object door de twee beoordelaars aan dezelfde categorie (1, 2, 3 of 4) wordt toegewezen. Er is sprake van perfecte beoordelaarsovereenstemming.

Ofschoon tabel 12.5 perfecte niet-overeenstemming laat zien, wijst het voorkomen van associatie er op dat er toch een bepaalde samenhang is tussen de oordelen van de beoordelaars. Een nadeel van κ is dat alle gevallen van niet-overeenstemming gelijk worden behandeld omdat alleen naar de proporties op de diagonaal van de overeenstemmingsmatrix wordt gekeken. Daarom heeft Cohen (1968) een overeenstemmingsindex voorgesteld die aan dit bezwaar tegemoet komt. Deze index bespreken we in de volgende paragraaf.

12.3 Beoordelaarsovereenstemming bij data van ordinaal niveau

Beoordelingsdata van ordinaal meetniveau betreffen vaak beoordelingen naar de mate van aanwezig zijn van een eigenschap of kenmerk. Denk daarbij bijvoorbeeld aan Likertschalen, waarbij gegradeerde kwalificaties gegeven worden zoals 'slecht', 'matig', 'redelijk', 'voldoende', 'goed'. We spreken dan over een classificatieschema met geordende categorieën, waarbij overigens geen gelijke afstanden tussen de schaalpunten worden verondersteld. Deze ordening maakt het mogelijk rekening te houden met de mate van niet-overeenstemming. Daartoe maken we gebruik van het begrip gedeeltelijke of partiële overeenstemming. Twee beoordelaars die een object respectievelijk classificeren als 'voldoende' en 'goed' stemmen meer met elkaar overeen dan twee beoordelaars die een object beoordelen als respectievelijk 'slecht' en 'goed'.

Gewogen coëfficiënt kappa

Een maat voor beoordelaarsovereenstemming bij data van ordinaal meetniveau is de gewogen coëfficiënt kappa κ_w. Twee kenmerken van deze coëfficiënt zijn dat niet alleen gecorrigeerd wordt voor de mate van overeenstemming tussen beoordelaars die op basis van louter toeval verwacht kan worden, maar dat ook met partiële overeenstemming rekening wordt gehouden. Voor dat laatste wordt een gewichtenmatrix gebruikt. Een voorbeeld van een gewichtenmatrix staat in tabel 12.7.

Tabel 12.7

Gewichtenmatrix voor κ_w

	1	2	.	j	.	c
1	w_{11}	w_{12}				w_{1c}
2	w_{21}					
.						
i				w_{ij}		
.						
c	w_{c1}					

De symbolen in tabel 12.7 hebben de volgende betekenis:

c = het aantal beoordelingscategorieën,

i = categorie-index voor beoordelaar 1, met $i = 1, ..., c$,

j = categorie-index voor beoordelaar 2, met $j = 1, ..., c$,

w_{ij} = gewicht behorend bij toewijzingen aan categorie i en j.

De gewichten in de matrix moeten liggen tussen 0 en 1. Cellen die volledige overeenstemming representeren (gelijke classificaties) geven we het gewicht 1. Het gewicht 1 moet daarom altijd worden toegekend aan cellen die op de diagonaal van de matrix liggen, dus $w_{ii} = 1$. Het gewicht 0 wordt toegekend aan cellen die volledige niet-overeenstemming betreffen (classificaties die maximaal verschillen). Verder moet de gewichtenmatrix symmetrisch zijn ($w_{ij} = w_{ji}$) en er moet gelden $0 \leq w_{ij} \leq 1 = w_{ii}$.

Indien in de gewichtenmatrix alle cellen op de diagonaal het gewicht 1 bevatten en alle overige cellen het gewicht 0, is de gewogen coëfficiënt kappa gelijk aan κ. Coëfficiënt κ kan dan ook als een speciaal geval van κ_w opgevat worden. Beschouw nu tabel 12.8.

Tabel 12.8

Beoordeling door twee beoordelaars van werkstukken van vijf personen op een beoordelingsschaal (1 = matig; 2 = redelijk; 3 = uitstekend)

persoon	beoordelaar 1	beoordelaar 2
1	1	1
2	2	2
3	1	2
4	1	2
5	3	3

We geven nu eerst de bij deze tabel behorende overeenstemmingstabel 12.9.

Tabel 12.9

Overeenstemmingstabel van classificaties van twee beoordelaars van werkstukken van vijf personen

		Beoordelaar 2			
		1	2	3	
Beoordelaar 1	1	.20	.40	.00	.60
	2	.00	.20	.00	.20
	3	.00	.00	.20	.20
		.20	.60	.20	$n = 5$

De definitie van κ_w is: $\kappa_w = (P_o - P_e) / (1 - P_e)$ (12.3)

waarin $P_o = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij} P_{ij}$ de gewogen proportie overeenstemming is die we observeren

en $P_e = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij} P_{i\cdot} P_{\cdot j}$ de gewogen proportie toevalsovereenstemming is.

De bepaling van de gewichten in de buitendiagonale cellen van de gewichtenmatrix kan op verschillende manieren gebeuren. We noemen er drie. In de eerste methode krijgen (net als de diagonale cellen) bepaalde buitendiagonale cellen op inhoudelijke

gronden het gewicht 1, de andere het gewicht 0. Dit is het geval wanneer een onderzoeker bijvoorbeeld bij nader inzien van mening is dat categorieën met verschillende labels in feite toch hetzelfde kenmerk van een object representeren. Dit is equivalent aan een hercodering van de data, waarbij categorieën worden samengevoegd. Een voorbeeld van een op deze wijze opgestelde gewichtenmatrix bij overeenstemmingstabel 12.9 geeft tabel 12.10.

Tabel 12.10
Voorbeeld van een gewichtenmatrix van κ_w

	1	2	3
1	1.00	1.00	.00
2	1.00	1.00	.00
3	.00	.00	1.00

Hier zien we dat door de gewichtentoekenning in feite de categorieën 1 en 2 worden samengenomen. De tweede methode bestaat uit het via een algoritme bepalen van zogenaamde lineaire gewichten. Dergelijke gewichten, onder andere voorgesteld door Cicchetti (1972, p. 17), worden bepaald volgens de regel:

$$w_{ij} = 1 - \left\{ |i - j| \ / \ |c - 1| \right\}.$$

Het gewicht 1 wordt toegekend aan cellen die betrekking hebben op volledige overeenstemming, waarbij dus de twee beoordelaars een object aan dezelfde categorie toewijzen. Het gewicht 0 wordt toegekend aan die cellen waarbij de (scores van) twee beoordelingen maximaal verschillen. Toepassing van deze regel op tabel overeenstemmingstabel 12.9 geeft tabel 12.11.
Het lineair gewicht in de gearceerde cel w_{12} wordt berekend als

$$w_{12} = 1 - \left\{ |1 - 2| \ / \ |3 - 1| \right\} = 1 - (1 \ / \ 2) = .50.$$

Tabel 12.11
Voorbeeld van een matrix met lineaire gewichten

	1	2	3
1	1.00	.50	.00
2	.50	1.00	.50
3	.00	.50	1.00

Bij de derde methode worden zogenaamde kwadratische gewichten (Cohen, 1968) aan de buitendiagonale cellen toegekend. Een onderzoeker vindt bijvoorbeeld dat een relatief kleine afstand tussen beoordelaars als een behoorlijke mate van overeenstemming kan worden beschouwd, maar een grotere afstand nauwelijks meer mag meetellen. Kwadratische gewichten worden bepaald volgens de regel:

$$w_{ij} = 1 - \left\{(i-j)^2 / (c-1)^2\right\}.$$

Toepassing van deze regel op overeenstemmingstabel 12.9 geeft tabel 12.12.

Tabel 12.12
Voorbeeld van een matrix met kwadratische gewichten

	1	2	3
1	1.00	.75	.00
2	.75	1.00	.75
3	.00	.75	1.00

Het kwadratisch gewicht in de gearceerde cel w_{12} wordt berekend als

$$w_{12} = 1 - \left\{(1-2)^2 / (3-1)^2\right\} = 1 - (1 / 4) = .75.$$

We geven nu een voorbeeld van de berekening van κ_w waarbij gebruik wordt gemaakt van lineaire gewichten. Tabel 12.8 bevat de ruwe data voor twee beoordelaars die van vijf personen de kwaliteit van een werkstuk beoordeelden. Elk werkstuk is aan een van $c = 3$ beoordelingscategorieën toegewezen. Tabel 12.9 is de

overeenstemmingstabel en tabel 12.11 bevat de lineaire gewichten. De proportie gewogen overeenstemming, P_o, berekenen we als:

$$P_o = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij} P_{ij} = w_{11}P_{11} + w_{12}P_{12} + w_{13}P_{13} + w_{21}P_{21} + w_{22}P_{22} + w_{23}P_{23} + w_{31}P_{31} + w_{32}P_{32} + w_{33}P_{33} = .20 + .20 + .00 + .00 + .20 + .00 + .00 + .00 + .20 = .80.$$

De proportie gewogen toevalsovereenstemming P_e is:

$$P_e = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij} P_{i\cdot} P_{\cdot j} = .56.$$

De gewogen coëfficiënt kappa, κ_w, met lineaire gewichten, is gelijk aan:

$$\kappa_w = (P_o - P_e) / (1 - P_e) = (.80 - .56) / (1 - .56) = .24 / .44 = .55.$$

Merk op dat voor de data in tabel 12.9 de ongewogen coëfficiënt κ gelijk is aan .44, waarbij P_o = .60 en P_e = .28. Het is eenvoudig in te zien dat weging altijd leidt tot een waarde voor de overeenstemmingsindex die gelijk is aan of hoger is dan de ongewogen kappa. Zouden we kwadratische gewichten hebben toegepast, dan zou gewogen kappa .67 hebben bedragen, met P_o = .90 en P_e = .70.

Betrouwbaarheidsinterval voor kappa

De variantie van κ_w, $\sigma^2_{\kappa_w}$ (voor twee beoordelaars), is (Fleiss, Cohen & Everitt, 1969; Popping, 1983, 1992):

$$\frac{\sum_{i=1}^{c} \sum_{j=1}^{c} P_{ij} [(1 - P_e) w_{ij} - (1 - P_o)(w_{i\cdot} + w_{\cdot j})]^2 - (P_o P_e - 2P_e + P_o)^2}{n(1 - P_e)^4}$$

waarin $w_{i\cdot} = \sum_{j=1}^{c} w_{ij} P_{\cdot j}$ en $w_{\cdot j} = \sum_{i=1}^{c} w_{ij} P_{i\cdot}$.

Op basis van deze variantie kunnen de betrouwbaarheidsgrenzen voor kappa berekend worden. De betrouwbaarheidsgrenzen voor kappa geven aan binnen welke waarden kappa kan varieren, wanneer we het onderzoek met andere beoordelaars zouden herhalen. Deze grenzen worden bij benadering (Popping, 1989, p. 37) gegeven door

$$\left[\kappa_w \left(- z_{(1 - \frac{1}{2}\alpha)} \sigma_{\kappa_w} \right) , \quad \kappa_w \left(+ z_{(1 - \frac{1}{2}\alpha)} \sigma_{\kappa_w} \right) \right],$$

waarin $\sigma_{\kappa_w} = (\sigma^2_{\kappa_w})^{\frac{1}{2}}$ en z de standaard normale afwijking behorend bij gegeven significantie-niveau α is.

Coëfficiënt κ_w voor meer dan twee beoordelaars

Coëfficiënt κ_w is eenvoudig uit te breiden naar situaties dat er m beoordelaars zijn, met $m > 2$. In een situatie met meer dan twee beoordelaars zijn er $m(m - 1)/2$ oftewel $\binom{m}{2}$ paren beoordelaars die beschouwd kunnen worden. We kunnen dan bijvoorbeeld het gemiddelde van alle κ_w, $\overline{\kappa}_w$, berekenen van alle mogelijke paren beoordelaars. Popping (1983, p. 32) stelt echter voor te middelen bij het berekenen van P_o en P_e. Voor elk paar beoordelaars g en h worden dan $P_{o_{gh}}$ en $P_{e_{gh}}$ bepaald volgens formule (12.5). De gemiddelde gewogen kappa, $\overline{\kappa}_w$, is dan gelijk aan formule (12.3), met

$$P_o = \sum_{g=1}^{m-1} \sum_{h=g+1}^{m} P_{o_{gh}} \Big/ \binom{m}{2} \quad \text{en} \quad P_e = \sum_{g=1}^{m-1} \sum_{h=g+1}^{m} P_{e_{gh}} \Big/ \binom{m}{2}.$$

De variantie van $\overline{\kappa}_w$ voor meer dan twee beoordelaars is afgeleid door Popping (1983).

Aantal benodigde observaties

Cicchetti (1976) heeft onderzocht hoeveel observaties, in relatie met het aantal categorieën in het classificatieschema, vereist zijn om staat te kunnen maken op de berekende waarde voor kappa. Hij adviseert voor het aantal te beoordelen objecten: $n > 2c^2$, met c het aantal categorieën. Dus bij $c = 3$ beoordelingscategorieën moet het aantal observaties groter zijn dan 18 en bij $c = 7$ moet het aantal observaties groter zijn dan 98.

12.4 Beoordelaarsovereenstemming bij data van intervalniveau

Maten voor beoordelaarsovereenstemming bij data van intervalniveau zijn veelal gedefinieerd als ratio's van variantiecomponenten (zie ook hoofdstuk 3). In de literatuur (Haggard, 1958) worden dergelijke ratio's gewoonlijk aangeduid als intraklassecorrelatiecoëfficiënten. Shrout en Fleiss (1979) bespreken schattingen van intraklassecorrelatiecoëfficiënten voor drie soorten beoordelingssituaties. In deze paragraaf beperken we ons tot de meest voorkomende, namelijk de situatie waarbij een aselecte steekproef van objecten beoordeeld wordt door een aselecte steekproef van beoordelaars. Tabel 12.13 bevat de formele structuur van de datamatrix bij een dergelijk design.

Tabel 12.13
Datamatrix voor een gekruist design met twee factoren

		Beoordelaars					
		1	2	.	b	.	k	
Objecten	1	X_{11}	X_{12}				X_{1k}	$X_{1.}$
	2	X_{21}						$X_{2.}$
	.							.
	p				X_{pb}			$X_{p.}$
	.							.
	n	X_{n1}						$X_{n.}$
		$X_{.1}$	$X_{.2}$	.	$X_{.b}$	.	$X_{.k}$	$X_{..}$

In tabel 12.13 hebben de gebruikte symbolen de volgende betekenis:

k = aantal beoordelaars,
n = aantal beoordeelde personen of objecten,
p = index voor personen of objecten, met $p = 1, ..., n$,
b = index voor beoordelaars, met $b = 1, ..., k$,
X_{pb} = score voor object p van beoordelaar b,
$X_{p.}$ = somscore, over beoordelaars, voor object p,
$X_{.b}$ = somscore, over objecten, voor beoordelaar b,
$X_{..}$ = som van alle scores, over objecten en beoordelaars.

De beoordeling (score) van een persoon door een beoordelaar, X_{pb}, schrijven we als:

$$X_{pb} = \mu + (\mu_p - \mu) + (\mu_b - \mu) + (X_{pb} - \mu_p - \mu_b + \mu).$$

In dit lineaire model onderscheiden we naast het algemene gemiddelde μ, een persoonseffect, $\mu_p - \mu$, een beoordelaarseffect, $\mu_b - \mu$, en een residueel effect, $(X_{pb} - \mu_p - \mu_b + \mu)$. Elk van deze drie effecten of componenten heeft een variantie die we aanduiden met de term variantiecomponent.

Het schatten van variantiecomponenten

In hoofdstuk 3 is uiteengezet hoe de variantiecomponenten van een gekruist design met twee factoren geschat kunnen worden. In dat hoofdstuk is bij de berekening van de kwadratensommen uitgegaan van afwijkingsscores. Hier laten we zien dat we voor de berekening van kwadratensommen ook van de ruwe data kunnen uitgaan.

De totale kwadratensom, SS_{tot}, voor een gekruist design met twee factoren kan geschreven worden als:

$$SS_{tot} = SS_p + SS_b + SS_{res}$$

waarin:

$$SS_{tot} = \sum_{p=1}^{n}\sum_{b=1}^{k} X_{pb}^2 - \frac{X_{..}^2}{nk} \quad = \text{kwadratensom totaal}$$

$$SS_p = \frac{1}{k}\sum_{p=1}^{n} X_{p.}^2 - \frac{X_{..}^2}{nk} \quad = \text{kwadratensom personen}$$

$$SS_b = \frac{1}{n}\sum_{b=1}^{k} X_{.b}^2 - \frac{X_{..}^2}{nk} \quad = \text{kwadratensom beoordelaars}$$

$$SS_{res} = SS_{tot} - (SS_p + SS_b) \quad = \text{kwadratensom residu}$$

Door de kwadratensommen te delen door de vrijheidsgraden verkrijgen we de gemiddelde kwadratensommen:

$$MS_p = SS_p / (n-1) \quad = \text{gemiddelde kwadratensom personen}$$

$$MS_b = SS_b / (k-1) \quad = \text{gemiddelde kwadratensom beoordelaars}$$

$MS_{res} = SS_{res} / \{(n-1)(k-1)\}$ = gemiddelde kwadratensom residu.

De schattingen voor de variantiecomponenten zijn nu:

$\hat{\sigma}_p^2 = (MS_p - MS_{res}) / k$ = variantiecomponent personen

$\hat{\sigma}_b^2 = (MS_b - MS_{res}) / n$ = variantiecomponent beoordelaars

$\hat{\sigma}_{res}^2 = MS_{res}$ = variantiecomponent residu.

Beoordelaarsovereenstemmingscoëfficiënt

De beoordelaarsovereenstemmingscoëfficiënt, $\hat{\rho}^2$, voor k beoordelaars, is gedefinieerd als:

$$\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \left(\hat{\sigma}_b^2 + \hat{\sigma}_{res}^2\right) / k}. \qquad (12.4)$$

Wanneer de beoordelingen van verschillende beoordelaars perfect overeenstemmen, dus per beoordeelde persoon of object identiek zijn, dan zijn $\hat{\sigma}_b^2$ en $\hat{\sigma}_{res}^2$ gelijk aan nul en is de coëfficiënt gelijk aan 1. De variantiecomponent voor beoordelaars, $\hat{\sigma}_b^2$, geeft aan in welke mate beoordelaarsgemiddelden verschillen. Hoe lager de overeenstemming, des te groter de variantiecomponenten $\hat{\sigma}_b^2$ en $\hat{\sigma}_{res}^2$ zijn in verhouding tot $\hat{\sigma}_p^2$. Een relatief grote $\hat{\sigma}_b^2$ is minder bezwaarlijk dan een grote $\hat{\sigma}_{res}^2$ indien voor verschillen in gemiddelden gecorrigeerd kan worden. Bij volledig gebrek aan overeenstemming heeft de coëfficiënt de waarde nul.

In welke mate het aantal beoordelaars de mate van overeenstemming beïnvloedt, kan met (12.4) worden geschat door verschillende waarden van k, het aantal beoordelaars, in de noemer in te vullen. De coëfficiënt kan geïnterpreteerd worden als een schatting van de mate van overeenstemming tussen de gemiddelde beoordeling van k willekeurig gekozen beoordelaars en de gemiddelde beoordeling van k andere, eveneens willekeurig gekozen beoordelaars. Indien $k = 1$, dan is de coëfficiënt een schatting van de overeenstemming tussen de beoordelingen van één willekeurig gekozen beoordelaar en de beoordelingen van één andere, willekeurig gekozen beoordelaar. Indien $k = 2$, dan is de coëfficiënt een schatting van de gemiddelde overeenstemming tussen de gemiddelde beoordeling van twee beoordelaars en de gemiddelde beoordeling van twee andere, willekeurige beoordelaars. Formule (12.4) kan ook rechtstreeks in termen van gemiddelde kwadratensommen geschreven worden als:

$$\hat{\rho}^2 = \frac{MS_p - MS_{res}}{MS_p + (k-1)MS_{res} + k(MS_b - MS_{res})/n}.$$

Overeenstemming en betrouwbaarheid

In tabel 12.14 geven we twee fictieve voorbeelden van beoordelingen van werkstukken van tien leerlingen met behulp van een schoolcijferschaal.

Tabel 12.14

Hypothetische scores ter illustratie van verschillende niveaus van beoordelaarsovereenstemming en beoordelaarsbetrouwbaarheid

	Voorbeeld A			Voorbeeld B		
	Beoordelaar			Beoordelaar		
Werkstuk	1	2	3	4	5	6
1	1	3	5	1	1	1
2	1	3	5	2	2	2
3	2	4	6	3	3	3
4	2	4	6	3	3	3
5	3	5	7	4	4	4
6	3	5	7	5	5	5
7	4	6	8	6	6	6
8	4	6	8	7	7	7
9	5	7	9	8	8	8
10	5	7	9	9	9	9
$\bar{X}$ =	3.0	5.0	7.0	4.8	4.8	4.8
s_x =	1.5	1.5	1.5	2.7	2.7	2.7

In voorbeeld A zien we dat de drie beoordelaars steeds elk werkstuk of object een andere score geven. Van overeenstemming is dus geen sprake. We zien echter ook dat in de data een bepaald patroon zit. Beoordelaar 2 geeft steeds twee scorepunten meer dan beoordelaar 1, en beoordelaar 3 geeft steeds twee scorepunten meer dan beoordelaar 2. Het verschijnsel dat per object de scores, op een constante na, aan elkaar gelijk zijn, wordt additieve bias genoemd. De spreiding van de scores is voor elke beoordelaar gelijk. De scores van de drie beoordelaars correleren perfect met elkaar, dat wil zeggen dat elke beoordelaar tot dezelfde rangordening van werkstukken komt. In voorbeeld A is sprake van wat we perfecte beoordelaarsbetrouwbaarheid noemen. Beoordelaarsbetrouwbaarheid wordt gedefinieerd als:

$$\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_{res}^2 / k}. \qquad (12.5)$$

Formule (12.5) verschilt van formule (12.4) door het ontbreken van $\hat{\sigma}_b^2$, de variantiecomponent beoordelaars. Merk op dat (12.5) gelijk is aan de definitie van Cronbachs alpha (zie hoofdstuk 3).

In voorbeeld B zien we dat de drie beoordelaars steeds elk werkstuk dezelfde, identieke, score toekennen. De gemiddelde scores van de beoordelaars en ook de spreidingen zijn gelijk. Er is hier sprake van perfecte beoordelaarsovereenstemming. We zien ook dat de scores van de drie beoordelaars perfect correleren, dus perfect betrouwbaar zijn. De twee voorbeelden laten zien dat een hoge beoordelaarsbetrouwbaarheid een noodzakelijke, maar geen voldoende voorwaarde is voor een hoge beoordelaarsovereenstemming.

Samenvattingsopdracht Nederlands

Sanders, Hendrix en Luijten (1984) trokken in het kader van hun onderzoek naar het functioneren van globale en analytische beoordelingsschema's een aselecte steekproef van dertig leerlingen die bij het centraal schriftelijk eindexamen voor het vak Nederlands een samenvattingsopdracht hadden gemaakt. Een samenvattingsopdracht houdt in dat van een langere betogende tekst een sterk verkorte, maar adequate, samenvatting moet worden gemaakt van maximaal 500 woorden. Globale beoordelingsschema's omvatten niet meer dan enkele beknopte algemene richtlijnen voor de beoordelaars. In dit geval bijvoorbeeld onder andere de instructie dat beoordeeld moet worden of de samenvatting representatief is voor de oorspronkelijke tekst en gevolgd kan worden door een lezer die de oorspronkelijke tekst niet kent. Daarbij dient de beoordelaar zijn waardering rechtstreeks uit te drukken in een cijfer. Een analytische beoordelingsschema daarentegen geeft veel meer gedetailleerde aanwijzingen en vereist dat de beoordelaar per te beoordelen aspect, zoals tekststructuur, tekstlengte, inhoud en formulering een afzonderlijke beoordelingsscore toekent. Vervolgens worden de scores op de aspecten gewogen naar hun relatieve belang en daarna samengevat in een cijfer. De dertig samenvattingen werden door acht beoordelaars onafhankelijk van elkaar beoordeeld. Tabel 12.15 bevat de resultaten van de globale beoordeling van de acht beoordelaars (B1 - B8). We zien in tabel 12.15 dat het nogal wat uitmaakt door welke beoordelaar een leerling wordt beoordeeld. Leerling 3 krijgt van beoordelaar 3 het cijfer 2.0 en van beoordelaar 8 het cijfer 6.0. Over het geheel genomen oordelen

beoordelaars 1 en 5 wat milder, terwijl beoordelaar 6 en 7 als strenge beoordelaars gekenmerkt kunnen worden. Tabel 12.16 geeft de resultaten van de variantie-analyse voor de data in tabel 12.15.

Tabel 12.15

De globale beoordeling van dertig samenvattingen door acht beoordelaars

Leerling	B1	B2	B3	B4	B5	B6	B7	B8	Som
1	6.0	6.0	8.0	6.5	9.0	6.0	7.0	7.0	55.5
2	6.5	6.0	7.0	6.0	6.5	4.0	7.0	7.0	50.0
3	4.0	5.5	2.0	5.0	3.0	4.0	4.0	6.0	33.5
4	7.5	5.0	6.0	5.0	8.5	5.0	7.0	6.0	50.0
5	6.5	4.5	4.5	4.0	6.5	4.0	4.0	6.0	40.0
6	6.0	6.0	7.0	5.5	7.5	5.0	5.0	7.0	49.0
7	7.0	5.0	3.8	5.0	7.0	4.0	6.0	7.0	44.8
8	7.0	7.5	7.0	7.0	7.0	4.0	6.0	8.0	53.5
9	7.0	6.0	6.8	6.0	7.0	5.0	6.0	7.0	50.8
10	6.5	5.0	6.8	5.5	6.5	6.0	7.0	8.0	51.3
11	8.5	7.5	7.0	8.0	10.0	7.0	5.0	9.0	62.0
12	8.0	6.0	7.5	5.5	7.5	6.0	3.0	7.0	50.5
13	7.5	6.0	6.5	6.0	7.5	7.0	6.0	6.0	52.5
14	6.0	6.0	7.0	5.5	5.0	6.0	5.0	6.0	46.5
15	8.0	6.0	6.5	6.0	6.5	6.0	3.0	6.0	48.5
16	6.5	7.0	6.5	6.5	7.0	5.0	3.0	5.0	46.5
17	9.0	5.0	7.0	5.5	7.5	4.0	7.0	7.0	52.0
18	7.5	6.0	8.0	6.5	6.5	5.0	6.0	5.0	50.5
19	7.0	5.0	6.0	5.0	8.0	6.0	5.0	6.0	48.0
20	4.0	6.5	4.0	6.0	4.5	5.0	3.0	4.0	37.0
21	4.0	6.0	3.0	6.0	4.0	5.0	4.0	4.0	36.0
22	6.0	6.0	7.0	5.5	7.5	7.0	8.0	5.0	52.0
23	4.0	4.0	5.0	4.0	4.0	5.0	7.0	6.0	39.0
24	6.5	6.0	7.0	6.5	7.5	6.0	8.0	6.0	53.5
25	7.5	6.0	8.0	6.0	5.0	5.0	6.0	4.0	47.5
26	8.0	7.5	7.5	7.0	7.0	6.0	7.0	6.0	56.0
27	5.0	4.0	4.5	3.0	6.0	3.0	3.0	5.0	33.5
28	3.0	5.0	1.0	5.0	3.0	3.0	5.0	3.0	28.0
29	5.0	4.5	6.0	4.0	5.0	4.0	6.0	5.0	39.5
30	4.0	5.5	4.0	5.0	4.0	3.0	5.0	4.0	34.5
Som	189	172	177.9	168	191.5	151	154	178	1391.4

In tabel 12.16 zien we dat de residuele component de grootste variantiecomponent is. De variantiecomponent beoordelaars daarentegen is relatief gering.

Tabel 12.16

Resultaten van de variantie-analyse voor de gegevens van de globale beoordeling van dertig werkstukken door acht beoordelaars

Effecten	Vrijheids-graden	Kwadraten-sommen	Gemiddelde kwadratensommen	Schattingen van variantiecomponenten
Personen (*p*)	29	236.31	8.15	$\hat{\sigma}^2_p$ = .876 (40%)
Beoordelaars (*b*)	7	41.05	5.86	$\hat{\sigma}^2_b$ = .157 (7%)
Residu (*res*)	203	231.92	1.14	$\hat{\sigma}^2_{res}$ = 1.143 (53%)

De beoordelaarsovereenstemmingscoëfficiënt voor $k = 8$ beoordelaars is gelijk aan:

$$\hat{\rho}^2 = \frac{.876}{.876 + (.157 + 1.143)/8} = .84 .$$

Het doel van het gebruik van beoordelingsschema's is het realiseren van een objectieve beoordeling. Dat wil zeggen dat we ernaar streven een beoordelingsschema te maken dat een zo hoog mogelijke beoordelaarsovereenstemming oplevert bij zo weinig mogelijk beoordelaars. Het zou ideaal zijn om in de beoordelingsprocedure slechts één beoordelaar in te hoeven inschakelen. In de praktijk zijn acht beoordelaars overigens meestal niet beschikbaar of betaalbaar. De geschatte overeenstemming voor het geval dat de samen-vattingen zouden worden beoordeeld door één beoordelaar is:

$$\hat{\rho}^2 = \frac{.876}{.876 + (.157 + 1.143)/1} = .40.$$

Een overeenstemmingscoëfficiënt van .40 betekent dat indien de werkstukken door één willekeurig gekozen beoordelaar beoordeeld worden, en deze beoordelingsscores zouden vergeleken worden met de scores van één andere willekeurige beoordelaar, we grote scoreverschillen zullen zien. In tabel 12.17 worden schattingen gegeven voor de overeenstemming bij gebruik van diverse aantallen beoordelaars.

In het genoemde onderzoek (Sanders et al., 1984) bleek dat met een analytisch beoordelingsschema een hogere beoordelaarsovereenstemming kon worden bereikt dan met een globaal beoordelingsschema. Bij een analytische, onafhankelijke beoordeling van samen-vattingen door twee beoordelaars kon dezelfde overeenstemming worden bereikt als met een globale beoordeling door drie onafhankelijke beoordelaars. Het behoeft geen betoog dat een beoordelingsprocedure waarin bij gelijkblijvende kwaliteit

van de beoordeling met minder beoordelaars kan worden volstaan, uit logistiek en kostenoogpunt de voorkeur verdient.

Betrouwbaarheidsinterval voor de overeenstemmingscoëfficiënt

De overeenstemmingscoëfficiënt $\hat{\rho}^2$ die we berekenen is een schatting. Bij replicaties van het onderzoek met andere steekproeven van kandidaten en beoordelaars verwachten we niet dezelfde resultaten te vinden. Het is daarom van belang het betrouwbaarheidsinterval voor de overeenstemmingscoëfficiënt ρ^2 te berekenen.

De methode voor het bepalen van een dergelijk betrouwbaarheidsinterval voor de overeenstemmingscoëfficiënt is ontleend aan Fleiss en Shrout (1978, 1979). Het betrouwbaarheidsinterval kan als volgt benaderd worden. Het aantal vrijheidsgraden, v, is gelijk aan:

$$v = \frac{(k-1)(n-1)\left\{k\hat{\rho}^2 F_b + n\left[1+(k-1)\hat{\rho}^2\right] - k\hat{\rho}^2\right\}}{(n-1)k^2\hat{\rho}^{2^2}F_b{}^2 + \left\{n\left[1+(k-1)\hat{\rho}^2\right] - k\hat{\rho}^2\right\}}.$$

In bovenstaande formule is $F_b = MS_b / MS_{res}$. Als we nu uit de *F*-verdeling de waarden definiëren $F^* = F_{1-½\alpha}[(n-1), v]$ en $F_* = F_{1-½\alpha}[v, (n-1)]$, dan zijn de grenzen van het $(1-\alpha)\times 100\%$ betrouwbaarheidsinterval voor ρ^2:

$$\left(\frac{n(MS_p - F^* MS_{res})}{F_*[kMS_b + (kn-k-n)MS_{res}]nMS_p} \quad , \quad \frac{n(F_* MS_p - MS_{res})}{kMS_b + (kn-k-n)MS_{res} + nF_* MS_p}\right).$$

Het minimum aantal beoordelaars

De ondergrens van het betrouwbaarheidsinterval van ρ^2 is richtinggevend voor het antwoord op de vraag hoeveel beoordelaars minimaal nodig zullen zijn om in vervolgsituaties, dus bij hernieuwd beoordelen (andere kandidaten, andere beoordelaars), een bepaalde zekerheid te hebben over de te verwachten beoordelaarsovereenstemming. We zullen dat hier aan de hand van het voorbeeld van de samenvattingsopdracht Nederlands toelichten. De beoordelaarsovereenstemming voor acht beoordelaars bedroeg .84, terwijl de grenzen voor het 90% betrouwbaarheidsinterval bij benadering .76 en .91 zijn. Stel nu dat een onderzoeker aanbevelingen wil doen voor toepassing in de praktijk van de onderzochte beoordelings-

procedure, maar bijvoorbeeld, mede gelet op het kostenaspect, tevreden zou zijn met een beoordelaarsovereenstemming van .60. De beoordelaarsovereenstemmings-coëfficiënt en de daarbij geschatte betrouwbaarheidsintervallen bij verschillende aantallen beoordelaars staan in tabel 12.17. Het betreft hier opnieuw de gegevens voor de globale beoordeling van dertig samenvattingen door acht beoordelaars.

Tabel 12.17

Schattingen van de beoordelaarsovereenstemming bij diverse aantallen beoordelaars en de grenzen voor een 90% betrouwbaarheidsinterval

Aantal beoordelaars	Beoordelaars-overeenstemming	Intervalgrenzen 90% betrouwbaarheidsinterval
1	.40	.29 - .55
2	.57	.44 - .71
3	.67	.55 - .78
4	.73	.62 - .83
5	.77	.67 - .86
6	.80	.71 - .88
7	.83	.74 - .89
8	.84	.76 - .91

Inspectie van tabel 12.17 leert dat bij vier beoordelaars het interval tussen .62 en .83 ligt.
Op grond hiervan kan de conclusie worden getrokken dat voor de beoordeling van een nieuwe reeks objecten kan worden volstaan met een beoordeling door vier beoordelaars.

12.5 Lage beoordelaarsovereenstemming: oorzaken en remedies

Oorzaken

Er zijn diverse factoren denkbaar die de beoordelaarsovereenstemming nadelig beïnvloeden. Saal, Downey en Lahey (1980) geven een overzicht en merken op dat er weinig overeenstemming schijnt te bestaan over de conceptuele definities met betrekking tot de criteria voor de kwaliteit van beoordelingen en over operationele definities voor die criteria. We kunnen een onderscheid maken tussen niet-systematische en systematische invloeden. Niet-systematisch noemen we toevallige en fluctuerende invloeden op de beoordelaar en diens beoordeling. We kunnen hierbij

denken aan vermoeidheid, schrijffouten, telfouten, onoplettendheid, verstoringen van de beoordeling door lawaai en temperatuur. Systematische invloeden maken dat de beoordelingen van een beoordelaar op een systematische manier afwijken van de beoordelingen die andere beoordelaars geven.

Een bekende systematische afwijking is 'restriction of range'. Hiervan is sprake wanneer sommige beoordelaars niet alle beschikbare categorieën in een classificatieschema benutten. Twee bekende vormen hiervan zijn mildheid en centrale tendentie. Van mildheid is sprake wanneer beoordelaars de neiging hebben relatief lage of juist relatief hoge scores te geven. Zo geven sommige docenten nooit cijfers hoger dan 8 en anderen nooit cijfers lager dan 4, ongeacht het bereik van de schoolcijferschaal of de prestaties van hun leerlingen. Saal et al. (1980) geven drie operationele definities voor dit effect. Sommige beoordelaars neigen ertoe geen expliciete uitspraken te willen doen. Ze vermijden extreem geformuleerde categorieën en zitten met hun beoordelingen steeds rond het midden van de beoordelingsschaal. Dit verschijnsel wordt wel centrale tendentie genoemd.

We spreken van een halo-effect wanneer beoordelaars hun oordeel mede laten afhangen van voor de meting niet terzake doende kenmerken van degene die beoordeeld wordt of van diens product, zoals uiterlijk, kleding of de netheid van het handschrift. Zo valt de beoordeling van een prestatie of werkstuk van een vriendelijk en beleefd persoon soms hoger uit dan de beoordeling van een prestatie van een persoon die in dit opzicht afwijkt van wat de beoordelaar als normaal beschouwt. Saal et al. (1980) beschrijven het halo-effect als het onvermogen van een beoordelaar om onderscheid te maken tussen verschillende aspecten van het gedrag van de persoon die beoordeeld wordt. Ze presenteren daarbij overigens vier verschillende operationele definities.

De neiging van een beoordelaar om zich in de strengheid van zijn beoordelingen aan te passen aan het gemiddelde niveau van de te beoordelen objecten staat bekend als normverschuiving. Hoe goed of hoe slecht een schoolklas als geheel ook is voor een bepaald vak, vaak zien we dat de percentages onvoldoendes bij elke klas voor een vak gelijk zijn.

Van een sequentie-effect spreken we wanneer de beoordeling die de beoordelaar aan een object geeft mede tot stand komt op basis van de nawerking van een beoordeling die net tevoren is gegeven. De middelmatige prestatie van een leerling die wordt beoordeeld net nadat een of meer zeer slecht presterende leerlingen zijn beoordeeld, wordt dan hoger gescoord dan in het omgekeerde geval, wanneer de beoordeling van een middelmatige leerling zou volgen op de beoordeling van een of meer excellente leerlingen.

Als laatste noemen we het signifisch effect. Hiervan is sprake wanneer beoordelaars de beoordelingstaak verschillend opvatten, omdat ze de nadruk leggen op verschillende aspecten. Bij de beoordeling van het opstel zien we bijvoorbeeld dat sommige docenten meer op stijl letten, anderen op inhoud, weer anderen op structuur, terwijl de ene docent spel- en schrijffouten in de beoordeling betrekt en de andere docent weer niet.

Remedies

Constateert men een te lage beoordelaarsovereenstemming, dan zijn er verschillende manieren om er voor te zorgen dat bij herhaling van de beoordelingsprocedure betere resultaten te verwachten zijn. Bepaalde maatregelen zijn eveneens mogelijk indien herhaling van de beoordelingsprocedure niet mogelijk is. Dit laatste betreft dan met name correcties op basis van aanwijsbare systematische fouten, zoals mildheid.

Dat het inzetten van meer beoordelaars de beoordelaarsovereenstemming kan verhogen is in het voorgaande al uitvoerig besproken. Merk echter op dat ook hier de wet van de verminderende meeropbrengst van toepassing is: de winst die elke toegevoegde beoordelaar oplevert in termen van verbetering van de overeenstemming begint op een gegeven ogenblik af te nemen, meestal na twee of drie beoordelaars.

Een duidelijke verbetering van de beoordelaarsovereenstemming kan worden verwacht wanneer beoordelaars worden getraind voor hun taak, bijvoorbeeld door met hen enkele proefbeoordelingen te doen en deze te bespreken. Van de proefobjecten moet bij voorkeur het resultaat bekend zijn van een standaardbeoordeling, zodat de beoordelaars hun eigen beoordelingsscores met deze standaard kunnen vergelijken.

Men dient er voor te zorgen dat beoordelaars werkelijk onafhankelijk van elkaar werken. Overleg tussen beoordelaars gedurende de uitvoering van de beoordelingstaak draagt het risico in zich dat oneigenlijke factoren (dominantie, senioriteit, status, argumentatievermogen) het overleg en daarmee de meting beïnvloeden.

Belangrijk is ook een merkbare controle op het werk van de beoordelaars. Indien beoordelaars weten dat hun werk wordt gecontroleerd, zullen ze zich minder afwijkingen van het beoordelingsschema en de bijbehorende instructies veroorloven. In veel beoordelingssituaties komt het voor dat op een of andere wijze de beoordelaars belang hebben bij de uitslag van de beoordeling.

Beoordelaarsovereenstemming is ook afhankelijk van de kwaliteit van beoordelaars-instructies. Gezorgd dient te worden voor duidelijke en hanteerbare beoordelaarsinstructies die, bijvoorbeeld bij een beoordelaarstraining, met de beoordelaars besproken worden. Beoordelaarsinstructies hebben bijvoorbeeld betrekking op de volgorde waarin objecten worden beoordeeld, de inrichting van de beoordelingssituatie (plaats, licht, geluid), op zaken zoals 'geen aantekeningen maken op schriftelijke werkstukken' om een mogelijke tweede beoordelaar niet te beïnvloeden. Zorg daarnaast voor een helder en functioneel classificatie-schema, zodanig dat alle beoordelaars op dezelfde wijze begrijpen wat de erin voorkomende categorieën betekenen. Beperk het aantal categorieën tot maximaal zeven (James et al., 1984; Cicchetti, 1976). Belangrijk is een duidelijk scoringsvoorschrift, dat wil zeggen een overzicht van het aantal scorepunten dat gegeven dient te worden aan bijvoorbeeld een goed, een minder goed en een fout antwoord. Geef bij globale of holistische beoordelingen een overzicht waarin wordt aangegeven op welke beoordelingsaspecten gelet moet worden. Gebruik waar mogelijk analytische beoordelingsschema's. Overweeg om beoordelaars die extreem afwijkende scores te zien geven te verwijderen uit de groep beoordelaars die bij de beoordeling wordt betrokken. Is van een beoordelaar systematisch afwijkend beoordelaarsgedrag bekend, met name mildheid of strengheid, overweeg dan aanpassing van diens scores.

12.6 Tot besluit

In de beoordelingssituaties die we in dit hoofdstuk beschreven hebben, had overeenstemming altijd betrekking op overeenstemming tussen beoordelaars. Overeenstemming tussen beoor delaars wordt in de literatuur vaak aangeduid als interbeoordelaarsovereenstemming. In beoordelingssituaties waarbij één beoordelaar een reeks personen of objecten op twee verschillende tijdstippen beoordeelt, kunnen we de overeenstemming tussen de scores op de twee tijdstippen uitrekenen. In dat geval spreken we over intrabeoordelaarsovereenstemming. Wanneer er sprake is van beoordelingssituaties waarbij de overeenstemming berekend wordt tussen beoordelaars en een standaard, spreken we van accuraatheid (Suen & Ary, 1989). Deze term is ontleend aan onderzoek dat in de exacte disciplines plaatsvindt en waarbij 'echte' standaarden worden gebruikt. Zo kan de overeenstemming berekend worden tussen metingen met verschillende duimstokken ('beoordelaars') die in de handel zijn en de 'echte' meetlat of standaard. Een hoge overeenstemming tussen een bepaalde duimstok en de standaard betekent dat die duimstok valide is voor het meten van lengte. In de

sociale wetenschappen is het soms mogelijk om voor bepaalde beoordelingssituaties standaards te gebruiken, bijvoorbeeld de oordelen van enkele deskundige beoordelaars aan wiens oordeel niet getwijfeld kan worden. Het gebruik van een standaard heeft als voordeel dat beoordelingssituaties vermeden worden waarbij we een hoge beoordelaarsovereenstemming vinden terwijl de groep beoordelaars collectief verkeerd beoordeeld heeft.

De bespreking van de overeenstemmingscoëfficiënten bij data van intervalniveau beperkte zich in de vorige paragraaf tot een design met twee factoren. In paragraaf 3.13 van hoofdstuk 3 is een gekruist design met drie factoren, in de generaliseerbaarheidstheorie een design met twee facetten genoemd, besproken. Daar zagen we dat in een gekruist design met drie factoren behalve de score X_{pvb}, de score die persoon *p* voor het antwoord op vraag *v* van beoordelaar *b* ontvangen heeft, zes gemiddelde scores onderscheiden worden. Twee voorbeelden zijn de gemiddelde score van vraag *v* (gemiddeld over alle personen en alle beoordelaars) en de gemiddelde score van beoordelaar *b* (gemiddeld over alle personen en alle beoordelaars). Overeenstemmingscoëfficiënten voor designs met drie factoren zijn afgeleid door Maxwell en Pilliner (1968). Hun afleiding is gebaseerd op het concept 'replicatie van het experiment'. Dit concept gebruikt ook Mellenbergh (1977) bij zijn afleiding van wat hij replicatiecoëfficiënten noemt. Een replicatiecoëfficiënt is gedefinieerd als de correlatie tussen bijvoorbeeld de gemiddelde score van beoordelaar(s) *b* bij een beoordelingsprocedure of 'experiment' en de gemiddelde score van beoordelaar(s) *b* bij een herhaling of replicatie van de beoordelingsprocedure. Een replicatiecoëfficiënt kan geschreven worden als een ratio van variantiecomponenten. Hoe variantiecomponenten geschat kunnen worden, is uitgebreid in hoofdstuk 3 beschreven. Voor een gekruist design met drie factoren kunnen in totaal 19 replicatiecoëfficiënten geschat worden. Voor details verwijzen we naar het artikel van Mellenbergh (1977, p. 380).

In de praktijk komt het regelmatig voor dat niet alle beoordelaars alle personen (kunnen) beoordelen. Behalve het weglaten van de objecten met ontbrekende scores, bespreekt Popping (1983, p. 46) nog andere methoden om in zulke gevallen kappa te berekenen. Voor data van intervalniveau hebben Houston, Raymond en Svec (1991) een drietal methoden ontwikkeld voor het schatten van beoordelaarseffecten in het geval dat beoordelingen ontbreken. Hierdoor is het toch mogelijk te corrigeren voor verschillen in strengheid van beoordelaars. De methoden zijn verwant aan de methoden die in hoofdstuk 7 besproken zijn. Van belang is op te merken dat statistische pakketten (bijvoorbeeld Dixon, 1992) tegenwoordig programma's bevatten waarmee variantiecomponenten van incomplete gegevensverzamelingen geschat kunnen worden.

In dit hoofdstuk hebben we ons beperkt tot overeenstemmingscoëfficiënten die hun bruikbaarheid bewezen hebben. Daarnaast zijn er de laatste jaren nog vele andere overeenstemmingscoëfficiënten voorgesteld. Zegers (1991) bespreekt de eigenschappen van zogenaamde associatiecoëfficiënten. Uebersax (1991) laat zien dat het ook mogelijk is beoordelaarsovereenstemming te modelleren en te berekenen met behulp van latente klassen-modelllen, loglineaire modellen, itemresponsmodellen, correspondentie- en homogeniteits-analyse.

13

Schalen, normen en cijfers

Een toets hoort te worden afgesloten met een heldere en duidelijke presentatie van het toetsresultaat, die de ruimte voor misverstanden tot een minimum beperkt. In het voorliggende hoofdstuk worden manieren besproken waarmee dit doel dichterbij kan worden gebracht. We gaan we ervan uit dat de antwoorden van de persoon op de items op papier zijn gecodeerd als itemscores en dat we dus de beschikking hebben over een vector van itemscores, ook kortweg antwoordpatroon genoemd. Zolang het antwoordpatroon niet expliciet met een of enkele kwaliteitsoordelen is samengevat, is het antwoordpatroon op zich niet erg informatief over het niveau van de geleverde prestatie. Zo'n samenvattend kwaliteitsoordeel noemen we een schaalwaarde of liever een cijfer. Het cijfer moet snel een zo nauwkeurig mogelijke indruk geven van het niveau van het resultaat. Voor een correcte interpretatie van het cijfer moet het natuurlijk duidelijk zijn waarvoor de toets valide is. De validiteit van toetsscores is eerder afdoende aan de orde geweest, zodat in de volgende vijf paragrafen aandacht kan worden geschonken aan andere aspecten van het rapporteren van toetsresultaten. In paragraaf 13.1 wordt het schaalniveau van cijfers behandeld. Het schaalniveau van de cijfers, bijvoorbeeld ordinaal of interval, moet worden vermeld, en verantwoord, om te voorkomen dat er onjuiste conclusies aan cijfers worden verbonden. Men kan niet volstaan met alleen te vermelden welk schaalniveau de cijfers hebben. Ook aan de manier waarop dit schaalniveau is bereikt en met welke veronderstellingen hoort aandacht te worden besteed. In paragraaf 13.2. behandelen we cijfers waarmee het niveau van de prestatie gemakkelijk kan worden vergeleken met prestaties in een of meer groepen. In paragraaf 13.3 behandelen we beheersingsschalen. Dit zijn cijferschalen waarmee het niveau van een prestatie wordt weergegeven als de mate waarin een vaardigheid wordt beheerst. De nauwkeurigheid van het cijfer kan op meerdere manieren in de rapportage worden verwerkt. In paragraaf 13.4 worden daarvoor enige suggesties gedaan. Het nemen van beslissingen op grond van cijfers is het onderwerp van paragraaf 13.5. De manier waarop dit gebeurt moet in de rapportage worden verantwoord. Bij de beslissing of een leerling slaagt of zakt voor een examen

moet bijvoorbeeld duidelijk zijn waarom een bepaald cijfer is aangewezen als de laagste voldoende.

13.1 Het niveau van de schaal

Cijfers winnen aan informatieve waarde naarmate de schaal waarop wordt gerapporteerd een hoger meetniveau heeft. In hoofdstuk 2 zagen we dat naarmate het meetniveau hoger is, de verzameling transformaties naar equivalente schalen kleiner is. Stel dat bijvoorbeeld ruwe scores zouden worden gerapporteerd op een schooltoets die wordt afgenomen voordat de leerstof is behandeld en die na de behandeling nog een keer wordt gemaakt. Kees behaalt de scores 24 en 30 en Hendrik 26 en 32. Het ligt voor de hand om te denken dat beide personen evenveel vooruit zijn gegaan. Echter, het schaalniveau van ruwe scores is lager dan intervalniveau. Daarom kunnen deze twee verschillen op verschillende plaatsen van de ruwe scoreschaal niet zonder meer met elkaar worden vergeleken. We zullen hierna evenwel zien dat met een geschikte theorie de interval-informatie die ruwe scores kunnen bevatten zichtbaar gemaakt kan worden. Het meetniveau van cijfers verkrijgen we door een psychometrische theorie of model over het ontstaan van een antwoordpatroon. Zonder enige theorie hebben we van een groep personen alleen hun antwoordpatronen op de toets of, nog erger, op verschillende toetsen. Twee personen met een verschillend antwoordpatroon op dezelfde toets, bijvoorbeeld 111000 en 001110, worden daarom verschillend beoordeeld. We weten echter niet of het eerste antwoordpatroon een betere, een slechtere of een gelijke prestatie weer-spiegelt als het tweede antwoordpatroon. Zelfs is niet duidelijk of het antwoordpatroon 111100 een grotere prestatie weergeeft dan 000111. Alleen als de antwoordpatronen van twee personen op dezelfde toets gelijk zijn, dan worden hun toetsprestaties gelijk beoordeeld. Indien dat niet het geval is dan moeten de oordelen over hun prestaties verschillen. Zonder enige veronderstelling komen we dus met antwoordpatronen niet verder dan een nominale schaal. Twee antwoordpatronen van verschillende toetsen maken natuurlijk geen enkele onderlinge vergelijking mogelijk. Zonder enige verdere veronderstelling over antwoord-patronen is hun informatieve waarde dus zeer beperkt.

In de klassieke testtheorie wordt dit probleem opgelost door simpelweg te stellen dat de toetsprestatie wordt weergegeven door de som van de itemscores of de ruwe score. De persoon wordt gekarakteriseerd met een ware score op de toets en de ruwe score is daarvan een schatter. Hoe hoger de ruwe score des te groter de toetsprestatie. Alle antwoordpatronen met dezelfde ruwe score zijn daarmee equivalent verklaard en de ruwe score geeft ordinale informatie over de toetsprestatie. De twee eerder genoemde

antwoordpatronen 111000 en 001110 vertegenwoordigen voor de klassieke testtheorie dus een gelijke toetsprestatie, en 011101 een hogere. Door deze afspraak is score 4 hoger dan score 3, en score 3 is hoger dan score 2. Echter, het verschil in niveau tussen de scores 2 en 3 en dat tussen 3 en 4 is niet vergelijkbaar. Immers, de ordinale cijfers 2, 3 en 4 zijn equivalent met bijvoorbeeld 1, 2, 100 en ook met 1, 99, 100. Maar toch, een aanzienlijke winst in de informatieve waarde van het ordinale cijfer ten opzichte van alleen het antwoordpatroon. Het is wel vreemd dat door af te zien van de rijke variëteit aan antwoordpatronen, en grote groepen daarvan als equivalent te beschouwen, het niveau van nominaal naar ordinaal stijgt, en dat we dus aan informatie winnen.

Voorwaarde voor de ordinale informatie van ruwe scores is dat ze op dezelfde toets behaald zijn. Scores op verschillende toetsen zijn niet zonder meer vergelijkbaar. Het ligt voor de hand dat een persoon met een ware score 7 op een toets van 10 items, een hogere ware score heeft op een toets van 20 ongeveer even moeilijke items. Dat zal ongeveer 14 zijn. Voor het probleem van de onderlinge vergelijkbaarheid van scores op verschillende toetsen zijn in het kader van de klassieke testtheorie vele equivaleringsmethoden ontwikkeld (zie hoofdstuk 8).

De introductie van itemresponsmodellen in de psychometrie kan als een belangrijke kwaliteitsimpuls worden beschouwd. We vatten de voordelen van de latente variabele in een itemresponsmodel ten opzichte van de ware score in de klassieke testtheorie nog eens kort samen. Om te beginnen is de waarde van de latente variabele exclusief gekoppeld aan de persoon en niet afhankelijk van de toets zoals de ware score. De toets waarmee de latente vaardigheid wordt geschat, is niet van belang voor de interpretatie van de waarde van de schatter maar alleen voor de nauwkeurigheid daarvan. Voorwaarde is wel dat de items alle-maal afkomstig zijn uit dezelfde verzameling gecalibreerde items of itembank. De geschatte vaardigheden van personen die zijn geschat met hun toetsresultaten op verschillende toetsen uit zo'n verzameling zijn direct vergelijkbaar. Bovendien is het bereikte meetniveau hoger dan het ordinale niveau van de toetsscore. Hoe moeten we begrijpen dat het ordinale niveau van de ruwe score wordt verhoogd naar het intervalniveau van de latente variabele? In de eerste plaats is er het formele argument dat alleen lineaire transformaties van de latente schaal equivalent zijn met de gekozen latente schaal. In de tweede plaats volgt hieruit de meer informele interpretatie dat een bepaalde verhoging van de latente vaardigheid overal op de schaal dezelfde interpretatie toelaat. Gegeven (een verhoging van) de latente vaardigheid kennen we van ieder item (de verandering van) de verdeling van de itemscores, en daarmee bijvoorbeeld ook (van) de verwachte itemscore. Het lijkt erop dat we daarmee niet erg veel opschieten. De itemscores zijn immers van ordinaal

niveau. Lood om oud ijzer dus? We proberen hierna aan te tonen waarom deze vraag ontkennend moet worden beantwoord.

Eerder gaven we het voorbeeld dat de itemscores 1, 2, 3 equivalent zijn met 1, 2, 100, maar ook met 1, 99, 100. Intuïtief voelt iedereen wel aan dat hiermee informatie in de item-scores wordt genegeerd. Bij de introductie van itemscores werd gesteld dat zij in principe ordinaal zijn, evenals toetsscores. Maar toetsconstructeurs kennen bij het opstellen van de scoringsvoorschriften wel degelijk ook informatie toe aan het verschil tussen itemscores. Voor hen zijn 1, 2, 3 en 1, 2, 100 niet hetzelfde. Evenwel, het ontbreekt op het moment van de constructie van de scoringsvoorschriften nog aan een theorie om deze verschillen tussen itemscores meettheoretische betekenis te geven. Daarom kunnen itemscores op dat moment alleen nog maar ordinaal worden geïnterpreteerd. Niet omdat itemscores geen interval-informatie bevatten, maar omdat die er nog niet kan worden uitgehaald. Als er vanaf het begin geen informatie in de verschillen tussen itemscores had gezeten, dan had geen enkele theorie die er uit kunnen halen. Itemresponsmodellen, zoals het Raschmodel of OPLM, kunnen de informatie in de verschillen tussen toetsscores zichtbaar maken.

De parameters in het Raschmodel of OPLM zijn van intervalniveau, of, na een exponentiële transformatie van de modelparameters van log-intervalniveau. Schalen die via een transformatie in elkaar over te voeren zijn, bijvoorbeeld log-interval en interval, worden isomorf genoemd (Stine, 1989). Dit betekent dat zij dezelfde informatieve waarde hebben. Wanneer voor een verzameling items het Raschmodel geldt, kan een transformatie $\hat{\theta}(r)$ worden vastgelegd van toetsscores naar een variabele θ van intervalniveau. Deze transformatie is maar ten dele bepaald door de keuze van het Raschmodel. De schattingsprocedure voor de itemparameters (CML, MML) en de schattingsprocedure voor de persoonsparameters (ML, WML, EAP) zijn mede bepalend voor deze transformatie van toetsscores naar een latente variabele van intervalniveau. We moeten derhalve concluderen dat, wanneer het Raschmodel geldt, ruwe scores isomorf zijn met een schaal van intervalniveau, en derhalve informatie van dit niveau bevatten. Dit betekent echter ook dat de itemscores interval-informatie bevatten. Immers, kies een willekeurig item. Zij r de score van een persoon op de toets zonder het item. Gegeven de score r, wordt de intervalinformatie tussen score 0 en 1 op het item, zichtbaar gemaakt in het verschil tussen $\hat{\theta}(r)$ en $\hat{\theta}(r+1)$.

De eerstvolgende betekenisvolle verhoging van het schaalniveau wordt verkregen door de introductie van een vast nulpunt. Echter, zolang er geen natuurlijk absoluut nulpunt van vaardigheid of itemmoeilijkheid wordt ontdekt, zal het niveau van de schalen in de psychometrie niet boven het intervalniveau uitstijgen.

13.2 Normschalen

Door het cijfer voor een toetsprestatie te laten afhangen van een vergelijking van deze prestatie met de prestaties van een belangrijke groep personen kan de relatieve waarde van de prestatie beter worden beoordeeld. De vergelijkingsgroep wordt een normgroep of referentiepopulatie genoemd, en een cijferschaal waarop de prestaties van een normgroep zijn af te lezen heet een normschaal. De cijfers op een normschaal noemen we normcijfers ter onderscheiding van de cijfers op basis waarvan de normschaal wordt geconstrueerd. Dit kunnen ruwe of gewogen scores zijn, maar ook latente vaardigheidsschattingen. We veronderstellen dat deze cijfers minimaal van ordinaal niveau zijn.

Voor de constructie van een normschaal moet een zogenaamd normeringsonderzoek worden uitgevoerd. Hiertoe moet in de eerste plaats een normgroep ondubbelzinnig worden afgebakend. Een normgroep is bijvoorbeeld alle kinderen in Nederland in groep 8 die niet hebben gedoubleerd. Het is belangrijk dat een normgroep nauwkeurig is omschreven, zodat precies duidelijk is wie er wel en wie er niet toe behoort. Verder moet zij betekenisvol zijn in relatie tot de toetsresultaten. Als de toets bijvoorbeeld is gericht op het meten van de rekenvaardigheden in groep 5 van de basisschool voor de kerstvakantie, dan kan de normgroep precies deze groep bevatten. Echter, als de normschaal beter interpreteerbaar zou worden door alleen de leerlingen te nemen die niet zijn blijven zitten, dan verdient dit de voorkeur.

Vervolgens vereist de constructie van een normschaal dat de frequentieverdeling van de cijfers in de normgroep wordt geschat. Hiertoe moet een representatieve steekproef uit de normgroep worden getrokken. De schatting van de frequentieverdeling is het uitgangspunt voor een ruime keuze aan normschalen. We bespreken vier hoofdtypen van normschalen: cumulatieve verdelingen, genormeerde lineaire transformaties, genormaliseerde schalen en ontwikkelingsschalen.

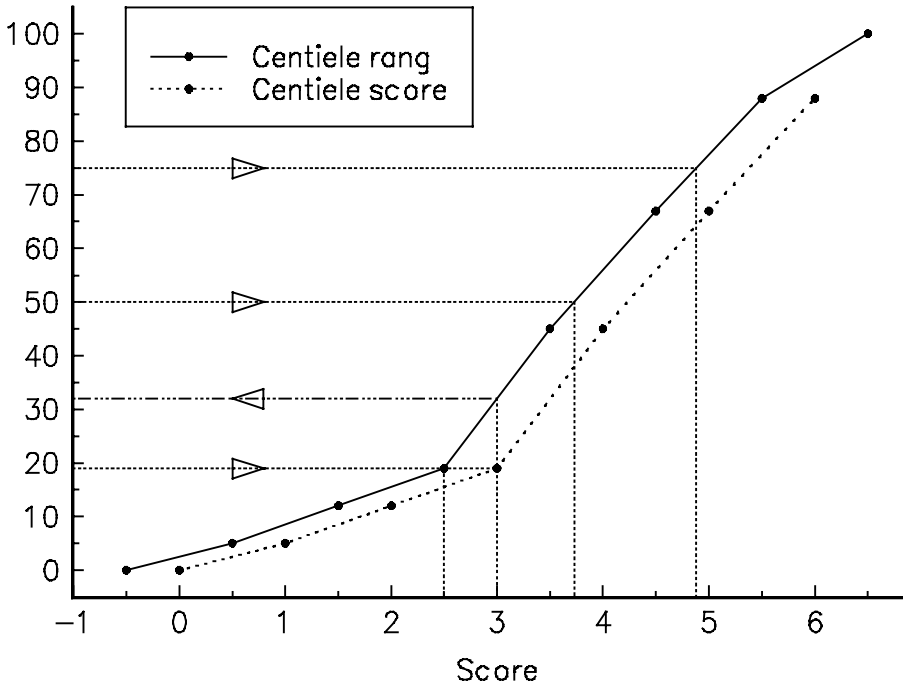
13.2.1 Cumulatieve verdelingen

Afgezien van de onenigheid onder de geleerden over de terminologie is de eenvoudigste normschaal de centiel- of percentielschaal. Uitgangspunt voor een centielschaal is een tabel met (schattingen van) de cumulatieve percentages van de scores op een toets in een normgroep, zoals bijvoorbeeld weergegeven in tabel 13.1.

Tabel 13.1

Cumulatieve percentages van de scores op een toets met zes dichotome items

Scores	Cumulatieve percentages
0	5
1	12
2	19
3	45
4	67
5	88
6	100



Figuur 13.1
Cumulatieve verdelingen en centielschalen bij discrete scores als continue variabele

Op basis van tabel 13.1 zijn er in figuur 13.1 met behulp van lineaire interpolatie twee grafieken voor de verdeling van de scores getekend. De score wordt hier als een continue variabele opgevat en kan derhalve worden gerepresenteerd met een horizontale lijn. De percentages worden op de verticale as afgezet. In figuur 13.1 laten we zien dat voor het tekenen van een verdeling van continue scores meerdere keuzes mogelijk zijn. Het is gebruikelijk in de statistiek om in verband met de zogenaamde correctie voor continuïteit, bijvoorbeeld het percentage 19 bij score 2 op de score-as af

te beelden op 2.5, precies tussen de bijbehorende score en zijn eerstvolgende waarde in. In figuur 13.1 is deze procedure weergegeven met de linker doorgetrokken lijn. Deze lijn wordt gebruikt voor het berekenen van de centiele rang. In figuur 13.1 kan men zien hoe de centiele rang bij score 3 door lineaire interpolatie wordt bepaald. We vinden dat de centiele rang bij score 3 gelijk is aan 19 + (45-19)/2 = 32. De centiele rang wordt ook wel centiele score genoemd (Drenth & Sijtsma, 1990). Hoewel niet de belangrijkste, is een van de oorzaken van de eerder genoemde verwarring het feit dat er in de psychometrie nog een tweede methode wordt gebruikt. Met deze tweede methode beeldt dan het percentage 19 af op de score 3.0. Hieraan wordt wel de naam verbonden van centiel of ook weer centiele score. Een andere benaming is percentiel. Uit tabel 13.1 zien we 19 als cumulatief percentage bij score 2. Dat het centiel 19 bij score 3 hoort, betekent derhalve dat 19% in de normgroep lager scoort dan 3. In de figuur is het enige effect van dit tweede alternatief dat de eerste curve een half scorepunt op de schaal naar rechts is verschoven. Een zekerder interpretatie kan als excuus worden aangevoerd om toch van deze tweede mogelijkheid gebruik te maken. Als het centiel bij score 3 gelijk is aan 19 dan weet men zeker dat men hoger heeft gescoord dan 19% van de normgroep. Bij de centiele rang van 32 bij score 3 is de interpretatie minder duidelijk. Bij een meer gedifferentieerde scoreschaal dan die in het voorbeeld van 0 tot 6 weegt dit voordeel minder zwaar, omdat de afstand tussen de curven voor de centiele score en de centiele rang kleiner is, en gaat het nadeel van een grotere kans op verwarring zwaarder tellen.

Men zegt ook wel dat een score in het zoveelste centiel ligt. Dit woordgebruik verdient enige toelichting. Het eerste centiel loopt van de centielen 0.0 tot 1.0, het tweede van 1.0 tot 2.0, enzovoort. Omdat het centiel van score 2 gelijk is aan 12.0 ligt score 2 dus in het dertiende centiel. Behalve de indeling van de verdeling van de scores in 100 gelijke stukjes, gebruikt men ook andere indelingen. Decielen bijvoorbeeld hebben een vergelijkbare betekenis als centielen, behalve dat de eenheid 10% is in plaats van 1%. In figuur 13.1 delen we de verticale as in tien gelijke delen in. De waarde van het deciel verkrijgen we door de laatste 0 van de getallen langs de verticale as in figuur 13.1 weg te laten. Bij score 2 met centiel gelijk aan 12.0 hoort dan een deciel gelijk 1.2. Omdat het eerste deciel loopt van deciel 0.0 tot 1.0 en het tweede deciel van deciel 1.0 tot 2.0, zegt men ook wel dat score 2, met deciel 1.2, in het tweede deciel ligt. Bij kwartielen is de eenheid 25%. Delen we het centiel van een score door 25 dan verkrijgen we het kwartiel. Het kwartiel bij centiel 12.0 is derhalve 0.48. Ronden we het kwartiel af dan zeggen we dat score 2 in het eerste kwartiel ligt. De algemene benaming voor centielen, decielen enzovoort is quantielen. Het Leerlingvolgsysteem rapporteert bijvoorbeeld in kwartielen per afnamemoment

(normgroep), waarbij het laagste kwartiel nog eens is onderverdeeld in de laagste 10% en de overige 15%. Beelden we bij het verkrijgen van centielen en centiele rangen continue scores af op percentages, voor centiele scores (terminologie van Guilford & Fruchter, 1978), ook wel centiel, centiel punt of centiele rang genoemd, gaan we de andere kant op. Dus van de percentage-schaal naar de continue scoreschaal. We kiezen eerst een percentage *p*, bijvoor-beeld *p* = 75, en zoeken, zoals in figuur 13.1 door lineaire interpolatie, de bijbehorende score. Meestal gebruikt men hiervoor de linker curve voor de centiele rang. Dit is in figuur 13.1 afgebeeld met de lijn die begint bij cumulatief percentage 75. De centiele score bij cumulatief percentage 75 is gelijk aan 4.5 + (75-67)/(88-67) = 4.88. Een andere centiele score is de mediaan. Hiervoor doet men hetzelfde als zojuist bij het percentage 75, maar nu voor het percentage 50. We beginnen dus bij de lijn die begint bij het cumulatieve percentage 50 en vinden dan dat de mediaan gelijk is aan 3.5 + (50-45)/(67-45) = 3.73. Voor de bepaling van de centiele scores wordt ook wel de andere curve genomen.

Uit het voorgaande blijkt dat de naamgeving bij deze schalen in de literatuur onzorgvuldig is. De hoofdbron van de verwarring lijkt te zijn dat er onvoldoende rekening mee wordt gehouden dat een transformatie een relatie tussen twee verzamelingen definieert: een element uit het domein wordt afgebeeld op een element uit de beeldverzameling. Men moet zich derhalve steeds goed realiseren welke twee verzamelingen bij de transformatie zijn betrokken en of bijvoorbeeld de scores op percentages worden afgebeeld of andersom. Hier is hoofzakelijk de terminologie aangehouden zoals gegeven in Guilford en Fruchter (1978). Door de rommelige terminologie bij deze schalen is het geen overbodige luxe om bij een rapportage op een dergelijke schaal zich goed te realiseren wat er is bedoeld. Gelukkig zijn de gehanteerde concepten eenvoudig, zodat de context en de gehanteerde waarden mogelijk de gevraagde helderheid verschaffen. Om misverstanden te voorkomen zou men er goed aan doen termen als centiel, centiele score en centiele rang te vermijden en gewoon te beschrijven hoe de waarden van een schaal zijn berekend.

13.2.2 Genormeerde lineaire transformaties

De algemene gedaante van een lineaire transformatie *s* van een cijfer *r* naar een cijfer *s(r)* is $s(r) = ar + b$. Het cijfer *s* is een normcijfer wanneer de transformatieconstanten *a* en *b* op basis van de frequentieverdeling van *r* zo zijn gekozen dat de prestatie bij een normcijfer gemakkelijk met de prestaties in de normgroep kan worden vergeleken. Omdat met een lineaire transformatie alleen het gemiddelde en de schaaleenheid van

de oorspronkelijke cijferschaal kunnen worden veranderd, worden alleen het gemiddelde en de standaarddeviatie van de frequentieverdeling van de cijfers in de normgroep gebruikt. Een eenvoudig te interpreteren transformatie is die naar standaardscores. De transformatieconstanten *a* en *b* worden zodanig gekozen dat in de normgroep het gemiddelde van de normcijfers *s* gelijk is aan 0 en de standaarddeviatie gelijk is aan 1. Het gemiddelde en de standaarddeviatie van *r* in de normgroep noteren we respectievelijk met μ_r en σ_r. Het is eenvoudig na te gaan dat $a = 1/\sigma_r$ en $b = -\mu_r/\sigma_r$ het gewenste resultaat geven, zodat $s_{(0,1)}(r) = (r-\mu_r)/\sigma_r$. Een standaardscore van $s = 1.0$ betekent derhalve dat men een standaarddeviatie boven het gemiddelde van de normgroep heeft gescoord.

Behalve een gemiddelde van 0 en een standaarddeviatie van 1, zijn vele andere waarden in gebruik, bijvoorbeeld een gemiddelde van 250 en een standaarddeviatie van 10. De waarden voor *a* en *b* die dit bewerkstelligen, verkrijgt men door $s_{(1,0)}$ met 10 te vermenig-vuldigen en er 250 bij op te tellen:

$$s_{(250,\,10)}(r) = 10\frac{(r-\mu_r)}{\sigma_r} + 250 \;\Rightarrow\; a = \frac{10}{\sigma_r},\; b = -\frac{10\mu_r}{\sigma_r} + 250.$$

Toetsscores worden ook vaak lineair getransformeerd naar de nederlandse schoolcijferschaal van 1 tot 10. Ook hier kan de frequentieverdeling van een normgroep aan ten grondslag liggen. Een voorbeeld. Op de cijferschaal wordt de grens tussen voldoende en onvoldoende meestal gelegd bij 5.5. Nemen we aan dat de cijfers worden gerapporteerd met een decimaal. Als men vindt dat 25% van de normgroep hoort te zakken, dan moet de centielscore bij 25%, zeg 17.83, worden afgebeeld op het normcijfer 5.5 - 0.05 = 5.45. Hiermee hebben we het eerste van de twee punten gevonden die de gezochte lineaire transformatie bepalen. Het tweede punt kunnen we vinden door bijvoorbeeld vast te stellen dat niet meer dan 25% van de normgroep een normcijfer 8.0 of hoger mag krijgen. Dan moeten we derhalve zorgen dat centielscore bij 75%, zeg 46.12, wordt afgebeeld op 8.0 - 0.05 = 7.95. De gewenste transformatie krijgen we door het volgende stelsel van twee vergelijkingen op te lossen:
$7.95 = a \times 46.12 + b$ en $5.45 = a \times 17.83 + b$. We vinden dan $a = (7.95 - 5.45)/(46.12 - 17.83)$ en $b = 5.45 - a \times 17.83$. Als de normcijfers niet lager dan 1.0 en niet hoger dan 10.0 mogen zijn, dan rapporteert men 1.0 voor alle cijfers die beneden 1.0 worden afgebeeld en 10.0 voor alle cijfers die boven 10.0 worden afgebeeld.

Een bekend voorbeeld is de ’standaardscore’ die de Eindtoets Basisonderwijs rapporteert voor een leerling (Uiterwijk & Engelen, 1993). Dit zijn geen standaardscores zoals zojuist vermeld, met gemiddelde 0.0 en standaarddeviatie 1.0. De (Eindtoets)standaardscores van een standaardjaar, voor de Eindtoets van 1990 is het standaardjaar 1985, hebben een gemiddelde van 535 en een standaarddeviatie van 10.

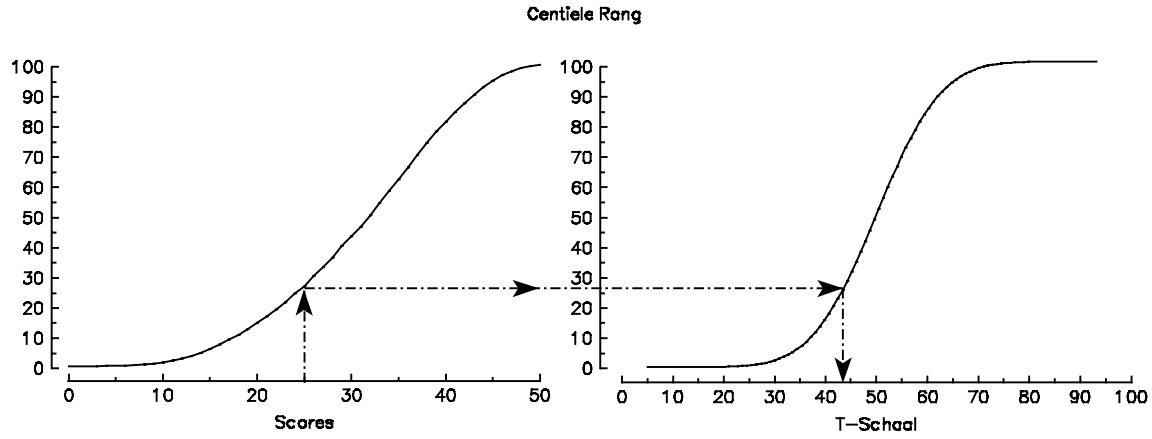
De toetsen na 1985 zijn middels een lineaire equivaleringsprocedure naar de schaal van het standaardjaar getransformeerd.

13.2.3 Genormaliseerde schalen

Tot nu toe werd geen enkele aanname gedaan over de vorm van de verdeling van de normcijfers in de normgroep. Dit lijkt misschien minder relevant, maar het is goed te beseffen dat daardoor de interpretatie van de waarde van een toetsresultaat er flink naast kan zitten. Neem bijvoorbeeld aan dat de cijfers volgens de Beta-verdeling in figuur 13.3 erg scheef naar links verdeeld zijn . De schaal van deze verdeling loopt van 0.0 tot 1.0 en de verdeling heeft een gemiddelde van 0.65 en een standaarddeviatie van 0.23. Stel dat we van een leerling in dit geval een standaardscore van 1.52 zouden rapporteren ($0.65 + 1.52 \times 0.23 \approx 1.00$, dus hoger kan niet). Over het algemeen zal dit worden geïnterpreteerd, weliswaar onterecht maar toch met de normale verdeling in het achterhoofd, als een goede prestatie, behorend tot het hoogste deciel. Deze interpretatie is weliswaar niet onjuist, maar miskend dat de prestatie tot het hoogste centiel van de Beta-verdeling behoort. Deze onjuiste interpretatie wordt vermeden door een genormaliseerde schaal te kiezen. De cijfers op een genormaliseerde schaal zijn verdeeld volgens de normale verdeling. Niet omdat de vaardigheid op de toets zo verdeeld zou zijn in de normgroep, maar eenvoudig omdat de schaal zo is geconstrueerd. Bijvoorbeeld, op een genormaliseerde standaardschaal betekent 1.52 dat precies 94% van de normgroep gelijk of lager scoorde. Het hoogste centiel op een genormaliseerde schaal is pas bereikt bij een cijfer 2.62. Bovendien is het aardige van een aanname over de vorm van de verdeling, dat daarmee een intervalschaal wordt gecreëerd, wanneer men daarbij tenminste ook een dichtheidsfunctie veronderstelt. Een ééndimensionale verdeling en een daarbij behorende dichtheidsfunctie veronderstellen noodzakelijkerwijs een lengtemaat op de intervallen van zijn domein. Was dat niet het geval, dan zou de dichtheids-functie niet zijn gedefinieerd. De dichtheidsfunctie is immers de afgeleide van de verdeling naar de maat op het domein. Wanneer het niveau van de oorspronkelijke cijfers niet van intervalniveau is, dan is men vrij om een dergelijke aanname te maken omdat zij op geen enkele manier getoetst en verworpen kan worden. Wanneer de oorspronkelijke schaal wel van intervalniveau is, dan is een hypothese over de verdeling wel toetsbaar. We komen hier nog op terug.

In de sociale wetenschappen gebruikt men graag de normale verdeling. Het hoeft ons dan ook niet te verbazen dat vaak wordt verondersteld dat de normcijfers normaal zijn verdeeld met een vrij te kiezen gemiddelde en standaarddeviatie (μ,σ). Veel

voorkomende genormaliseerde schalen zijn de T-schaal, de C-schaal en de Stanine schaal. Voor de T-schaal kiest men $(\mu, \sigma) = (50,10)$, voor de C-schaal en de Stanines $(\mu, \sigma) = (5,2)$. Voor deze laatste twee schalen komt daar nog bij dat alleen gehele getallen worden gerapporteerd. Voor de C-schaal lopen die getallen van 0 tot 10. Stanines zijn identiek aan de C-schaal, behalve dat de C-schaalcijfers 0 en 1 worden samengevoegd tot Stanine 1 en de C-schaalcijfers 9 en 10 tot Stanine 9.



Figuur 13.2

Bepaling van T-schaal bij een toetsscore. Links staan centiele rangen van een referentiepopulatie bij de toetsscores. Rechts is de cumulatieve normale verdeling N(50, 10^2) weergegeven

Het algemene principe voor de berekening van genormaliseerde schalen is als volgt (zie figuur 13.2). Zij G een cumulatieve verdelingsfunctie, bijvoorbeeld de cumulatieve normale verdeling $N(50, 10^2)$. Dan is het genormaliseerde cijfer $n(r)$ van cijfer r met centiele rang $c(r)$ gelijk aan $n(r) = G^{-1}(c(r))$, dus $G(n(r)) = c(r)$. Oftewel de cumulatieve verdelingsfunctie met als argument het genormaliseerde cijfer is gelijk aan de centiele rang van het cijfer. De linker grafiek representeert de centiele rangen bij de cijfers, de functie $c(r)$. De rechter grafiek toont de cumulatieve normale verdelingsfunctie met gemiddelde 50 en standaarddeviatie 10. In figuur 13.2 is de bepaling van de T-score bij cijfer 25 grafisch weergegeven. Daartoe zoeken we eerst de centiele rang p_{25} bij cijfer 25. Dit is weergegeven in het linkerdeel van figuur 13.2. Daar kunnen we zien dat p_{25} ongeveer gelijk is aan 26. Vervolgens zoeken we bij p_{25} de T-schaalwaarde, zoals weergegeven in het rechterdeel van figuur 13.2. Daar zien we dat de T-schaalwaarde bij score 25 ongeveer gelijk is aan 43.

T-schaalcijfers worden niet altijd gebaseerd op centiele rangen. Men gebruikt ook wel het cumulatieve percentage van de toetsscore lager dan de betreffende toetsscore (centiel), en soms ook wel inclusief de betreffende toetsscore zelf.

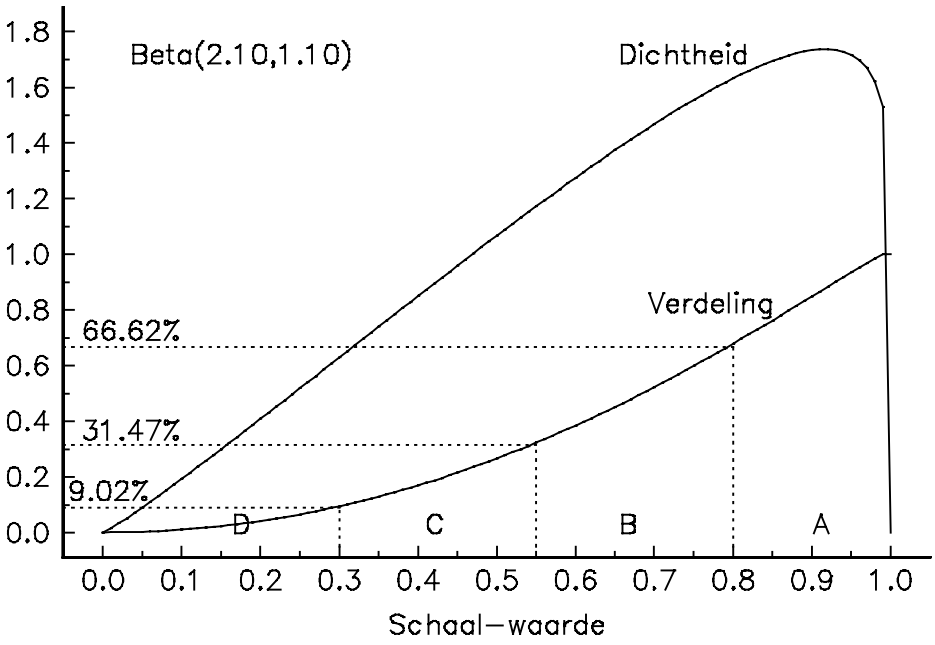
Tabel 13.2

Bovengrenzen van genormaliseerde standaardscores en centiele rangen voor de C-schaal

C-schaal waarde	Genormaliseerde standaardscore	Centiele rang
0	-2.25	1.2
1	-1.75	4.0
2	-1.25	10.6
3	-0.75	22.7
4	-0.25	40.1
5	0.25	59.9
6	0.75	77.3
7	1.25	89.4
8	1.75	96.0
9	2.25	98.8
10	∞	100.0

In tabel 13.2 zijn de bovengrenzen van de centiele rangen opgenomen voor de C-schaal. Het C-schaalcijfer van een cijfer wordt gevonden bij de kleinste bovengrens groter dan de centiele rang van het cijfer. Als bijvoorbeeld cijfer 25 een centiele rang heeft van 26.5, dan is het C-schaalcijfer voor cijfer 25 gelijk aan 4, omdat 40.1 de kleinste bovengrens is groter dan 26.5. De onderlinge afstand tussen C-schaalcijfers komt overeen met 0.50 standaarddeviatie. Natuurlijk kunnen we de C-schaalcijfers door een lineaire transformatie afbeelden naar een schaal met gemiddelde 0 en standaarddeviatie 1.0. Daartoe trekken we van het C-schaalcijfer 5 af en delen het resultaat door 2. We verkrijgen dan de genormaliseerde versie van de eerder genoemde standaardscores. Genormaliseerde standaardscores zijn per definitie normaal verdeeld. Daarentegen heeft de verdeling van de eerder genoemde standaardscores dezelfde vorm als die van de oorspronkelijke cijfers. Let wel dat tabel 13.2 de bovengrenzen van de genormaliseerde standaardscores bij de C-schaal bevat. Bij de C-schaalwaarde 5 hoort bijvoorbeeld een genormaliseerde standaardscore van 0.0, de bovengrens is echter 0.25. Een niet onbelangrijk voorbeeld van een genormaliseerde schaal is de deviatie-IQ-schaal. Dit IQ is in iedere normgroep (leeftijdsgroep) normaal verdeeld met een gemiddelde van 100 en een standaarddeviatie van 15. De gemiddelde intelligentie, voor zover gemeten door de Stanford-Binet IQ-tests, neemt na het vijftiende levensjaar niet meer toe (Linn, 1989). Een willekeurige steekproef uit de populatie van volwassenen en een willekeurige steekproef van vijftienjarigen hebben dezelfde gemiddelde ruwe score op de Stanford-Binet test. Linn vermeldt niet of de variantie boven deze leeftijd

onveranderd blijft. Voor de SON (Snijders-Oomen Non-verbale intelligentietest, 1991) zijn normschalen gepubliceerd voor de nederlandse populatie voor leeftijden van 5.5 tot 16.5 jaar. Deze schalen laten nog een progressie zien tot en met de hoogste leeftijdsgroep.



Figuur 13.3
De Beta-getransformeerde schaal van de Entreetoets

Een vergelijkbare procedure is gevolgd bij de Entreetoets van het Cito (Moelands, 1988). Dit is overigens net als de Eindtoets, een hele batterij van toetsen die samen een groot deel van de leerstof van het laatste jaar van de basisschool dekken. Voor de schalen van de toetsen in de Entreetoets werd echter geen normale verdeling gekozen maar de Betaverdeling B(2.10, 1.10). Voor de verdeling in figuur 13.3 kan men de cijfers 0.0 t/m 1.0 langs de verticale as lezen als centiele rangen gedeeld door honderd. Figuur 13.3 is dan een Beta-equivalent van het rechterdeel van figuur 13.2. We hebben hier dus geen genormaliseerde schaal maar een 'Beta-getransformeerde' schaal. Deze verdeling werd gekozen omdat zij redelijk aansloot bij de wens de totale schaal in vier hoofdcategorieën (A, B, C, D) in te delen die respectievelijk de 30% hoogste scoorders bevat (A), de middelste 40% (B), 20% lagere (C) en de 10% laagste (D). Verder wenste men de Beta-schaal in te delen in 20 intervallen ter grootte van 0.05, zodanig dat de verdeling van deze intervallen over de hoofd categorieën D t/m A gelijk is aan 6, 5, 5, 4. Hoofdcategorie D bevat de eerste 6 van deze eenheden, B en C ieder 5, en

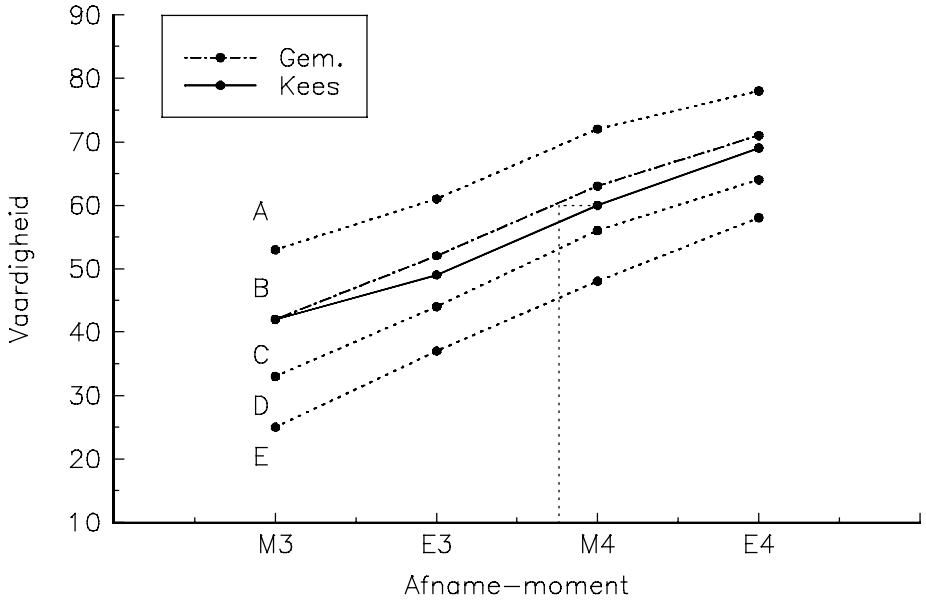
A de hoogste 4 (17 t/m 20). De genoemde B(2.10, 1.10) voldeed ongeveer aan deze merkwaardig gedetailleerde wensen. Zoals in figuur 13.3 te zien is, leidt deze transformatie tot een aan de onderkant enigszins uitgerekte, maar overigens bijna lineaire transformatie van de percentielschaal. Door deze aan de onderkant gerekte schaal wordt bereikt dat de cijfers op de twintigpuntsschaal vooral differentiëren tussen zwakkere leerlingen: de eerste elf van de twintig punten wordt verdeeld onder de 31.5% laagst scorende leerlingen. Dit laatste is in overeenstemming met het doel van de Entreetoets om vooral te letten op het lagere deel van de schaal: het detecteren van zorgwekkend lage niveaus in het vaardigheidsprofiel van een leerling.

13.2.4 Ontwikkelingsschalen

De intelligentietests van Binet-Simon (Drenth & Sijtsma, 1990) rapporteerden intelligentie als het quotiënt van mentale leeftijd en kalenderleeftijd maal 100: het intelligentiequotiënt. De mentale leeftijd van een kind met cijfer *r* is de leeftijdsgroep met gemiddeld cijfer *r*. De mentale leeftijd is een voorbeeld van een ontwikkelingsschaal. De constructie van een ontwikkelingsschaal vereist grootschalig onderzoek. Men kiest een normgroep met een voldoende range aan leeftijden, bijvoorbeeld de populatie van het basisonderwijs. Men groepeert de leeftijden in deze normgroep in een aantal categorieën. Bijvoorbeeld de leeftijdscategorie 6 bevat alle leerlingen die op het moment van de toetsafname tussen de 5½ en 6½ jaar oud zijn. De leeftijdsgroep 6 zouden we dan een deelnormgroep kunnen noemen. Uit alle leeftijdsgroepen trekt men een representatieve steekproef. Voor iedere leeftijdsgroep wordt het gemiddelde cijfer bepaald, eventueel de mediaan. Vervolgens wordt bijvoorbeeld door lineaire interpolatie een regressiefunctie van de cijferschaal naar de leeftijdsschaal verkregen. Deze regressiefunctie geeft bij ieder cijfer een leeftijdsaanduiding, bijvoorbeeld bij cijfer 25 de leeftijd 5;7 jaar. Zou men de Binet-Simon manier van rapporteren kiezen en stel dat het kind met score 25 de leeftijd heeft van 5;5 jaar, dan is de quotiëntscore (5 7/12) : (5 5/12) x 100 = 103.

Het rapporteren van toetsresultaten op een ontwikkelingsschaal is tamelijk problematisch en de rapportage op een quotiëntschaal dus ook. Verschillende vaardigheden kunnen zich met verschillende snelheid ontwikkelen ten opzichte van de spreiding in een normgroep. Het gemiddelde verschil in leesvaardigheid tussen zeven en negen jaar kan bijvoorbeeld maar een standaarddeviatie op de schaal van zevenjarigen groot zijn, terwijl dit voor rekenen gelijk zou kunnen zijn aan twee standaarddeviaties. Rekenen is voor achtjarigen bijvoorbeeld al een standaarddeviatie

hoger. Dergelijke verschillen in ontwikkelingssnelheid leiden tot oneven-wichtigheid in de rapportage. Neem een kind van zeven jaar dat zowel op een leestoets als op een rekentoets een standaarddeviatie boven het gemiddelde van zijn leeftijdsgroep scoort. Dit kind krijgt voor lezen het leeftijdscijfer 9 en voor rekenen een 8. Dit wekt de indruk dat het kind met lezen meer presteert dan met rekenen. Het is gemakkelijk dit voorbeeld zo extreem te maken dat men wel moet concluderen dat deze indruk onterecht is.



Figuur 13.4
Het grafische LVS rapport van de ontwikkeling van Kees

De bovengenoemde problemen kunnen worden opgelost door een rapportagevorm te vinden waarin zowel de ontwikkeling van de normgroep, als de plaats van de persoon in zijn huidige normgroep tot zijn recht komt. Nog beter is het wanneer ook de ontwikkeling van de persoon kan worden weergegeven. Deze vorm heeft men in het Leerlingvolgsysteem (Jansen e.a., 1992) weten te realiseren, hoewel een adequate schatting van de ontwikkeling van de persoon technische problemen oplevert (zie hoofdstuk 10). Figuur 13.4 laat het grafische rapport zien van de prestaties van Kees op de rekentoetsen voor de afnamemomenten Medio Groep 3 (M3) tot en met Eind Groep 4 (E4). De gebieden A, B en C bevatten de drie bovenste kwartielen van de centielschaal, waarvan A (boven de bovenste lijn) het hoogste deel. D en E bevatten samen het onderste kwartiel, waarvan E de laagste 10%. Voor Kees zijn in de grafiek niet alleen zijn positie binnen zijn groep duidelijk, maar ook zijn 'Groepsequivalenten'. Bijvoorbeeld, het snijpunt van de horizontale lijn door zijn positie op M4 met de lijn

voor het gemiddelde levert zijn Groepsequivalent op M4. Dit ligt ongeveer op een kwart van de afstand (E3, M4) onder M4 (figuur 13.4). Nemen we aan dat de tijd tussen E3 en M4 een half leerjaar bedraagt, dan zou men kunnen zeggen dat hij op M4 een vaardigheid heeft die gelijk is aan het gemiddelde in de normgroep van ongeveer een achtste leerjaar geleden, of dat hij op M4 ten opzichte van het gemiddelde in zijn groep een achtste leerjaar achterloopt. De bepaling van dit snijpunt lukt natuurlijk niet voor alle gevallen. Voor een leerling die op M3 beneden het gemiddelde scoort, bestaat zo'n snijpunt niet. Dit is echter een probleem dat aan alle ontwikkelingsschalen kleeft en is niet uniek voor de schalen van het Leerling-volgsysteem.

13.2.5 De nauwkeurigheid van normschalen

Normschalen zijn gebaseerd op een schatting van de frequentieverdeling in een normgroep. De schatting van deze frequentieverdeling is natuurlijk behept met steekproeffouten. Met name wanneer er nonrespons te verwachten is die samenhangt met de te normeren schaal kan de schattingsfout van de frequentieverdeling aanzienlijk zijn. Wanneer bijvoorbeeld in een normeringsonderzoek van een rekentoets vooral de slecht presterende scholen niet meedoen, dan zal de resulterende normschaal een te somber beeld geven van de prestaties van de leerlingen. De schatting van het gemiddelde cijfer van de toets in de normgroep zal dan bijvoorbeeld hoger uitvallen dan in werkelijkheid het geval is. Een leerling die in werkelijkheid gemiddeld scoort, zal een normcijfer krijgen dat aangeeft dat hij beneden het gemiddelde presteert. De steekproeffouten kunnen worden verkleind door een gestratificeerde steekproef te trekken waarin bijvoorbeeld de percentages jongens en meisjes gelijk zijn aan die in de normgroep. Een belangrijke overweging voor de keuze van stratificatievariabelen is de beschikbaarheid van de verdeling uit een andere bron, bijvoorbeeld het Centraal Bureau voor de Statistiek (CBS). De tweede overweging voor de keuze van een stratificatievariabele is een verwachte samenhang met een dreigende nonrespons. Wanneer de stratificatievariabelen aan beide voorwaarden voldoen, dan kan de representativiteit van de steekproef en de mogelijke invloed van nonrespons worden ingeschat en eventueel worden gecorrigeerd. Angoff (1971) bespreekt overwegingen rond steekproef-trekking en vereiste nauwkeurigheid van normschalen. Zijn aanbevelingen komen erop neer dat de steekproeffouten van de normschaal ten opzichte van de meetfouten van de normcijfers verwaarloosbaar horen te zijn. In de rapportage over een normschaal mag een verslag over de representativiteit van de steekproef niet ontbreken. Hierin wordt de verdeling van belangrijke

achtergrondvariabelen in de steekproef vergeleken met de verdeling in de normgroep, voor zover bekend uit bijvoorbeeld CBS-publikaties.

13.3 Beheersingsschalen

Hoewel voor veel schoolvakken een normcijfer een belangrijke indicatie is voor het niveau van de prestatie, zijn er ook situaties waar het er minder toe doet welk percentiel van een normgroep aan de prestatie van een persoon gehecht moet worden. Piloten moeten een vliegtuig veilig aan de grond zetten. Het doet er niet toe of 90% van de kandidaten daartoe in staat is of maar 1%. Zoiets geldt ook voor loodgieters en bruggenbouwers. Hun produkten moeten gewoon voldoen aan de eisen die daaraan gesteld moeten worden. In dit soort gevallen geeft een normschaal niet de gewenste informatie. Een normcijfer geeft geen inzicht in het niveau van de prestatie. Hoe goed kan een persoon rekenen die een centiel van 80 scoort in groep 4? Hoeveel procent van de aftreksommen met getallen van vier cijfers maakt zo'n leerling goed? Hoeveel procent van de deelsommen? Dit type informatie wordt gegeven door een beheersingsschaal. Het kan zowel gaan om een indicatie van de huidige beheersing, alsook voor een te verwachten beheersing in de toekomst. Beheersingsschalen geven een cijfer betekenis door dit te transformeren naar een maat die aangeeft in welke mate de persoon een leerstofonderdeel beheerst of zal beheersen. We noemen deze maat verder het beheersings- cijfer. De psychometrie van beheersingsschalen werd met name in de jaren 70 ontwikkeld. Men noemt beheersingsschalen ook wel criterium-georiënteerde schalen (Van der Linden, 1982).

Het eerste probleem bij de constructie van een beheersingsschaal is het afbakenen van het leerstofdomein. Zolang hierover onduidelijkheid bestaat kan aan geen enkel beheersingscijfer een ondubbelzinnige betekenis worden gegeven. Het probleem voor de afbakening is de veelal grote keuze aan invalshoeken en begrenzingen. Deze kunnen leerstofgericht zijn of gebaseerd zijn op cognitief psychologische onderscheidingen. Ook het onderscheid tussen kennis, toepassing en inzicht wordt hier vaak gehanteerd. Daar komt nog bij dat vele van deze onderscheidingen erg vaag zijn. Het lijdt bijvoorbeeld geen twijfel, dat een toepassing toch vaak ook inzicht vereist. En kan een leerling inzicht hebben zonder dat deze evidente toe-passingen ziet? Ook een inhoudelijke afbakening laat echter vaak meerdere interpretaties toe. Zo hebben bijvoorbeeld de schoolvakken aardrijkskunde en wiskunde de laatste decennia grote veranderingen ondergaan. Maar niet duidelijk is of leerstofonderdelen die nu expliciet tot de leerstof worden gerekend, er tevoren, impliciet of in de praktijk, ook al niet toe behoorden.

Het probleem van de afbakening van een leerstofdomein is concreter wanneer men niet alleen over tamelijk abstracte leerdoelen praat, maar ook over een concrete verzameling items. Eerst maakt men afspraken waarover de items zullen gaan, maar daarna kan worden volstaan met de vraag of een bepaald item nu wel of niet tot het domein kan worden gerekend. Bovendien kan men lacunes in de itemverzameling opsporen, daar weer items bij schrijven, enzovoort. Zo kan een itembank ontstaan waar over men het gemakkelijker over eens kan worden dat hiermee een leerstofdomein kan worden gemeten. Een groot voordeel van de constructie van een dergelijke itembank is de duidelijke betekenis die daarmee aan een beheersingscijfer kan worden gehecht. Men kan bijvoorbeeld rapporteren welk percentage van deze verzameling naar verwachting correct beantwoord zal worden. Binnen de klassieke testtheorie is dit zonder groot verlies van nauwkeurigheid (generaliseren) echter niet goed mogelijk. Daar beperkt men zich vaak tot het percentage van de items in de toets zelf. Als schatter van dit percentage neemt men dan eenvoudig $r/m \times 100$ %, waarin r de toetsscore en m de maximaal te behalen score op de toets. Deze oplossing heeft het bezwaar dat twee verschillende toetsen uit dezelfde itemverzameling kunnen verschillen in moeilijkheid. Een percentage beheersing op een gemakkelijke toets is dan een overschatting van het percentage beheersing van de itemverzameling en een percentage op een moeilijke toets een onderschatting. Binnen het kader van IRT vervalt dit bezwaar doordat voor iedere schatting van de latente vaardigheid het verwachte percentage correct op de complete gecalibreerde itemverzameling kan worden berekend.

Ook de Eindtoets rapporteert beheersingscijfers. Wegens het ontbreken van een gecalibreerde itembank hebben deze beheersingscijfers echter alleen betrekking op de gemaakte toetsen. Men rapporteert het percentage items uit de toets dat goed is beantwoord. Bij het Leerlingvolgsysteem wordt een fraai grafisch overzicht gepresenteerd van de beheersingsgraad van een leerling op de vaardigheidsschaal, waarop ook het interval tussen 50% en 80% kans op correct voor een selectie van de items is aangegeven.

13.4 Het rapporteren van meetnauwkeurigheid

Voor een goede interpretatie van cijfers is het belangrijk als de nauwkeurigheid gemakkelijk is af te lezen. Een algemeen raamwerk hiervoor wordt beschreven in Kolen (1986, 1988). Men kiest een cijferstap h en een $\gamma \times 100$ % betrouwbaarheidsinterval. Vervolgens wordt een transformatie $s(r)$ van de cijfers r geconstrueerd zodat bij iedere

s het interval $[s - h, s + h]$ een tweezijdig $\gamma\times100$ % betrouwbaarheidsinterval is. Kiest men bijvoorbeeld $h = 1.0$ en $\gamma = 0.50$, dan is voor een getransformeerd cijfer *s(r)* het interval [*s* - 1.0, *s* + 1.0] een 50% betrouwbaarheidsinterval rond *s*.

Als de standaardmeetfout σ_E van de cijfers *r* constant is over het bereik van *r*, dan is de transformatie *s* lineair. De coëfficiënt *b* van de lineaire transformatie *s(r) = ar + b* kan arbitrair worden gekozen terwijl *a* als volgt wordt bepaald. Laat z_γ het getal zijn waarvoor

$$(2\pi)^{-\frac{1}{2}} \int_{-z_\gamma}^{z_\gamma} \exp\left(\frac{-t^2}{2}\right) dt = \gamma, \tag{13.1}$$

dan is $a = h/(\sigma_E z_\gamma)$. Let wel dat het gebruik van (13.1) een normaal verdeelde meetfout veronderstelt.

Als de standaardmeetfout niet constant wordt verondersteld, maar een functie σ_E is van het cijfer *r*, dan wordt het ingewikkelder. Kolen (1986, 1988) behandelt de arcsinus-transformatie (Freeman & Tukey, 1950; Lord & Novick, 1968). De variantie van de arcsinustransformatie van de ruwe score is onder het binomiale of compound binomiale foutenmodel ongeveer constant. Het is op zich een interessant probleem om bij een willekeurige standaardmeetfout als functie van *r* een variantiestabiliserende transformatie te bedenken. Zij daarom de meetfouten van cijfer *r* verdeeld volgens G_r met standaarddeviatie $\sigma_E(r)$. Het meest voor de hand ligt om de functie $\sigma_E(r)$ te zien als een te corrigeren transformatie T^{-1} van de maat van de intervallen tussen de opeenvolgende cijfers *r*. Door de inverse transformatie te nemen kan de variabele standaarddeviatie constant worden gemaakt:

$$T(r) = \int_{r_0}^{r} \frac{1}{\sigma_E(v)}\, dv$$

waarin r_0 een willekeurig cijfer is. Hierna volgt een schets van het bewijs dat de meetfout van $T(r)$ ongeveer constant is. Het kwadraat van de standaardmeetfout van $T(r)$ is

$$\begin{aligned}
\sigma_E(T(r))^2 &= \int_R (T(u) - T(r))^2\, d\,G_r(u) \\
&= \int_R \left(\int_{r_0}^{u} \frac{1}{\sigma_E(v)}\, dv - \int_{r_0}^{r} \frac{1}{\sigma_E(v)}\, dv \right)^2 d\,G_r(u) \\
&= \int_R \left(\int_{u}^{r} \frac{1}{\sigma_E(v)}\, dv \right)^2 d\,G_r(u),
\end{aligned}$$

waarin R het domein van r. Veronderstellen we nu dat de standaardmeetfout $\sigma_E(v)$ voor v 'in de buurt van' r ongeveer gelijk is aan $\sigma_E(r)$, dan blijkt dat

$$\sigma_E(T(r))^2 \approx \int_R \left(\frac{u-r}{\sigma_E(r)} \right)^2 dG_r(u)$$

$$= \frac{\sigma_E(r)^2}{\sigma_E(r)^2} = 1,$$

ongeveer constant is. De uitdrukking 'in de buurt van r' moet men zien in relatie tot G_r. Het 'ongeveer gelijk aan' is in samenhang met 'in de buurt van r' preciezer te maken, maar dat is hier niet zo relevant.

Deze transformatie maakt het mogelijk om voor iedere cijferschaal waarvan de standaardmeetfout bekend is een schaal te construeren volgens het recept van Kolen. Zij bijvoorbeeld het cijfer een schatting $\hat{\theta}$ van de latente vaardigheid θ op een Raschschaal, geschat met een toets met informatiefunctie $I(\theta)$. Dan is de transformatie $T(\hat{\theta})$:

$$T(\hat{\theta}) = \int_{-\infty}^{\hat{\theta}} \sqrt{I(v)}\, dv. \tag{13.2}$$

Is het cijfer de ruwe score op deze toets, dan krijgen we de transformatie $T(r)$:

$$T(r(\hat{\theta})) = \int_{-\infty}^{\hat{\theta}} \frac{1}{\sqrt{I(v)}} dr(v), \tag{13.3}$$

waarin $r(\hat{\theta})$ de verwachte score op de toets voor latente vaardigheidsschatter $\hat{\theta}$. Deze transformatie kan voor toetsen, die aan het Raschmodel voldoen, in plaats van de bovengenoemde Freeman-Tukey arcsinustransformatie worden gekozen. Uiteraard leiden (13.2) en (13.3), als functie van $\hat{\theta}$, tot hetzelfde resultaat. Dit is ook als volgt in te zien. De informatiefunctie is gedefinieerd als:

$$I(\theta) \doteq \frac{\left(\frac{\partial r(\theta)}{\partial \theta} \right)^2}{\sigma_r^2(\theta)}.$$

Omdat in het Raschmodel $I(\theta) = \sigma_r^2(\theta)$, volgt dat $dr(\theta) = I(\theta)\, d\theta$, waarmee de identiteit van (13.2) en (13.3) is aangetoond.

Een reden die kan worden aangevoerd om te kiezen tussen bijvoorbeeld T-schaal, C-schaal of Stanines is de meetnauwkeurigheid. De algemene regel is om met de

rapportage van het cijfer geen grotere nauwkeurigheid te suggereren dan de standaardmeetfout van het cijfer toelaat. Deze enigszins vage regel wordt dan geconcretiseerd tot de vuistregel dat de cijfers moeten oplopen in stappen van ongeveer een standaardmeetfout. Kolen (1986, 1988) wijst erop dat deze procedure niet goed te verdedigen is. Immers, bij toepassing van de vuistregel voegt men dan door afronden maximaal een halve standaardmeetfout toe aan de meetfout, gemiddeld dus ongeveer een kwart van de standaardmeetfout. Natuurlijk moet er geen betekenisloze precisie worden gerapporteerd, maar een kwart van de standaardmeetfout lijkt te veel. Een betere richtlijn zou zijn om voor de rapportage een precisie te kiezen waarbij de door afronden toegevoegde meetfout verwaarloosbaar is ten opzichte van de meetfout. Men kan natuurlijk een kwart toegevoegde meetfout verwaarloosbaar vinden. Dit is evenwel niet goed te rijmen met de moeite en kosten die gepaard gaan met de constructie van zo nauwkeurig mogelijke meetinstrumenten. Dit betekent ook dat meetnauwkeurigheid minder belangrijk is voor de keuze tussen de zojuist genoemde drie schalen. Hoewel dit niet gebruikelijk is, kan men bijvoorbeeld C-schaalwaarden op een decimaal nauwkeurig rapporteren.

Duidelijker is het om het betrouwbaarheidsinterval van bijvoorbeeld een standaardmeetfout op de schaal zelf af te beelden (zie tabel 13.3). Dit verdient de voorkeur boven het kiezen van de schaaleenheid op basis van de meetnauwkeurigheid. Deze procedure wordt onder andere gevolgd bij de Eindtoets door het betrouwbaarheidsinterval van het cijfer van een leerling met enkele aaneengesloten sterretjes op de cijferschaal weer te geven.

Tabel 13.3

Rapportage van toetsresultaat en de nauwkeurigheid op een reeks van schalen

*	: puntschatting					
+++*+++	: 50% betrouwbaarheidsinterval					
---+++*+++---	: 90% betrouwbaarheidsinterval					
Aantal items goed:	10	12	14	16	18	20
Standaardscore	25	29	33	37	41	45
Percentiel	46	51	56	61	65	70
Groeps-equivalent	5:4	5:8	5:12	6:4	6:8	6:12
Beheersing %	50	59	67	77	86	93
Cijfer	5.2	5.5	6.5	7.5	8.5	9.5
Resultaat Kees	---------+++++++*+++++++---------					

Dit kan, met enige voorzichtigheid, in een keer voor meerdere typen schalen tegelijk. Stel dat de toetsresultaten worden gerapporteerd op de ruwe score-schaal *r*, een genormeerde lineaire transformatie-schaal, standaardscore genoemd, $s(r) = a \times r + b$, een centielschaal, een ontwikkelingsschaal waarop de basisschoolgroep en het aantal maanden van het schooljaar wordt weergegeven, een beheersingsschaal, en op een cijferschaal van 1 tot 10 die wordt verkregen met twee lineaire transformaties met 'de knik' bij het cijfer 5.5. Hoe het rapport er dan kan uitzien is in tabel 13.3 weergegeven. Hoe moeten we nu naar een dergelijk uitgebreid rapport kijken? De puntschatting geeft het behaalde resultaat weer, in eerste instantie de ruwe score, want dat is de schaal waarvan de overige schalen zijn afgeleid. Kees had 16 items goed en de puntschatting weergegeven met * moet dus precies onder het getal 16 in de ruwe scoreschaal staan. Neem aan dat de beide betrouwbaarheidsintervallen bepaald zijn met de standaardmeetfout van de ruwe score. Uit de tabel is te lezen dat het 50%-betrouwbaarheidsinterval van de score van Kees loopt van ongeveer 14 tot 18, het 90% betrouwbaarheidsinterval van ongeveer 12 tot 20. In een overzicht als het bovenstaande geldt dat het betrouwbaarheidsinterval voor alle lineaire transformaties eenvoudig kan worden afgelezen. Stel dat de ondergrens van het 50% interval iets boven de 14 ligt, bijvoorbeeld 14.5, dan ligt deze ondergrens voor de standaardscore ook precies op een kwart van het interval [33,37] vanaf 33, dus op 34. In principe moet men voorzichtiger zijn met niet-lineaire transformaties, omdat men eigenlijk volgens de transformatie zelf moet interpoleren. De bovenstaande schalen wijken over het algemeen, tussen de gespecificeerde cijfers in, zo weinig af van lineariteit dat lineaire interpolatie binnen de gespecificeerde intervallen geen foute interpretaties tot gevolg zal hebben. Bijvoorbeeld, bij een ondergrens van de ruwe score op 14.5, ligt de ondergrens op de beheersingsschaal ongeveer op 67+(77-67)/4 = 69.5. Wanneer een intervalgrens zich precies op een gespecificeerd cijfer bevindt maakt men, ook bij niet-lineaire schalen, geen interpretatie-fout. Als bijvoorbeeld de ondergrens van het 50%-betrouwbaarheidsinterval van de ruwe score precies gelijk is aan 14, dan is deze ondergrens voor de schaal met groepsequivalenten precies gelijk aan 5:12. Dit is ook het geval wanneer groeps-equivalenten niet lineair zijn met de ruwe scores.

Op deze plaats is ook een waarschuwing op zijn plaats in verband met de interpretatie van een score op een ontwikkelingsschaal en de nauwkeurigheid van het meetinstrument. Als de normgroep slechts langzaam groeit op het meetinstrument, kan men grote betrouwbaarheids-intervallen verwachten op de ontwikkelingsschaal, ook bij een relatief nauwkeurig meet-instrument. Kijken we in dit verband weer eens naar het rapport van Kees in figuur 13.4. Nemen we weer zijn resultaat op M4. Daarvan werd beschreven dat zijn resultaat impliceerde dat hij ongeveer een achtste leerjaar op zijn

normgroep achterloopt. Nemen we aan, wat niet onwaarschijnlijk is, dat het 50%-betrouwbaarheidsinterval van zijn meting op E4 ongeveer loopt van de helft van het interval B tot de helft van het interval D, dan is het 50%-betrouwbaarheidsinterval op de groepsequivalenten schaal ongeveer gelijk aan [E3,E4], oftewel een heel leerjaar. Erg veel zekerheid over de vermeende achtste jaar achterstand hebben we dus niet.

13.5 De cesuur voldoende/onvoldoende en andere normen voor cijfergeving

Onder cesuur verstaan we hier het laagste voldoende cijfer. Omdat de cesuur de grens markeert tussen voldoende en onvoldoende, is zij daarmee het belangrijkste cijfer van een schooltoets. Geen wonder dat daarover reeds veel is nagedacht en geschreven (Berk, 1986). De methodes voor cesuur bepaling die ons uit de literatuur bekend zijn, stammen grotendeels uit de zeventiger jaren waarin de beschikking van interactieve computerprogrammatuur niet vanzelfsprekend was, noch het beheer van gecalibreerde itembanken. Deze twee nieuwe mogelijkheden mogen bij de zo belangrijke cesuurbepaling niet worden genegeerd. Hetzelfde geldt evenwel voor de traditie. Daarom is het van belang een vruchtbare synthese tot stand te brengen tussen de concepten die ten grondslag liggen aan de traditionele methoden en de nieuwe mogelijkheden.

We behandelen om te beginnen de methoden die bekend zijn uit de literatuur. Ook de werkwijze bij de centrale eindexamens van het voortgezet onderwijs krijgt enige aandacht omdat die afwijkt van de bekende methoden en, wegens het belang van de examens, hier niet gemist mag worden. Daarna wordt onderzocht hoe de nieuwere mogelijkheden ons in staat stellen deze methoden verder te ontwikkelen. In het laatste deel van de paragraaf besteden we tevens aandacht aan andere onderscheidingen die in een cijferschaal kunnen worden aangebracht, zoals het onderscheid tussen (ruim) voldoende en goed.

13.5.1 Traditionele methoden van cesuurbepaling

Alle methoden voor cesuurbepaling steunen op het gecombineerde oordeel van een groep van 'deskundigen'. Deze deskundigen kunnen uit meerdere groepen afkomstig zijn. Natuurlijk uit het betreffende onderwijs zelf, maar ook de groepen die belang hebben bij het niveau en het aantal geslaagde kandidaten, zoals werkgevers, de overheid, de beroepsgroep, of het vervolg-onderwijs. De methoden voor cesuurbepaling

leveren de deskundigen een methode voor het systematisch specificeren van hun oordelen en het combineren daarvan voor het verkrijgen van een cesuur. Berk (1986) beschrijft 38 methoden voor cesuurbepaling. Hier bespreken we de meest bekende methoden. Al deze methoden hebben betrekking op een toets, dus niet op een itembank of itemdomein.

De methoden voor cesuurbepaling kan men indelen in een groep die alleen gebruik maakt van de 'grenspersoon' en de rest die de hele verdeling van cijfers in de populatie in het proces betrekt. Met een grenspersoon wordt een kandidaat bedoeld die zich precies op de grens tussen zakken en slagen bevindt. De methoden van Angoff, Ebel, Nedelsky en de 'borderline group' methode van Livingston en Zieky behoren tot de eerste groep die zich alleen op de grenspersoon richt. De methoden van Beuk, Hofstee en de 'contrasting groups' methode van Livingston en Zieky maken gebruik van de verdeling van de cijfers in de populatie.

Besliskunde

Omdat de cesuur het criterium is op grond waarvan men beslist of iemand slaagt of zakt, is het zinvol de vaststelling van een cesuur ook vanuit besliskundig oogpunt te bekijken (Hambleton & Novick; 1973, Van der Linden, 1982). De besliskundige benadering van de cesuurbepaling houdt expliciet rekening met het toevallige karakter van het toetscijfer, dat slechts een onnauwkeurig beeld van de ware vaardigheid van een persoon kan geven. Daarom moet er in de eerste plaats een conceptueel onderscheid worden gemaakt tussen de cesuur of het grenscijfer en de grensvaardigheid. Met het grenscijfer of de cesuur x_g bedoelen we de grens op de cijferschaal bijvoorbeeld de ruwe sore of $\hat{\theta}$. Een cijfer lager dan het grenscijfer betekent dat de kandidaat is 'gezakt'. Het onderliggende ware cijfer van een persoon v noemen we zijn vaardigheid en noteren we met ξ_v. De ware score τ is een voorbeeld van een vaardigheid, evenals de persoonsparameter θ op een Raschschaal. De grensvaardigheid wordt genoteerd als ξ_g. Een persoon v met vaardigheid $\xi_v < \xi_g$ verdient te zakken. Heeft persoon v een hogere vaardigheid dan verdient hij te slagen. Het is de bedoeling een cesuur zo te kiezen dat zo goed mogelijk onderscheid wordt gemaakt tussen degenen die verdienen te slagen en degenen die verdienen te zakken. Maar, omdat het (geobserveerde) cijfer niet alleen van de vaardigheid afhangt, maar behept is met een meetfout, lukt het niet altijd om een juiste beslissing te nemen. Zelfs met een optimaal gekozen cesuur kan het voorkomen dat iemand ondanks een vaardigheid $\xi < \xi_g$ toch een voldoende cijfer $x \geq x_g$ behaalt. Zo iemand slaagt onterecht. Als het omgekeerde

het geval is, zakt men onterecht. Beide foute beslissingen kan men in verschillende mate schadelijk vinden. Zo kan men het erger vinden om een ongeschikte kandidaatpiloot te laten slagen dan een geschikte te laten zakken. Ook kan men het erger vinden om een kandidaat met een vaardigheid ruim boven de grensvaardigheid te laten zakken, dan een kandidaat wiens vaardigheid vlak boven de grensvaardigheid ligt. De besliskunde levert een raamwerk om, gegeven een grensvaardigheid ξ_g, een grenscijfer x_g te vinden met een zodanige verhouding tussen de twee soorten verkeerde beslissingen, dat de beslissingen in een bepaalde zin optimaal zijn.

Een eerste stap naar de bepaling van een cesuur is derhalve het vaststellen van de grensvaardigheid ξ_g, de vaardigheid op de grens tussen geslaagd en gezakt. Daarna kan dan het optimale grenscijfer x_g worden bepaald. Helaas zijn veel methoden voor cesuurbepaling tot stand gekomen zonder besliskundige overwegingen. Dit ziet men alleen al daaraan dat het onderscheid tussen cesuur en grensvaardigheid niet wordt gemaakt. Die twee worden min of meer als identiek beschouwd. Toch is meestal duidelijk welke van de twee een bepaalde methode oplevert, een grenscijfer of een grensvaardigheid. We zullen daar steeds op wijzen.

Grensgroepmethoden

De grensgroepmethoden van Angoff, Ebel en Nedelsky, verlangen van deskundigen zich een idee te vormen over een grenspersoon. Vervolgens moeten zij voor ieder item in de toets een oordeel geven over de kans op een correct antwoord voor een grenspersoon. In de methode van Angoff (1971) wordt dit precies zo gevraagd, terwijl Ebel (1972) dit oordeel over items opbouwt in twee stappen. Eerst moet de deskundige de items groeperen volgens een tweeweg-classificatie naar moeilijkheid (makkelijk, gemiddeld, moeilijk) en relevantie voor de te meten vaardigheid (essentieel, belangrijk, acceptabel, twijfelachtig). Daarna wordt voor ieder van de twaalf categorieën items bepaald welk percentage een grenspersoon hiervan goed moet beantwoorden. Nedelsky's (1954) methode is alleen toepasbaar op meerkeuzevragen. De deskundigen moeten voor ieder item aangeven welke afleiders een grenspersoon als fout moet kunnen aanwijzen. Door de aanname dat het antwoord volgens toeval uit de overblijvende alternatieven wordt gekozen, volgt dan de kans op een goed antwoord voor een grenspersoon. Over het algemeen wordt aanbevolen om de deskundigen met elkaars oordelen te confronteren en erover te discussiëren. Daarna krijgen zij de gelegenheid eventueel hun oordelen te herzien.

Ieder van deze drie methoden leidt zo voor iedere deskundige, tot een score op de toets die zij verwachten van een grenspersoon. Deze scores kunnen worden gecombineerd tot de uiteindelijke cesuur door te middelen, eventueel na uitsluiting van extremen, of, door de mediaan te nemen.

Uit de beschrijving blijkt dat deze drie methoden de verwachte ruwe score en daarmee de ware score van een grenspersoon opleveren. Dit is derhalve een grensvaardigheid. Een kandidaat met een vaardigheid beneden de vaardigheid van een grenspersoon, de grens-vaardigheid, hoort te zakken. Deze oorspronkelijke drie methoden nemen echter zonder verdere besliskundige overwegingen de laagste score die niet kleiner is dan de grensvaardigheid als de cesuur. Deze cesuur is over het algemeen in besliskundige zin niet optimaal.

De borderline group methode vereist alleen dat een deskundige de grenspersonen aanwijst, zonder hun toetsresultaat te kennen. De mediaan van de toetsscores van deze groep is de cesuur voor deze deskundige. Noch Zieky (1987), noch Livingston en Zieky (1982) vermelden hoe de cesuren van de deskundigen worden samengevoegd. Men zou ook de mediaan kunnen nemen van de cijfers van alle grenspersonen, waarbij het cijfer van een persoon die door k deskundigen als grenspersoon is aangewezen, k keer meetelt. Een nadeel van deze methode is dat de groep grenspersonen meestal klein is.

Dit nadeel heeft de contrasting group methode niet. Een deskundige geeft voor iedere kandidaat aan of hij moet slagen of zakken, eventueel zonder zijn cijfer te kennen. Men mag echter hopen dat de kans om als voldoende te worden geclassificeerd sterk positief samenhangt met het cijfer. Voor ieder cijfer c telt men het aantal foute beslissingen: het aantal voldoende personen met een cijfer lager dan c en het aantal onvoldoende personen met een cijfer hoger dan c. De cesuur voor deze deskundige is het cijfer met het kleinste aantal foute beslissingen. Deze methode heeft als bijkomend voordeel dat kan worden meegewogen hoeveel erger men het vindt om iemand onterecht te laten slagen dan iemand onterecht te laten zakken. Stel dat men onterecht zakken (een voldoende persoon scoort lager dan c) tweemaal zo erg vindt als onterecht slagen. Men geeft dan de personen die de deskundige als voldoende beoordeelde het gewicht 2, de andere personen het gewicht 1, en summeert de gewichten van de personen, die bij een bepaalde cesuur onterecht als voldoende of onvoldoende worden geklassificeerd.

Uit deze laatste eigenschap blijkt een bepaalde besliskundige benadering. Zoals Van der Linden (1984) opmerkt, wordt hier dan ook een echte cesuur gekozen. Men kan dat als volgt zien. Het oordeel van de deskundige over een kandidaat identificeren we met het gegeven dat de (ware) vaardigheid van de beoordeelde persoon groter of kleiner is dan ξ_g, evenwel zonder dat er expliciet een ξ_g is gekozen. Bij de hier impliciet

gevolgde besliskundige procedure, gebaseerd op drempelutiliteit, is dat echter niet meer relevant zodra bekend is of de vaardigheid onder of boven ξ_g ligt. Drempelutiliteit wordt gebruikt wanneer men vindt dat de afstand van de vaardigheid van een persoon tot de grensvaardigheid voor het nemen van een beslissing van geen belang is. Het wordt bijvoorbeeld even erg geacht iemand onterecht te laten zakken ongeacht of deze nu een vaardigheid heeft net boven de grens-vaardigheid, of ver daarboven. Dit klinkt misschien vreemd, maar men dient hierbij wel te bedenken dat iemand met een vaardigheid ver boven de grensvaardigheid maar zeer zelden zal zakken.

De borderline group methode levert echter, in tegenstelling tot wat Van der Linden (1984) beweert, en in overeenstemming met wat hij 'common belief' noemt, wel degelijk een grens-vaardigheid ξ_g op. De verkregen grensscore is de mediaan van de geobserveerde scores van een groep van min of meer identieke (exchangeable) personen die de deskundige een vaardigheid gelijk aan ξ_g toedicht. Onder een model met normaal verdeelde fouten gegeven de ware score is deze mediaan gelijk aan de verwachte score gegeven ξ_g en derhalve gelijk aan ξ_g.

De laatste twee methoden hebben als nadeel dat de deskundigen de personen moeten beoordelen (natuurlijk) zonder kennis van hun toetsresultaat. Dit impliceert dat de deskundigen de personen op het gebied van de te meten vaardigheid op een andere manier goed moeten kennen. In de praktijk zal het erop neerkomen dat de 'groep' deskundigen beperkt zal zijn tot de eigen (vak)docent. Geen breed samengestelde groep van deskundigen dus.

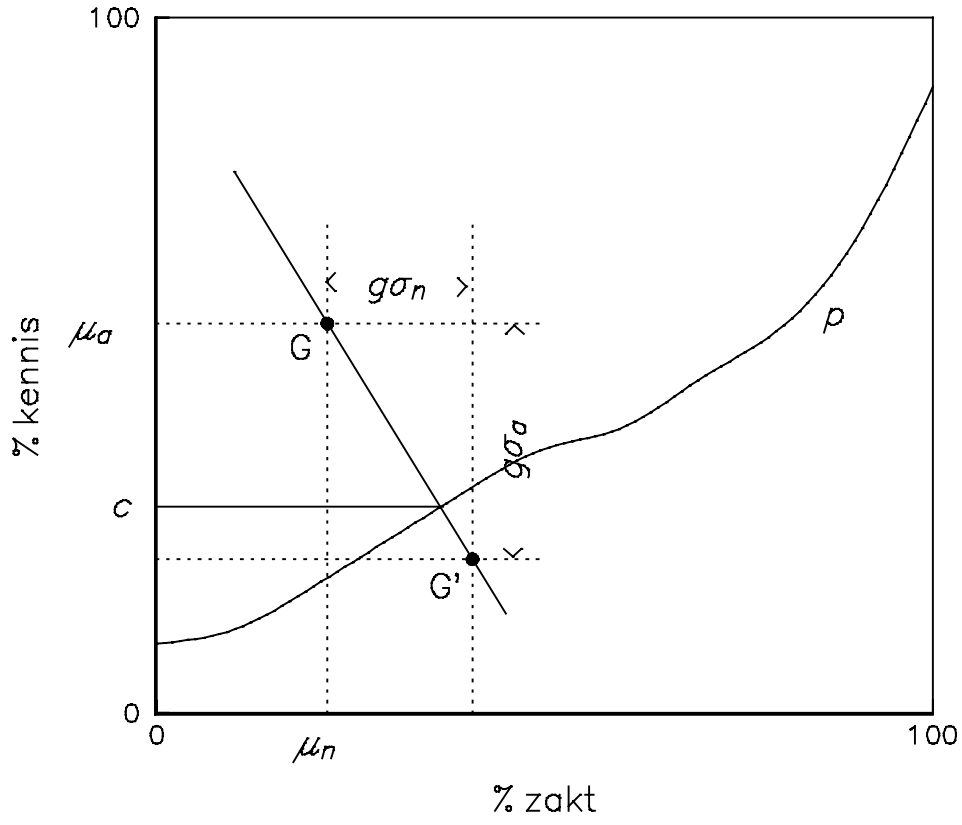
Compromismethoden

De zogenaamde compromismethoden kennen het zojuist genoemde nadeel niet. Iedereen die op de hoogte is met de betreffende vaardigheid en met de populatie van kandidaten kan hier als deskundige zijn oordeel geven. Maar het belangrijkste kenmerk van de compromis-methoden ten opzichte van al de voorgaande is dat er niet alleen naar een acceptabel prestatieniveau wordt gekeken, maar ook naar een acceptabel percentage kandidaten dat zakt. Men zoekt een compromis tussen een absolute cesuur en een normatieve cesuur. Bij een volledig normatieve cesuur telt alleen de verdeling van de cijfers. De cesuur wordt gelegd bij een vooraf bepaald percentage geslaagden, bijvoorbeeld 75%. In dat geval slagen de 75% hoogste cijfers, de overige 25% zakt. Overigens moet men zich niet voorstellen dat dit onderscheid erg hard is te maken. Bij de voorgaande methoden moesten de deskundigen zich immers een grenspersoon voorstellen. Het is haast niet te vermijden dat deze voorstelling mede wordt ingegeven

door een idee over de prestaties in de populatie. Zo spelen normatieve elementen daar ook mee. Vandaar dat we hier niet de strakke indeling volgen die wel eens wordt gemaakt tussen absoluut en normatief normeren bij het behandelen van methoden voor cesuurbepaling.

Bij de compromismethoden van Beuk en die van Hofstee worden de absolute cesuren eerst op een schaal gebracht die het percentage kennis in het getoetste domein weergeeft. Voor toetsen met open vragen is het percentage kennis bij cesuur *c* gelijk aan 100 x c/c_{max}%. Bij meerkeuzevragen wordt gecorrigeerd voor gokken. Als bijvoorbeeld het verwachte cijfer bij puur gokken gelijk is aan c_g, dan is het percentage kennis bij cesuur *c* gelijk aan $100(c\text{-}c_g)/(c_{max}\text{-}c_g)$. Op deze manier worden open vragen en meerkeuzevragen gelijk behandeld. De normatieve cesuur is het percentage van de kandidaten dat zakt.

Volgens de methode van Beuk (1984) wordt van iedere deskundige een absolute cesuur en een normatieve cesuur gevraagd. De deskundige moet de vraag beantwoorden welk percentage kennis hij precies voldoende vindt. Dit is zijn absolute cesuur. Vervolgens moet hij aangeven welk percentage hij vindt dat er moet zakken. Dit is zijn normatieve cesuur.

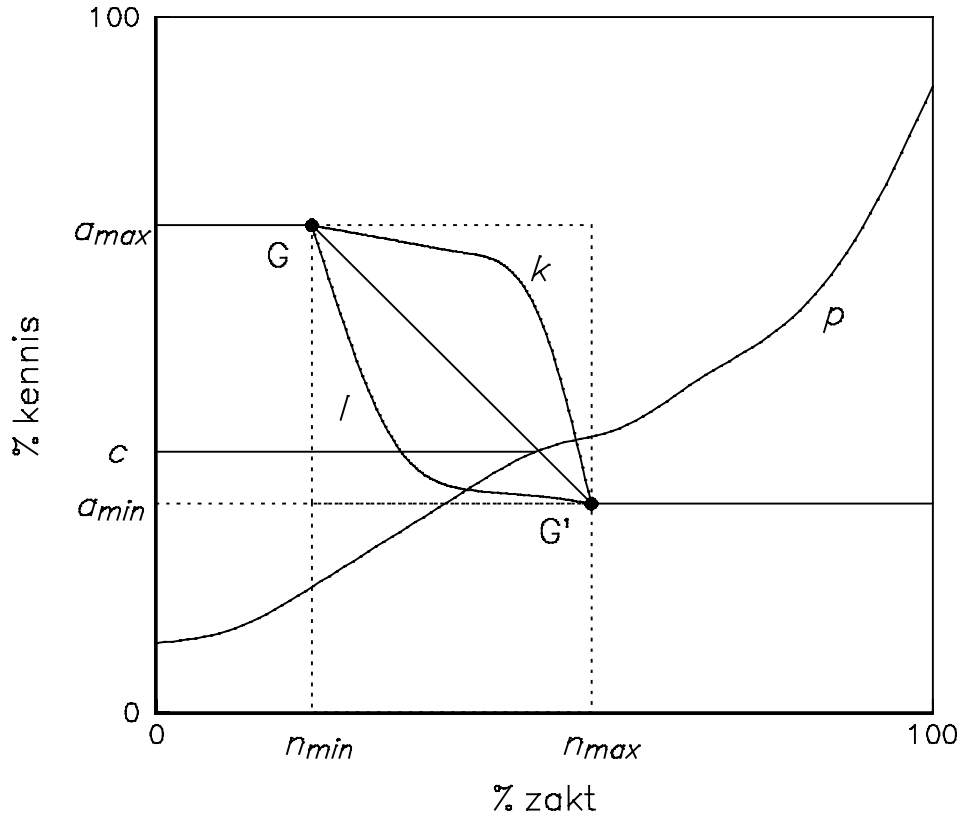


Figuur 13.5

Tabel 13.3
De gewenste absolute en normatieve cesuren van vijf fictieve deskundigen

	1	2	3	4	5	μ	σ	5σ
n% zakt	10	15	15	20	20	16	3.74	18.7
a% kennis	50	60	65	65	70	62	6.78	33.9

Daarna wordt het gemiddelde μ_a bepaald van de absolute cesuren van de deskundigen, en het gemiddelde μ_n van hun normatieve cesuren. In figuur 13.5 is op de horizontale as het percentage gezakten uitgezet en op de verticale as het percentage kennis. In de figuur is het punt (μ_n, μ_a) aangeven met de letter G. Het voorbeeld in figuur 13.5 is gebaseerd op vijf fictieve deskundigen waarvan de gegevens in tabel 13.3 zijn opgenomen. Deskundige 1 vindt bijvoorbeeld dat er 10% moet zakken en dat er minimaal 50% kennis moet worden gevraagd.
Nadat de toets is afgenomen bij de kandidatenpopulatie kent men de verdeling van de percentages kennis, zoals gemeten door de toets. Deze verdeling is in figuur 13.5 aangegeven met de lijn *p*. Een willekeurig punt (*n*, *a*) op lijn *p* betekent dat *n*% van de populatie *a*% kennis of minder heeft, en dus zou zakken als de cesuur bij *a*% zou liggen. Nu zal punt G over het algemeen niet op de lijn *p* liggen. Was dat wel het geval dan waren we klaar. Voor het verkrijgen van de cesuur moeten we vanaf G naar *p* toe schuiven in een richting waarbij de absolute en de normatieve cesuur in een bepaalde zin gelijkwaardig veranderen. Om het begrip 'gelijkwaardig' een precieze inhoud te geven, kiest Beuk voor de mate waarin de deskundigen het onderling eens zijn over beide cesuurtypen. Daartoe berekenen we de standaarddeviaties σ_n van de normatieve cesuren en σ_a van de absolute cesuren. In het voorbeeld in tabel 13.3 is σ_n = 3.74 en σ_a = 6.78. Het punt G' is nu gedefinieerd als $(\mu_n + g\sigma_n, \mu_a - g\sigma_a)$ voor een willekeurige *g* (in figuur 13.5 is *g* = 5). We bepalen vervolgens het snijpunt van GG' en *p*. Dit snijpunt bepaalt het compromis tussen absolute en normatieve cesuurwensen van de deskundigen: het minimaal geëiste kennispercentage *c* om te slagen. Het laagste cijfer op de toets dat hoort bij een kennispercentage groter of gelijk aan *c* is de laagste voldoende.



Figuur 13.6 De cesuur bepaling volgens Hofstee

De methode Hofstee (1977, 1983; De Gruijter, 1985), weergegeven in figuur 13.6, vraagt van elke deskundige twee absolute cesuren en twee normatieve cesuren. Ten eerste de minimum absolute cesuur $a_{\min}$, het percentage kennis dat minimaal wordt geëist ook al zou iedereen zakken en de maximum absolute cesuur $a_{\max}$, het percentage dat men maximaal eist ook al zou iedereen slagen. Vervolgens moet de deskundige het percentage $n_{\max}$ gezakten aangeven dat hij binnen de absolute kennisgrenzen maximaal accepteert. Als $n_{\max}$ of minder procent van de populatie $a_{\min}$ of minder kennis zou hebben dan zou hij zijn eisen tot $a_{\min}$ laten zakken. Tenslotte moet hij het percentage $n_{\min}$ opgeven dat hij minimaal accepteert binnen $a_{\min}$ en $a_{\max}$. Als het percentage gezakten bij $a_{\max}$ als cesuur lager uitvalt dan $n_{\min}$ dan wordt $a_{\max}$ als cesuur genomen. Zij nu G het punt $(n_{\min}, a_{\max})$ en G' het punt $(n_{\max}, a_{\min})$ dan noemt Hofstee het lijnstuk GG' de verzameling acceptabele compromissen. Het snijpunt van *p* en GG' levert dan het feitelijk compromis met cesuur *c*.

Drie opmerkingen over de methode Hofstee. Ten eerste zegt geen enkele van de geraad-pleegde publikaties iets over de manier waarop de oordelen van meer dan een deskundige worden gecombineerd. Men kan op beide assen het minimum van de minima en het maximum van de maxima nemen, maar ook hun gemiddelde of mediaan,

en daarmee de lijn GG' bepalen. Mocht het maximum van de minima kleiner zijn dan het minimum van de maxima, dan zou men ook daarmee de cesuur kunnen bepalen. In dat geval zijn alle deskundigen tevreden met de cesuur als *p* het lijnstuk GG' snijdt. Men zou ook voor iedere deskundige een cesuur kunnen bepalen en daarvan het gemiddelde of de mediaan kiezen. De tweede opmerking betreft de situatie die zich voordoet wanneer *p* het lijnstuk GG' niet snijdt. Mills en Melican (1987) vinden dat er dan opnieuw een cesuur moet worden vastgesteld. Echter, uit de definities van $a_{\min}$ en $a_{\max}$ blijkt dat dan, afhankelijk van heel slechte of juist heel goede prestaties, respectievelijk $a_{\min}$ of $a_{\max}$ de cesuur zal moeten zijn. De derde opmerking betreft de tamelijk willekeurige keuze van de rechte lijn GG' als verzameling acceptabele compromissen. GG' is de lijn waarin normatieve en absolute overwegingen precies gelijk worden gewogen. In principe is echter ieder punt acceptabel dat ligt in de rechthoek waarvan GG' de diagonaal is. In figuur 13.6 representeert de lijn *k* een situatie waarin men aan de absolute cesuur hogere prioriteit geeft dan aan de normatieve, terwijl dit voor de lijn *l* andersom is.

Van deze twee compromismethoden lijkt, ondanks de gesignaleerde onduidelijkheden, die van Hofstee het meest rationeel. In de methode van Hofstee geeft iedere deskundige zijn onderhandelingsruimte duidelijk aan. In de methode van Beuk, daarentegen, worden twee zaken vermengd die niet vermengd lijken te mogen worden. De 'gelijkwaardige' verandering van normatieve en absolute wensen van de deskundigen en de mate waarin zij het onderling eens zijn worden als hetzelfde beschouwd. Hoe meer zij het eens zijn over een van de twee cesuren des te kleiner de relatieve verschuiving. Over het algemeen zal een gelijkwaardige bijstelling echter door andere factoren zijn bepaald. Een klein voorbeeld kan dit verduidelijken. Stel er zijn twee deskundigen die beiden een normatieve cesuur van 25% kiezen, maar ieder een verschillende absolute cesuur, respectievelijk 60% en 70%. Volgens de methode Beuk zakt in dit geval altijd 25% van de kandidaten, ook als de absolute cesuur daarmee bijvoorbeeld op 40% of nog lager zou komen te liggen. Waarschijnlijk vinden de deskundigen 40% kennis als minimale eis niet acceptabel. Zij zouden beiden liever een groter percentage kandidaten laten zakken om zo dichter in de buurt van hun gewenste absolute cesuren te komen.

Het zou beter zijn wanneer iedere deskundige, naast zijn voorkeurspunt, ook twee richtingen van gelijkwaardige verandering zou preciseren, een richting voor een verhoging en een voor een verlaging van de absolute cesuur. Men zou dan het gemiddelde voorkeurs-punt van de deskundigen kunnen bepalen, en ook de twee gemiddelde richtingen. Vervolgens kan men de twee lijnen met deze richtingen vanuit het ideaalpunt trekken en het snijpunt met *p* bepalen voor de cesuur. Een voorbeeld

kan dit verduidelijken. De deskundige ziet het bepalen van de cesuur als een onderhandeling tussen hemzelf en een vertegenwoordiger van de kandidaten. De deskundige bepaalt zijn positie voor de onderhandelingen als volgt. Hij vindt dat 50% kennis is vereist en accepteert daarbij dat 10% van de kandidaten zakt. Mochten er evenwel bij 50% kennis meer dan 10% van de kandidaten zakken dan is hij bereid de absolute cesuur te laten zakken, maar de kandidatenpopulatie moet voor iedere 1% verlaging van de kenniseis genoegen nemen met 9% meer gezakten dan de voorgestelde 10%. Een verlaging van de kenniseis weegt dus negen keer zo zwaar als een verhoging van de normatieve eis. Mochten er bij 50% kennis minder dan 10% van de kandidaten zakken dan is er ruimte voor een kwaliteitsverhoging van het diploma. De deskundige is bereid om in ruil voor iedere 1% verhoging van de absolute cesuur 1% minder kandidaten te laten zakken dan de voorgestelde 10%.

De Gruijter (1985) doet een voorstel waar dit voorstel op het eerste gezicht enigszins op lijkt. Hij hanteert evenwel geen richtingen van verandering maar een Euclidische metriek. Deze metriek is gebaseerd op de relatieve onzekerheid die een deskundige heeft ten aanzien van beide cesuren, niet aan het relatieve belang dat wordt gehecht aan een verhoging of verlaging. In die zin lijdt zijn voorstel aan dezelfde conceptuele verwarring als de methode van Beuk. Er wordt eveneens geen onderscheid gemaakt tussen onzekerheid en bereidheid tot verandering. De Gruijter substitueert alleen een individuele onzekerheid voor de collectieve onzekerheid van Beuk. Bovendien is 'onzekerheid' symmetrisch, zodat geen onderscheid wordt gemaakt tussen verhoging en verlaging van de absolute cesuur. Doordat deze methode geen richting van verandering gebruikt maar een afstandsmaat, heeft zij de vreemde eigenschap dat het kan voorkomen dat de absolute cesuur flink wordt verlaagd, zonder dat daar een noemenswaardige verhoging van het percentage gezakten tegenover staat. Immers, als p onder het ideaalpunt doorloopt en daar niet of nauwelijks stijgt, dan kan het punt op p met de kleinste afstand tot het ideaalpunt, daar bijna loodrecht onder liggen.

Het aanwijzen van een minimaal vereist percentage kennis, komt in het besliskundig raamwerk uiteraard overeen met het aanwijzen van de grensvaardigheid ξ_g. Echter de invloed van de verdeling van de cijfers op de uiteindelijke cesuur, het normatieve element in deze methoden, is precies omgekeerd aan de invloed van het normatieve element in besliskundige procedures. Van der Linden (1984) wijst erop dat besliskundige procedures er toe leiden dat hoe hoger de prestaties in een groep zijn hoe lager de cesuur zal uitvallen. Dit is een fenomeen dat voortvloeit uit het Bayesiaanse karakter van besliskundige procedures.

De centrale eindexamens

Bij de centrale eindexamens wordt de cesuur niet met een van de eerder genoemde methoden bepaald. Hoewel er bij de examens, afhankelijk van het type vragen, zes verschillende gevallen van cesuurbepaling worden onderscheiden, wordt in essentie een enkele methode gevolgd. Om te beginnen wordt er voor ieder examen, voordat de scoreverdeling bekend is op basis van een inschatting van de moeilijkheid van het examen, de laagste voldoende ruwe score gekozen. Als de scoreverdelingen bekend zijn bespreken deskundigen hoe acceptabel deze voorafgekozen cesuur is gezien het percentage kandidaten dat zou zakken bij deze cesuur. Als het examen onverhoopt moeilijker uitvalt dan gedacht, en dus een hoog percentage gezakten zou opleveren bij de vooraf vastgestelde cesuur, dan kan de cesuur binnen bepaalde restricties worden verlaagd. Wanneer het examen makkelijker blijkt dan verwacht, en er dus veel leerlingen slagen bij de vooraf gekozen cesuur, dan mag men de voorafgekozen cesuur meestal niet verhogen.

De cesuurbepaling bij de examens komt het dichtst in de buurt van de compromismethoden. Zij mist echter een duidelijk omschreven procedure voor het afwegen van absolute en normatieve wensen. De voorafgekozen cesuur lijkt het meest op een minimaal vereist percentage kennis, een grensvaardigheid ξ_g. Ook de richting van de invloed van het niveau van de prestatie van de groep lijkt enigszinsop die van de compromismethoden. Een lage prestatie kan worden beloond met een verlaging van de cesuur. Het bestraffen van een hoge prestatie is daarentegen meestal niet toegestaan.

Naar aanleiding van een advies van het Cito over normhandhaving, is er een onderzoek gedaan (Inspectierapport, 1992) naar de gelijkwaardigheid van de examencijfers over een aantal jaren heen. Hieruit bleek dat de moeilijkheid van de examens van jaar tot jaar sterk uiteen liep. Dit is natuurlijk niet zo erg. Door equivalering kan hiervoor immers worden gecorrigeerd. Er bleek echter ook dat de cesuren van jaar tot jaar met sterk verschillende vaardigheden correspondeerden, ondanks de correcties van de cesuren door de deskundigen. Het rapport besluit dan ook met enkele suggesties voor verbetering. Pretesting en calibratie op een schaal met de eerdere examens van hetzelfde type maken er deel van uit.

Ter afsluiting van deze paragraaf behandelen we nog een aardig technisch probleem dat bijvoorbeeld bij examens ontstaat bij het toekennen van cijfers. Ruwe scores, en dus percentages goed op de toets, worden vaak afgebeeld op de gebruikelijke cijferschaal van 1 tot 10 via een of meer lineaire transformaties. De cijfers 1.0 tot en met 10.0 worden dan op een decimaal nauwkeurig gerapporteerd. Voor het vinden van de gewenste lineaire transformatie(s) gaat men als volgt te werk. Men kiest een score r_1,

die exact op het cijfer 5.5 (de laagste voldoende) moet worden afgebeeld. Verder wordt een score r_0 gekozen die op het laagste cijfer 1.0 wordt afgebeeld, en een score r_2 voor het hoogste cijfer 10.0. Dit levert twee lineaire transformaties van scores naar cijfers op, een naar de cijfers 1.0 t/m 5.5 en een naar de cijfers 5.5 t/m 10.0. Bij examens is het exacte cijfer dat men krijgt (uiteraard) erg belangrijk. Een tiende punt meer of minder kan het verschil tussen zakken of slagen uitmaken voor een bepaald vak. Bovendien is de procedure volgens welke de cijfers uit de scores worden berekend openbaar. Men kan zich dus niet veroorloven dat cijfers een tiende punt hoger of lager uitvallen door toevallige afwijkingen die ontstaan door de binaire floating point (drijvende komma) representatie van reële getallen in de computer. Deze ongewenste toevallige effecten zijn te vermijden door een algoritme voor de transformatie te gebruiken zonder floating point-getallen en -operaties. Het algoritme mag alleen met integer (gehele) getallen en integer operaties werken. Omdat de cijfers op 1 decimaal nauwkeurig worden gerapporteerd, verkrijgen we integer cijfers *f* door de oorspronkelijke cijfers met 10 te vermenigvuldigen waardoor *f* integer waarden aanneemt van 10 t/m 100. Beeld r_0 af op het cijfer f_0 en r_1 op f_1. Zij $a = f_1 - f_0$, $c = r_1 - r_o$ en $ar_1 - cf_1$, dan kan de lineaire transformatie $f = g(r)$ van scores *r* naar cijfers *f* geschreven worden met alleen integer getallen. De integer representatie *G(r)* van *g(r) = f = (ar + b)/c* is dan gegeven door:

$$cf \leq ar + b < c(f + 1). \tag{13.4}$$

Gegeven een score $r = r'$ zoekt men een f' die aan deze ongelijkheden voldoet. Als $ar' + b$ dichter bij cf' ligt dan bij $c(f' + 1)$ dan is $G(r') = f'$ anders is $G(r') = f' + 1$ ('afronden' gebeurt in het voordeel van de student). Cijfers kleiner dan het minimum (10) worden als 1.0 en cijfers groter dan het maximum (100) worden als 10.0 gerapporteerd. Bij alle overige cijfers wordt er een punt ingevoegd. Bijvoorbeeld als $f = 56$ wordt het gerapporteerde cijfer 5.6. Door gebruik te maken van integerdeling (genoteerd met \) is het eenvoudig een algoritme te construeren dat de functie *G(r)* berekent. Immers de f' die voor $r = r'$ voldoet aan de ongelijkheden in formule (13.4) is $f' = (ar' + b) \backslash c$.

13.5.2 Cesuurbepaling en overige cijfers binnen itemresponstheorie

Alle hierboven genoemde methoden voor cesuurbepaling kunnen gemakkelijk worden gegeneraliseerd naar een gecalibreerde itembank. Op het eerste gezicht lijkt deze

opmerking niet ter zake, omdat veel van de bovengenoemde methoden nu juist bedoeld waren voor de situatie dat er nog geen empirische gegevens over de items, of de toets bekend zijn. Laat staan dat men de beschikking heeft over een gecalibreerde itembank. Tegenwoordig zullen er echter bijna altijd empirische gegevens van de doelgroep beschikbaar zijn over items uit een leerstofdomein. Met deze gegevens kan men de items calibreren en de vaardigheids-verdeling van de doelgroep schatten. Op basis van deze gecalibreerde itembank kan men een grensvaardigheid ξ_g bepalen. De vaardigheidsverdeling van de doelgroep en een geschikte besliskundige procedure leveren nu voor iedere toets een optimale cesuur. Wanneer de toets of het examen moet bestaan uit nieuwe, niet eerder gebruikte items, dan kunnen die later gecalibreerd aan deze itembank worden toegevoegd.

Voor alle methoden van cesuurbepaling kiest men uit de itembank een reeks items waarvan men verwacht dat die de vaardigheid in de buurt van de nog nader te bepalen grens-vaardigheid ξ_g goed zal meten. Deze verzameling items noemen we de referentietoets. We veronderstellen dat het model voor de referentietoets een strikt monotone regressiefunctie $r(\theta)$ van de latente vaardigheid naar de verwachte ruwe score definieert. Voor het Raschmodel en OPLM is dit altijd het geval. Daarmee bestaat dus ook de inverse functie $r^{-1}(r) = \theta(r)$ van scores naar de latente vaardigheid. De methoden van Angoff, Ebel en Nedelsky leveren een verwachte ruwe score r_g voor de grenspersoon, en daarmee de minimaal voldoende vaardigheid $\theta_g = \theta(r_g)$. De borderline group methode van Livingston en Zieky is gebonden aan een groep personen die bij de deskundigen bekend zijn, echter ook deze methode kan eenmalig worden toegepast voor het vinden van een minimaal vereiste θ_g. De contrasting groups methode resulteert niet in een grensvaardigheid, maar in een echte cesuur op de referentietoets. Willen we bij deze cesuur een grensvaardigheid verkrijgen, dan moet de beslissingsprocedure worden omgekeerd. Normaal zoeken we een optimale cesuur bij een gegeven grensvaardigheid. Nu moeten we een grensvaardigheid vinden waarvoor deze cesuur op de referentietoets optimaal is.

Met een gecalibreerde itembank en een schatting van de verdeling van de vaardigheden kunnen de beide compromismethoden worden vervangen door een veel directer alternatief. Bij iedere θ is niet alleen het kennispercentage op de referentietoets bekend, maar ook het percentage kennis op de hele itembank. Bovendien staat de verdeling van vaardigheden in de doelgroep ter beschikking. Daardoor kent men bij ieder kennispercentage, dus bij iedere mogelijke grensvaardigheid, het percentage in de doelgroep dat verdient te zakken. Men kan er derhalve mee volstaan om iedere deskundige direct op de curve *p* in de figuren 13.5 en 13.6 zijn combinatie van absolute en relatieve cesuur te laten aangeven. Voor het

combineren van verschillende keuzen op de lijn *p* zijn dan meerdere voor de hand liggende oplossingen te bedenken. Een mogelijk probleem bij deze methode is, dat het percentage werkelijk gezakten bij een optimale cesuur over het algemeen niet gelijk zal zijn aan het percentage dat verdient te zakken.

Een gecalibreerde itembank kan ook worden ingezet voor het rapporteren op de schalen die behandeld zijn in paragraaf 13.2. De cumulatieve verdelingen, zoals centielen bij een geschatte vaardigheid zijn eenvoudig te berekenen. De informatiefunctie van de toets en de verdeling van de vaardigheden in de doelgroep bepalen de verdeling van de vaardigheids-schatter. Ook de genormeerde lineaire transformaties zijn daarmee eenvoudig op de latente schaal af te zetten. Alleen met de genormaliseerde schalen moeten we oppassen in verband met de eigenschap 'intervalniveau'. Hierboven werd gesteld dat de T-schaal (en de C-schaal en de Stanines) intervalniveau heeft en per definitie normaal is verdeeld in de referentie-populatie. Als de latente vaardigheidsschatter ook normaal is verdeeld, dan is de T-schaal een lineaire transformatie van de latente vaardigheidsschatter. Is deze laatste duidelijk niet normaal verdeeld, dan hebben we twee schalen van verondersteld intervalniveau, die geen lineaire transformatie van elkaar zijn. De conclusie moet zijn dat minstens een van de twee schalen er geen aanspraak op kan maken van intervalniveau te zijn.

Vele schoolgeneraties lang is het al gebruikelijk om de prestaties in ieder geval (ook) te rapporteren op een zogenaamde cijferschaal. In Nederland is dat de bekende schaal van 1 tot en met 10. Naast het rapporteren van een percentiel of T-schaalwaarde moet er dan ook een transformatie worden geconstrueerd van vaardigheidsschattingen naar de cijferschaal. We kunnen hier kort over zijn. In principe is iedere cijferovergang, bijvoorbeeld die van 7.9 naar 8.0, op een analoge manier te behandelen als de grensvaardigheid voor de cesuur. Alle methoden die men gebruikt voor het vaststellen van een grensvaardigheid, zijn ook toepasbaar voor de bepaling van een andere vaardigheidsgrens. Gelukkig hoeft niet voor alle 90 cijferovergangen op de schaal van 1.0 tot 10.0 afzonderlijk een grensvaardigheid te worden vastgesteld. Enkele belangrijke overgangen, zoals die tussen 7.9 en 8.0, of tussen 4.4 en 4.5, kan men zorgvuldig behandelen. De overige overgangen kan men vervolgens vastleggen door (lineaire) interpolatie. Is de cijferschaal eenmaal vastgelegd, dan kan vervolgens voor vele toekomstige examens die uit deze itembank worden samengesteld dezelfde automatisch geëquivaleerde cijferschaal worden gehanteerd.

Op basis van deze cijferschaal kunnen vervolgens de minimale psychometrische kwaliteiten worden gespecificeerd waaraan het examen in onze ogen moet voldoen. Uiteraard is de grens tussen voldoende en onvoldoende het punt waarnaar onze

grootste zorg zal uitgaan. Een kandidaat met een vaardigheid groter dan de minimale voldoende vaardigheid moet een zo klein mogelijke kans hebben om onvoldoende te scoren. Het is natuurlijk erger wanneer een kandidaat die een 7.0 verdient beneden de 5.5 scoort, dan wanneer dit een kandidaat overkomt die een 5.6 verdient. Zoeken we eerst het vaardigheidsinterval dat begrensd wordt door de ondergrens voor de 7.0 en de ondergrens voor de 7.1. Het midden, $\theta_{7.0}$, van dit interval representeert de vaardigheid van de kandidaten die een 7.0 verdienen. De kans dat met de vaardigheid $\theta_{7.0}$ beneden de 5.5 wordt gescoord neemt af naarmate het examen meer informatie bevat tussen de ondergrens van het interval 5.5 en $\theta_{7.0}$, terwijl tevens de informatie op $\theta_{7.0}$ zo laag mogelijk moet zijn (Verstralen & Verhelst, 1991). Als we er ook waarde aan hechten dat iemand die een 8.0 verdient een zo klein mogelijke kans heeft een 6.5 of minder te halen, dan kunnen deze twee wensen elkaar een beetje in de weg zitten. Verder kan uiteraard het aantal items niet onbeperkt groot gekozen worden. Er is programmatuur (Verschoor, 1990) die kan helpen bij het expliciteren van onze wensen met betrekking tot de lokale meetnauwkeurigheid van het examen en het vaststellen van de minimale informatiefunctie die daarbij hoort. Bij iedere informatiefunctie I kan worden gekeken hoeveel items ongeveer nodig zijn voor een toets met een informatiefunctie die groter is dan I. Bovendien kan worden beoordeeld of de conditionele verdelingsfunctie van een selectie van de cijfers gegeven θ, bijvoorbeeld $\theta = \theta_{7.0}$, acceptabel is. Als de selectie de cijfers 7.0 en 5.4 bevat, kunnen we zien hoe groot de kans is dat iemand die een 7.0 verdient, onvoldoende scoort. Hetzelfde kan ook voor andere vaardigheden worden bekeken. We kunnen bijvoorbeeld nagaan wat de kans is dat iemand die een 6.5 verdient een onvoldoende scoort. Maar ook hoe groot de kans is dat iemand die een 5.0 verdient een 6.0 of hoger haalt. Als we op deze manier onze psychometrische wensen, binnen de randvoorwaarden van het examen hebben vorm gegeven, kunnen we een examen samenstellen dat aan deze psychometrische eisen en de specificaties zoals neergelegd in een toetsmatrijs voldoet.

Gegeven een toets uit een Rasch- of OPLM-gecalibreerde itembank, kan er een functie $\hat{\theta}(s)$ van (gewogen) toetsscores naar vaardigheidsschattingen worden gevonden. We hadden met de cijferintervallen al een functie $c(\theta)$ van θ naar de cijfers van 1.0 tot en met 10.0 die θ afbeeldt op het cijfer van het interval waartoe het behoort. De samenstelling $d(s) = c(\hat{\theta}(s))$ genereert dan een transformatietabel van scores naar cijfers. Voor het bevorderen van een goed begrip van deze cijfers, kan bij ieder cijfer het centiel in een normgroep en het scorepercentage op de itembank en op de toets vermeld worden.

In de bovenbeschreven procedure voor de transformatie van scores naar cijfers is geen rekening gehouden met besliskundige aspecten. Hoewel dit in de praktijk niet

gemakkelijk zal zijn, is het principe niet ingewikkeld. Men bepaalt voor ieder van de 91 classificaties een utiliteitsfunctie $U_f(\theta)$ ($f = 1.0, \ldots, 10.0$). Met $U_f(\theta)$ geeft men aan welke waarde men eraan hecht om iemand met vaardigheid θ te classificeren als *f*. Men doet er uiteraard verstandig aan om in de serie functies U_f enige systematiek aan te brengen zodat er niet voor iedere *f* afzonderlijk nagedacht hoeft te worden. Bij iedere score *r* op de toets wordt de a posteriori verdeling g_r van θ bepaald. Vervolgens zoekt men de classificatie *f* met de grootste verwachte utiliteit over g_r. Eventueel kan men andere criteria hanteren in plaats van de grootste verwachte utiliteit (Berger, 1980).

Uiteraard hoort bij de resultaten van een meetprocedure ook een indicatie van de nauwkeurigheid. Gegeven een OPLM-gecalibreerd examen *b* en een vaardigheid θ_{vb} voor persoon *v* op deze OPLM-schaal, dan is de score op het examen een toevalsvariabele met een conditionele verdeling gegeven θ_{vb}. Omdat $\hat{\theta}_{vb} = \hat{\theta}(s_{vb})$ is ook $\hat{\theta}$ een toevalsvariabele met een conditionele verdeling gegeven θ_{vb}. De standaarddeviatie van deze verdeling is de lokale standaardschattingsfout van $\hat{\theta}_{vb}$. Deze lokale standaardschattingsfout kan ook rechtstreeks uit de informatiefunctie van het examen worden berekend als $I(\theta_{vb})^{-½} \approx I(\hat{\theta}_{vb})^{-½}$, en dus ook een 50% of 95% betrouwbaarheidsinterval. Via de hierboven genoemde transformatie *c(.)* verkrijgen we dan de overeenkomstige betrouwbaarheidsintervallen op de cijferschaal en tevens op de schalen die de interpretatie ondersteunen zoals het centiel in de referentiepopulatie. Tabel 13.5 bevat een voorbeeld van een rapportage voor de vakken Duits, Frans en Engels.

Tabel 13.5

Rapportage van cijfers en hun nauwkeurigheid van alle vakken gezamenlijk

Vak	Punt-schatting	Cijfer →					
	*	5.0	6.0	7.0	8.0	9.0	10.0
Duits	6.6		--++*++--				
Frans	6.3		--++*++--				
Engels	7.0			--++*++--			

De symbolen in tabel 13.5 hebben de volgende betekenis:

* : puntschatting (ook als getal afgedrukt onder *),

++*++ : 50% betrouwbaarheidsinterval,

--++*++-- : 95% betrouwbaarheidsinterval.

Daarna kunnen, zoals in tabel 13.3, voor ieder vak afzonderlijk, bijvoorbeeld voor Duits in tabel 13.6, de waarden van de cijfers op overige schalen, zoals norm- en

beheersingsschalen, worden gegeven waarmee de betekenis van de cijfers wordt verduidelijkt. De interpretatie van een dergelijk rapport is behandeld onder tabel 13.3.

Tabel 13.6

Rapportage per vak over meerdere schalen

Vak	Punt-schatting	Schaalwaarde →					
Duits	*						
score % itembank	72	52	66	78	86	93	99
score % examen	67	54	62	69	79	92	98
% populatie	74	63	69	77	87	98	100
cijfer	6.6	5.0	6.0	7.0	8.0	9.0	10.0
			– – + + * + + – –				

Het combineren van de resultaten op verschillende examens tot een zak/slaag-beslissing

Examens bestaan in het algemeen uit een reeks onderdelen die ieder een bepaald schoolvak als onderwerp hebben. In verband met de traditionele toekenning van diploma's, of meer in het algemeen voor een globale niveau-aanduiding, moeten de resultaten op al deze vakken worden gecombineerd tot een eindbeslissing. Over het algemeen bestaan er voor het combineren van de examenresultaten tot een beslissing over het toekennen van een bepaald diploma, allerlei compensatieregelingen. Al deze regelingen zijn echter vaak ad hoc, zodat meer gefundeerde methoden overwogen kunnen worden. Hieronder wordt een mogelijke aanpak geschetst.

Een Bayesiaanse benadering lijkt het meest aangewezen. Zij $\theta = (\theta_1,...,\theta_I)$ een vector van latente vaardigheden op de verschillende onderdelen $i, (i = 1,...,I)$ van het gehele examen. Zij $f(\theta)$, de a priori verdeling van θ, en $f(\theta \mid \boldsymbol{s})$ de a posteriori verdeling van θ, gegeven de vector $\boldsymbol{s} = (s_1,...,s_I)$ van (gewogen) scores op de I examenonderdelen. Noteer de door het model (OPLM) gegeven conditionele verdeling van de scores gegeven θ met $g(\boldsymbol{s} \mid \theta)$ en de marginale scoreverdeling met $g(\boldsymbol{s})$, dan is volgens de regel van Bayes:

$$f(\theta \mid \boldsymbol{s}) = \frac{g(\boldsymbol{s} \mid \theta) f(\theta)}{g(\boldsymbol{s})}. \tag{13.5}$$

Formule (13.5) kan als volgt uitgangspunt zijn voor het combineren van toetsuitslagen tot een beslissing over het algehele niveau.

Zij $\theta^{(5.5)} = (\theta_1^{(5.5)}, ..., \theta_I^{(5.5)})$ de vector van ondergrenzen van de intervallen voor de cijfers 5.5 op de verschillende examenonderdelen en $\Omega^{(5.5)}$ de deelverzameling van $\mathbb{R}^I$, waarin voor ieder element geldt dat alle componenten groter zijn dan het overeenkomstige element in $\theta^{(5.5)}$: $\theta \in \Omega^{(5.5)}$ als voor alle i $\theta_i > \theta_i^{(5.5)}$, dan is

$$P_{\boldsymbol{s}} = P\left(\theta > \theta^{(5.5)} \mid \boldsymbol{s}\right) = \int_{\Omega^{(5.5)}} f(\theta \mid \boldsymbol{s})\, d\theta,$$

de mate waarin we geloof kunnen hechten aan de bewering dat een persoon met scorevector $\boldsymbol{s}$ op alle onderdelen van het examen minstens een voldoende vaardigheid heeft bereikt, en $1 - P_{\boldsymbol{s}}$ dat dit op minstens een van de onderdelen niet het geval is. De ondergrens voor $P_{\boldsymbol{s}}$ waarboven tot toekenning van het diploma wordt besloten, is een subjectief besluit, waarin niet alleen de ernst van onterecht zakken of slagen moet worden verwerkt. Ook is enige ervaring met deze procedure vereist voor een afgewogen keuze.

Omdat het hier een beslissing over zakken of slagen betreft is er ook veel voor te zeggen om een besliskundige benadering te volgen, bijvoorbeeld op basis van de verwachte à posteriori utiliteit. Men kiest voor beide klassen zakken en slagen respectievelijk de utiliteitsfuncties $U_0(\theta)$ en $U_1(\theta)$ en berekent

$$U_i(\boldsymbol{s}) = \int_{\mathbb{R}^I} U_i(\theta)\, f(\theta \mid \boldsymbol{s})\, d\theta$$

voor $i = 0, 1$. Als $U_0(\boldsymbol{s}) > U_1(\boldsymbol{s})$ dan zakt een kandidaat met scorevector $\boldsymbol{s}$, anders slaagt hij. Het grootste probleem van deze benadering is de keuze van de beide utiliteitsfuncties. Men zou om te beginnen de utiliteitsfuncties kunnen bestuderen die impliciet waren in de beslisregels die bij vroegere examens zijn gehanteerd (Lord, 1983b).

Formule (13.5) kan ook de basis zijn voor nauwkeuriger puntschattingen van θ, dan wanneer de schatting per schaal afzonderlijk gebeurt. De verschillende examenonderdelen zullen immers in de a priori verdeling over het algemeen onderling gecorreleerd zijn. Het is dan evenwel beter en helderder om voor de itemcalibratie en de schattingen van persoons-parameters een multidimensioneel IRT-model te kiezen. Het is te verwachten dat dan met aanzienlijk minder dimensies kan worden volstaan

dan het aantal deelexamens, hetgeen in een overzichtelijker beschrijving van de data resulteert.

13.6 Conclusie

Over het algemeen wordt er bij de rapportage van testresultaten in voldoende mate gebruik gemaakt van de methoden en middelen die in de voorgaande paragrafen zijn besproken. Te vaak echter is het schoolrapport en de rapportage van eindexamenresultaten hierop een uitzondering. Ook de kwaliteiten van deze rapporten kunnen worden beoordeeld volgens de criteria die in het voorafgaande zijn besproken. Gezien de spaarzame informatie die het traditionele school- en eindexamenrapport biedt, valt echter niet te ontkennen dat het meten en rapporteren van het bereikte niveau van leerlingen in onze schoolcultuur geen hoge prioriteit heeft. Voor een deel is dit het gevolg van een aversie tegen het beoordelen en vergelijken van kinderen. Wat zou er echter tegen zijn om bijvoorbeeld normgegevens op te nemen met de klas, de regio, het land als normgroepen. Kinderen vergelijken hun rapportcijfers toch ook onderling. Ook beheersingsschalen zouden het informatiegehalte van schoolrapporten aanzienlijk kunnen verhogen. Met name echter, zou de meetnauwkeurigheid meer aandacht moeten krijgen. Een verandering van ruim voldoende naar zeer onvoldoende in een trimester op verschillende vakken moet bijvoorbeeld geweten worden aan een te lage betrouwbaarheid van de instrumenten, of er moet een andere reden zijn waarom de leerling niet zijn normale niveau heeft kunnen laten zien. Zo'n drastische verandering van resultaten mag echter niet zo maar worden geaccepteerd. Het rapporteren van de meetnauwkeurigheid, heeft niet alleen tot doel om ouders een betere inschatting te laten maken van de nauwkeurigheid van een resultaat. Belangrijker is dat een onderwijsinstelling meer geneigd zal zijn om de meetnauwkeurigheid van de rapportcijfers op een acceptabel niveau te houden of te krijgen.

Literatuur

Adema, J.J., & van der Linden, W.J. (1989). Algorithms for computerized test construction of parallel tests using classical item parameters. *Journal of Educational Statistics, 15,* 129-145.

Aitchison, J., & Silvey, S.D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics, 29,* 813-828.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Andersen, E.B. (1970). Asymptotic properties of conditional maximum likelihood estimation. *Journal of the Royal Statistical Society, Series B, 32,* 283-301.

Andersen, E.B. (1973a). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123-140.

Andersen, E.B. (1973b). *Conditional inference and models for measuring.* (Unpublished Ph.D. Thesis). Copenhagen: Mentalhygiejnisk Forlag.

Andersen, E.B. (1973c). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology, 26,* 31-44.

Andersen, E.B., & Madsen, M. (1977). Estimating the parameters of the latent population distribution. *Psychometrika, 42,* 357-374.

Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42,* 69-81.

Andersen, E.B. (1980). *Discrete statistical models with social science applications.* Amsterdam: North Holland.

Andersen, E.B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 46,* 443-459.

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Andrich, D. (1978b). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement, 38,* 665-680.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In: R.L. Thorndike (red.). *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Armstrong, R.D., Jones, D.H., & Wu, I. (1992). An automated test development of parallel tests from a seed test. *Psychometrika, 57,* 271-288.

Bartko, J.J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports, 19,* 3-11.

Bartko, J.J., & Carpenter, W.T. (1976). On the methods and theory of reliability. *The Journal of Nervous and Mental Disease, 163,* 307-317.

Bejar, I.I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7,* 303-310.

Bentler, P. M. (1985). *Theory and implementation of EQS: A structural equations program*. Los Angeles: BMDP Statistical Software.

Berger, J.O. (1980). *Statistical decision theory: Foundations, concepts and methods.* New York: Springer.

Berk, R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56,* 137-172.

Beuk, C.H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement, 21,* 147-152.

Bezembinder, Thom. G. G. (1970). *Van rangorde naar continuum*. Deventer: Van Loghum Slaterus.

Birnbaum, A. (1968). Some latent trait models. In: F.M. Lord, & M.R. Novick. *Statistical theories of mental test scores* (pp. 397-424). Reading: Addison-Wesley.

Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge: The MIT Press.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Bock, R.D. (1976). Basic issues in the measurement of change. In: D.N.M. de Gruijter, & L.J.Th. van der Kamp (red.). *Advances in psychological and educational measurement* (pp. 75-96). London: Wiley.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM-algorithm. *Psychometrika, 46,* 443-459.

Bock, R.D., Gibbons, R.D., & Muraki, E. (1988). Full-information factor analysis. *Psychological Measurement, 13,* 261-280.

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks. *Journal of Educational Statistics, 15,* 129-145.

Bol, E., & Verhelst, N.D. (1985). Inhoudelijke en statistische analyse van een leertoets. *Tijdschrift voor Onderwijsresearch, 10,* 49-68.

Bollen, K.A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bosch, L. van den, Gillijns, P., Krom, R., & Moelands, F. (1991). *Handleiding schaal vorderingen in spellingvaardigheid 1.* Arnhem: Cito.

Bradley, T.B. (1983). Remediation of cognitive deficits: A critical appraisal of the Feuerstein model. *Journal of Mental Deficiency Research, 27,* 79-92.

Braun, W.I., & Holland, P.W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In: P.W. Holland, & D.B. Rubin (red.). *Test equating* (pp. 9-49). New York: Academic Press.

Brennan, R.L. (1992). Elements of generalizability theory. Iowa City: ACT.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296-322.

Bügel, K. (1991). Sexeverschillen in onderwijsprestaties in Nederland: Een overzicht van de literatuur en enkele nieuwe gegevens. *Pedagogische Studiën, 68,* 350-370.

Bügel, K. (1993). Tekstbegrip moderne vreemde talen: De invloed van sekse en tekstonderwerp op de scores van centrale examens. *Tijdschrift voor Onderwijs-wetenschappen, 23,* 162-176.

Bügel, K., & Glas, C.A.W. (1991). Item specifieke verschillen in prestaties tussen jongens en meisjes bij tekstbegrip examens moderne vreemde talen. *Tijdschrift voor Onderwijsresearch, 16,* 337-351.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the mulititrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Campbell, D.T., & Stanley, J.C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Coombs, C.H. (1964). *A theory of data.* New York: Wiley.

Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement, 18,* 183-204; *19,* 331-332.

Cicchetti, D.V. (1972). A new measure of agreement between rank ordered variables. *In Proceedings of the 80th Annual Convention of the American Psychological Association 7,* 17-18.

Cicchetti, D.V. (1976). Assessing inter-rater reliability for rating scales: Resolving some basic issues. *British Journal of Psychiatry, 129,* 452-456.

Cochran, W. G. (1977). *Sampling techniques.* New York: Wiley.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psycholo-*
gical Measurement, 20, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provisions for scales disagreement of partial credit. *Psychological Bulletin, 70,* 213-220.

Cornfield, J., & J.W. Tukey (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics, 27,* 907-949.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Cronbach, L.J. (1971). Test validation. In: R.L. Thorndike (red.). *Educational Measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Cronbach, L.J., & Furby, L. (1970). How we should measure "change" - or should we? *Psychological Bulletin, 74,* 68-80.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Dirickx, Y.M.I., Baas, S.M., & Dorhout, B. (1987). *Operationele research*. Schoonhoven: Academic Service.

Divgi, D.R. (1981). *Two direct procedures for scaling and equating tests with item response theory*. Paper presented at the annual meeting of the National Council on Measurement in Education.

Dixon, W.J. (red.) (1992). *BMDP statistical software manual: Vol. 1 and 2.* Berkeley: University of California Press.

Dousma, T., & Horsten, A. (1989). *Tentamineren.* Groningen: Wolters-Noordhoff.

Drenth, P.J.D., & Sijtsma, K. (1990). *Testtheorie: Inleiding in de theorie van de psycholo- gische test en zijn toepassingen.* Houten: Bohn Stafleu Van Loghum.

Dunn, G. (1989). *Design and analysis of reliability studies: The statistical evaluation of measurement errors.* New York: Oxford University Press.

Ebel, R.L. (1967). The relation of item discrimination to test reliability. *Journal of Educational Measurement, 4,* 125-128.

Ebel, R.L. (1972). *Essentials of educational measurement.* Englewood Cliffs: Prentice-Hall.

Ebel, R.L. (1983). The practical validation of tests of ability. *Educational Measurement: Issues and Practice, 2,* 7-10.

Ebel, R.L., & Frisbie, D.A. (1986). *Essentials of educational measurement.* Englewood Cliffs: Prentice Hall.

Eggen, T.J.H.M. (1990). Innovative procedures in the calibration of measurement scales. In: W.H. Schreiber, & K. Ingenkamp (red.). *International developments in large scale*

assessment (pp.199-212). Windsor, Berkshire: NFER-NELSON.

Eggen, T.J.H.M., & Verhelst, N.D. (1992). *Item calibration in incomplete testing designs.* (Measurement and Research Department Reports 92-3). Arnhem: Cito.

Elliott, C.D., Murray, D.J., & Saunders, R. (1977). *Goodness of fit to the Rasch model as a criterion of test unidimensionality.* Manchester: University of Manchester.

Evers, A., Vliet-Mulder, J.C. van, & Laak, J. ter. (1992). *Documentatie van tests en testresearch in Nederland.* Amsterdam: Nederlands Instituut van Psychologen.

Fagot, R.F. (1991). Reliability of ratings for multiple judges: Intraclass correlation and metric scales. *Applied Psychological Measurement, 15,* 1-11.

Fagot, R.F. (1993). A generalized family of coefficients of relational agreement for numeri-
cal scales. *Psychometrika, 58,* 357-370.

Feldt, L.S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika, 30,* 357-370.

Feldt, L.S. (1993). The relationship between the distribution of item difficulties and test reliability. *Applied Measurement in Education 6,* 37-49.

Feldt, L.S., Steffen, M., & Gupta, N.C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement, 9,* 351-361.

Feldt, L.S, & Brennan, R.L. (1989). Reliability. In: R.L. Linn (red.). *Educational Measure-*
ment (3rd ed., pp. 105-146). Washington, DC: American Council on Education.

Ferguson, G.A., & Takane, Y. (1989). *Statistical analysis in psychology and education.* New York: McGraw-Hill.

Feuerstein, R. (1980). *Instrumental enrichment: An intervention program for cognitive modifiability.* Baltimore: University Park Press.

Fischer, G.H. (1972). *A step towards dynamic test-theory.* (Research Bulletin Nr. 10/72). Universität Wien: Psychologisches Institut.

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-373.

Fischer, G.H. (1974). *Einführung in die theorie psychologischer tests.* Bern: Huber.

Fischer, G.H. (1981). On the existence and uniqueness of maximum likelihood estimates in the Rasch model. *Psychometrika, 46,* 59-77.

Fischer, G.H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48,* 3-26.

Fischer, G.H. (in voorbereiding). Derivations of the Rasch model. In: G.H. Fischer, & I.W. Molenaar (red.). *Rasch models: Their foundations, recent developments and applica-*

tions.

Fischer, G.H., & Scheiblechner, H. (1970). Algorithmen und programme für das probabi- listische testmodell von Rasch. *Psychologische Beiträge, 12,* 23-51.

Flanagan, J.C. (1951). Units, scores and norms. In: E.F. Lindquist (red.). *Educational measurement* (pp. 695-763). Washington, DC: American Council on Education.

Fleiss, J.L. (1986). *The design and analysis of clinical experiments.* New York: Wiley.

Fleiss, J.L., Cohen, J., & Everitt, B.S. (1969) Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72,* 5, 323-327.

Fleiss, J.L., & Shrout, P.E. (1978). Approximate interval estimation for a certain intraclass correlation coefficient. *Psychometrika, 43,* 259-262.

Follman, D. (1988). Consistent estimation in the Rasch model based on nonparametric margins. *Psychometrika, 53,* 553-562.

Freeman, M.F., & Tukey, J.W. (1950). Transformations related to the angular and square root. *The Annals of Mathematical Statistics, 21,* 607-611.

Frisbie, D.A. (1988). Reliability of scores from teacher-made tests. *Educational Measure- ment: Issues and practice, 7,* 53-63.

Glas, C.A.W. (1981). *Het Raschmodel bij data in een onvolledig design*. (PSM-Progress reports, 81-1). Utrecht: Vakgroep PSM van de subfaculteit Psychologie.

Glas, C.A.W. (1989). *Contributions to estimating and testing Rasch models.* Arnhem: Cito.

Glas, C.A.W. (1992). A Rasch model with a multivariate distribution of ability. In: M. Wilson (red.). *Objective measurement: Theory into practice: Vol. 1* (pp. 236-258). Norwood: Ablex.

Glas, C.A.W., & Verhelst, N.D. (1989). Extensions of the partial credit model. *Psychometrika, 54,* 635-659.

Glas, C.A.W., & Verhelst, N.D. (in voorbereiding). Testing the Rasch model. In: G.H.Fischer, & I.W.Molenaar (red.). *Rasch models: Their foundations, recent developments and applications.*

Green, S.B., & Lissitz, R.W. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37,* 827-838.

Groot, A.D. de (1966). *Vijven en zessen.* Groningen: Wolters.

Groot, A.D. de, & Naerssen, R.F. (1973). *Studietoetsen, construeren, afnemen, analyseren: Deel I en II.* Den Haag: Mouton.

Gruijter, D.N.M. de (1985). Compromise models for establishing examination standards. *Journal of Educational Measurement, 22,* 263-269.

Guilford, J.P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education.* Tokyo: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Gustafsson, J.E. (1979). *PML: A computer program for conditional estimation and testing in the Rasch model for dichotomous items.* (Reports from the Institute of Education, nr. 63). Göteborg: University of Göteborg.

Guttman, L. A. (1950). The Basis of Scalogram Analysis. In: S.A. Stouffer, L.A. Gutmann, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, & J.A. Clausen (red.). *Measurement and prediction: Studies in social psychology in World War II: Vol. 4.* Princeton: Princeton University Press.

Guttman, L. A. (1954). A new approach to factor analysis: The radex. In: P.F. Lazersfeld (red.). *Mathematical thinking in the social sciences* (pp. 258-348). New York: Colombia University Press.

Haggard, E.A. (1958). *Intraclass correlation and the analysis of variance.* New York: The Dryden Press.

Hambleton, R.K., & Novick, M.R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10,* 159-170.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer Academic Publishers.

Hambleton, R.K., & Rogers, H.J. (1989). Detecting potentially biased test items: Compa- rison of IRT area and Mantel-Haenszel methods. *Applied Psychological Measurement, 2,* 313-334.

Harris, D.H., & Crouse, J.D. (1992). *A study of criteria used in equating.* Paper presented at the annual meeting of the National Council on Measurement in Education.

Heinen, T. (1993). *Discrete latent variable models.* Proefschrift, Katholieke Universiteit Brabant.

Henrysson, S. (1963). Correction of item-total correlations in item analysis. *Psychometrika, 28,* 211-218.

Hofstee, W.K.B. (1977). Cesuurprobleem opgelost. *Onderzoek van Onderwijs, 6/2,* 6-7.

Hofstee, W.K.B. (1981). *Psychologische uitspraken over personen.* Deventer: Van Loghum Slaterus.

Hofstee, W.K.B. (1983). The case for compromise in educational selection and grading. In Anderson, S.B., & Helmick, J.S. (red.). *On educational testing.* San Francisco: Jossey-Bass.

Hoijtink, H., & Boomsma, A. (1991). *Statistical inference with latent ability estimates.* (Prepublication Department of Statistics and Measurement Theory). Groningen: University of Groningen.

Hoijtink, H. (red.). (1993). *Kwantitatieve Methoden nr. 42.*

Holland, P.W., & Rubin, D.B. (1982). *Test equating.* New York: Academic Press.

Holland, P.W., & Thayer, D.T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In: H. Wainer, & H.I. Braun (red.). *Test validity* (pp.129-145). Hillsdale: Lawrence Erlbaum.

Hommel, G. (1983). Tests of the overall hypothesis for arbitrary dependence structures. *Biometrical Journal, 25,* 423-430.

Houston, W.M., Raymond, M.R., & Svec, J.C. (1991). Adjustments for rater effects in performance assessment. *Applied Psychological Measurement, 15,* 409-421.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). *Item response theory: Applications to psychological measurement.* Homewood: Dow-Jones Irwin.

Iker, H.P., & Perry, N.C.A. (1960). A further note concerning the reliability of the point-biserial correlation. *Educational and Psychological Measurement, 20,* 505-507.

Imbos, Tj. (1989). *Het gebruik van einddoel toetsen bij aanvang van de studie.* Proefschrift, Rijksuniversiteit Limburg.

Inspectierapport. (1992). *Examens op punten getoetst: Onderzoek naar de ontwikkeling van de normen bij de centrale examens in het voortgezet onderwijs.*

James, L.R., Demaree, R.G., & Wolf, G. (1984). Estimating within-group interrater relia- bility with and without response bias. *Journal of Applied Psychology, 69,* 85-98.

Jannarone, R.J. (1986). Conjunctive item response theory kernels. *Psychometrika, 51,* 357-373.

Jansen, G.G.H. (1979). *Het meten van veranderingen in de klassieke testtheorie.* (Bulletinreeks nr. 2). Arnhem: Cito.

Jarjoura, D. (1983). Best linear prediction of composite universe scores. *Psychometrika, 48,* 525-539.

Jazwinsky, A.H. (1970). *Stochastic processes and filtering theory.* New York: Academic Press.

Johnson, H.M. (1935). Some neglected principles in aptitude testing. *American Journal of Psychology, 47* 159-165.

Jonge, H. de (1963). *Inleiding tot de medische statistiek: Deel I.* Groningen: Wolters-Noordhoff.

Jöreskog, K.G. (1970). Estimation and testing of simplex models. *The British Journal of Mathematical and Statistical Psychology, 23,* 121-145.

Jöreskog, K.G., & Sörbom, D. (1989). *LISREL 7, user's reference guide.* Mooresville: Scientific Software.

Kamphuis, F.H., & Engelen, R.J.H. (in voorbereiding). Estimation and testing of structured latent ability covariance matrices in IRT models.

Kane, M.T. (1992). An argument-based approach to validation. *Psychological Bulletin, 112,* 527-535.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49,* 223-245.

Kelderman, H. (1988). *Loglinear multidimensional IRT model for polytomously scored items.* (Research Report 88-17). Enschede: Universiteit Twente.

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika, 54,* 681-697.

Kelderman, H., & Steen, R. (1988). *LOGIMO I: Loglinear item response theory modeling.* (Computer Program). Enschede: University of Twente, Department of Educational Technology.

Kelderman, H., & Macready, G.B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement, 27,* 307-327.

Kelley, T.L. (1947). *Fundamentals of statistics.* Cambridge: Harvard University Press.

Kendall, M., & Stuart, A. (1973). *The advanced theory of statistics: Vol. 2.* Londen: Griffin.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27,* 887-903.

Klauer, K.C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika, 56,* 213-228.

Kolen, M.J. (1988). Defining score scales in relation to measurement error. *Journal of Educational Measurement, 25,* 97-110.

Koppen, M.G.M. (1987). On finding the bidimension of a relation. *Journal of Mathematical Psychology, 3*1, 155-178.

Knol, D.L. (1986). *Een overzicht van meerdimensionale itemresponsmodellen*. (Rapport R-86-5). Enschede: Univeriteit Twente, Faculteit TO, vakgroep OMD.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement, 30,* 61-70.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Beverly Hills: Sage Publications.

Kuder, G.F., & Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151-160.

Lahey, M.A., Downey, R.G., & Saal, F.E. (1983). Intraclass correlations: There's more than meets the eye. *Psychological Bulletin, 93,* 586-595.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174.

Laros, J.A., & Tellegen, P.J. (1991). *Construction and validation of the SON-R 5½-17, the Snijders-Oomen non-verbal intelligence test.* Groningen: Wolters-Noordhoff.

Lazarsfeld, P.F. (1950). Logical and mathematical foundations of latent structure analysis. In: S.A. Stouffer. *Studies in social psychology in World War II, IV*. Princeton, NJ: Princeton University Press.

LBR (1988). *Psychologische tests en allochtonen.* Symposiumverslag 1987, LBR-Reeks nr. 6.

LBR (1990). *Toepasbaarheid van psychologische tests bij allochtonen.* Rapport van de testscreeningscommissie ingesteld door het LBR in overleg met het NIP, LBR-Reeks nr. 11.

Leeuw, J. de, & Verhelst, N.D. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics, 11*, 183-196.

Leeuwe, J.F.J. van (1990). *Probabilistic conjunctive models.* Proefschrift. Nijmegen: NICI.

Linden, W.J. van der (red.). (1982). Aspects of criterion-referenced measurement. *Evalua- tion in Education: An International Review Series, 5.*

Linden, W.J. van der (1983). *Van standaardtest naar itembank*. Universiteit Twente (oratie).

Linden, W.J. van der (1984). Some thoughts on the use of decision theory to set cutoff scores: Comment on De Gruijter and Hambleton. *Applied Psychological Measurement, 8,* 9-17.

Linden, W.J. van der (1985). Decision theory in educational research and testing. In: T. Husén, & T.N. Postlethwaite (red.). *International encyclopedia of education: Research and studies.* Oxford: Pergamon Press.

Linden, W.J. van der, & Boekkooi-Timminga, E. (1988). A zero-one programming approach to Gulliksen's matched random subtests method. *Applied Psychological Measurement, 12,* 201-209.

Linden, W.J. van der, & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika, 54,* 237-247.

Lindsay, B., Clifford, C.C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association, 86,* 96-107.

Linn, R.L. (red.). (1989). *Intelligence: Measurement, theory, and public policy*. Chicago: University of Illinois Press.

Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Livingston, S.A., & Zieky, M.J. (1982). *Passing scores: A manual for setting standards of performance on educational and performance tests.* Princeton, NJ: Educational Testing Service.

Lord, F.M. (1950). *Notes on comparable scales for test scores* (Research Bulletin 50-48). Princeton, NJ: Educational Testing Service.

Lord, F.M. (1952). The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika, 17*, 181-194.

Lord, F.M. (1953). On the statistical treatment of football numbers. *The American Psycholo- gist, 8,* 750-751.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale: Lawrence Erlbaum.

Lord, F.M. (1983a). Unbiased estimators of ability parameters, their variance and of their parallel-forms reliability. *Psychometrika, 48,* 233-245.

Lord, F.M. (1983b). *Estimating the imputed social cost of errors of measurement.* (Report RR-83-33-ONR). Princeton, NJ: Educational Testing Service.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading: Addison-Wesley.

Lord, F.M. & Wingerskey, M.S. (1983). Comparison of IRT true-score and equipercentile observed-score 'equatings'. *Applied Psychological Measurement, 8, 453-461.*

MacCann, R.G. (1990). Derivations of observed score equating methods that cater to populations differing in ability. *Journal of Educational Statistics, 15*, 146-170.

Maris, E. (1992). *Psychometric models for psychological processes and structures.* Proefschrift, Universiteit Leuven.

Martin-Löf, P. (1973). *Statistika Modeller: Anteckningar från seminarier Lasåret 1969-1970, utarbetade av Rolf Sunberg. Obetydligt ändrat nytryck, oktober 1973.* Stockholm: Institutet för Försäkringsmatematik och Matematisk Statistik vid Stockholms Universitet.

Martin-Löf, P. (1974). The notion of redundancy and its use as a quantitative measure if the discrepancy between a statistical hypothesis and a set of observational data. *Scandinavian Journal of Statistics, 1*, 3-18.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Masters, G.N., & Wright, B.D. (1984). The essential process in a family of measurement models. Psychometrika, 49, 529-544.

Maxwell, A.E., & Pilliner, A.E.G. (1968). Deriving coefficients of reliability and agreement. *The British Journal of Mathematical and Statistical Psychology, 21,* 105-116.

McKinley, R.L., & Reckase, M.D. (1983). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods and Instrumentation, 15,* 389-390.

Meerling (1981). *Methoden en technieken van psychologisch onderzoek: Deel 1.* Meppel: Boom.

Mellenbergh, G.J. (1977). The replicability of measures. *Psychological Bulletin, 84,* 378-384.

Mellenbergh, G.J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105-118.

Mellenbergh, G.J. (1983). Conditional item bias methods. In: S.H. Irvine, & W.J. Berry (red.). *Human assessment and cultural factors* (pp. 293-302). New York: Plenum Press.

Mellenbergh, G.J. (1985). Vraag-onzuiverheid: definitie, detectie en onderzoek. *Nederlands Tijdschrift voor Psychologie, 40,* 425-435.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In: H. Wainer, & H.I. Braun (red.). *Test validity* (pp.33-45). Hillsdale: Lawrence Erlbaum.

Messick, S. (1989). Validity. In: R.L. Linn (red.). *Educational Measurement* (3rd ed., pp. 13-103). Washington, DC: American Council on Education.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In: R.L. Linn (red.). *Educational Measurement* (3rd ed., pp. 335-366). Washington, DC: American Council on Education.

Mills, C.N., & Melican, G.J. (1987). *A preliminary investigation of three compromise methods for establishing cut-off scores.* (Report RR-87-14). Princeton, NJ: Educational Testing Service.

Mislevy, R.J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359-381.

Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177-195.

Mislevy, R.J., & Bock, R.D. (1986). *PC-BILOG: Maximum likelihood item analysis and test scoring with logistic models for binary items.* Mooresville: Scientific Software.

Mislevy, R.J., & Wu, P.K. (1988). *Inferring examinee ability when some item responses are missing.* (Research Report RR-88-48-ONR). Princeton, NJ: Educational Testing Service.

Mislevy, R.J., & Sheenan, K.M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54,* 661-680.

Moelands, A.H.J. (1988). *Entreetoets: Basisvaardigheden taal, rekenen en informatieverwerking (Verantwoording)*. Arnhem: Cito.

Mokken, R.J. (1971). *A theory and procedure of scale analysis.* Den Haag: Mouton.

Molenaar, I.W. (1981). *Programmabeschrijving van PML (versie 3.1) voor het Raschmodel.* (Heymans Bulletins Psychologische Instituten R.U.Groningen, nr. HB-81-538-RP). Groningen: Rijksuniversiteit Groningen.

Molenaar, I.W. (1983). *Item steps.* (Heymans Bulletins Psychologische Instituten R.U Groningen, nr. HB-83-630-EX). Groningen: Rijksuniversiteit Groningen.

Molenaar I.W., & Hoijtink, H (1990). The many null-distributions of person fit indices. *Psychometrika, 55,* 75-106.

Muskens, G.J. (1980). *Frames of meaning - are they measurable?* Proefschrift, Katholieke Universiteit Nijmegen.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika, 49,* 115-132.

Muthén, B. (1989). *LISCOMP: Analysis of linear structural equations with a comprehensive measurement model.* Mooresville: Scientific Software.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Nederlands Instituut van Psychologen. (1988). *Richtlijnen voor ontwikkeling en gebruik van psychologische tests en studietoetsen.* Amsterdam: Nederlands Instituut van Psychologen.

Novick, M.R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3,* 1-18.

Oud, J.H.L., & Mommers (1988). Longitudinale computerondersteunende ondersteuning van lees- en spellingsmoeilijkheden: Een toepassing van het Kalmanfilter in de onderwijs- praktijk. *Tijdschrift voor Onderwijsresearch, 13,* 31-50.

Pennings, A.H. (1988). The development of strategies in embedded figure tasks. *International Journal of Psychology, 23,* 65-78.

Pennings, A.H. (1991). *Individual differences in the development of the restructuring ability in children.* Proefschrift, Rijksuniversiteit Utrecht.

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, norming, and equating. In R.L. Linn (red.). *Eductional Measurement* (3rd ed., pp. 221-262). Washington, DC: American Council on Education.

Popping, R. (1983). *Overeenstemmingsmaten voor nominale data.* Proefschrift, Rijksuniversi- teit Groningen.

Popping, R. (1989). *AGREE: Computing agreement on nominal data, version 5.* (User's manual) Groningen: IEC ProGamma.

Popping, R. (1992). *Taxonomy on nominal scale agreement 1945 - 1990.* Groningen: IEC ProGamma.

Rao, C.R. (1948). Large sample tests of statistical hypothesis concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society, 44,* 50-57.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1961). On the general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* 321-333. Berkeley: University of California Press.

Rasch, G. (1977). *On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements.* Berkeley: University of California Press.

Read, T.R.C., & Cressie, N.A.C. (1988). *Goodness-of-fit statistics for discrete multivariate data.* New York: Springer.

Reckase, M.D., & Mckinley, R.L. (1985). Some latent trait theory in a multidimensional latent space. In: D.I. Weiss (red.). *Proceedings of the 1982 computerized adaptive testing conference* (pp. 151-177). Minneapolis: University of Minnesota.

Rigdon S.E., & Tsutakawa, R.K. (1983). Parameter estimation in latent trait models. *Psychometrika, 48,* 567-574.

Rigdon S.E., & Tsutakawa, R.K. (1986). Estimation for the Rasch model when both ability and difficulty parameters are random. *Journal of Educational Statistics, 12,* 76-86.

Roskam, E.E. (1982). Hypotheses non fingo, een methodologische gevalstudie over onderzoek van intelligentietests. *Nederlands Tijdschrift voor de Psychologie, 37,* 331-359.

Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63,* 581-592.

Rubin, D.B. (1980). Using empirical Bayes techniques in law school validity studies. *Journal of the American Statistical Association, 75,* 801-816.

Saal, F.E., Downey, R.G., & Lahey, M. (1980). Rating the ratings: Assessing the psycho- metric quality of rating data. *Psychological Bulletin, 88,* 413-428.

Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. *(Psycho- metric Monograph No. 17).* Psychometric Society.

Samejima, F. (1972). A general model for free response data. *(Psychometric Monograph No. 18)*. Psychometric Society.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38,* 203-219.

Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika, 42,* 193-198.

Sanders, P.F., Hendrix, A.C., & Luijten, A.J.M. (1984). De beoordeling van de samenvatting Nederlands. *Tijdschrift voor Taalbeheersing, 6,* 241-251.

Sanders, P.F., Theunissen, T.J.J.M., & Baas, S.M. (1989). Minimizing the number of observations: A generalization of the Spearman-Brown formula. *Psychometrika, 54,* 587-598.

Schouten, H.J.A. (1985). *Statistical measurement of interobserver agreement: Analysis of agreement and disagreement between observers.* Proefschrift, Rijksuniversiteit Utrecht.

Shavelson, R.J., & Webb, N.M. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology, 34,* 133-166.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer.* Newbury Park: Sage Publications.

Shepard, L.A. (1993). Evaluating test validity. In: L. Darling-Hammond (red.). *Review of research in education: Vol. 19* (pp.405-450). Washington, DC: American Educational Research Association.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliabi-
lity. *Psychological Bulletin, 86,* 420-428.

Shumway, R.H., & Stoffer, D.S. (1982). An approach to time series smoothing and forcasting using EM algorithm. *Journal of Time Series Analysis,* 3, 253-264.

Siegel, S., & Castellan, N.J.Jr. (1988). *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill.

Sijtsma, K., & Molenaar, I.W. (1987). Reliability of test scores in non-parametric item response theory. *Psychometrika, 52,* 79-97.

Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, series B, 13,* 238-241.

Sirotnik, K. (1970). An analysis of variance framework for matrix sampling. *Educational and Psychological Measurement, 30,* 891-908.

Sluijter, C., Boertien, H., de Klijn, W., & van Roosmalen, W. (1991). *De constructie van plaatsingstoetsen.* (Onderzoeksrapporten beginfase voortgezet onderwijs nr. 6). Arnhem: Cito.

Smith, P.L. (1978). Sampling errors of variance components in small sample multifacet generalizability studies. *Journal of Educational Statistics, 3,* 319-346.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271-295.

Staphorsius, G. (1992a). Welk boek is gemakkelijk, mijnheer ? *RAIN informatiebulletin, 2,* 7-10.

Staphorsius, G. (1992b). *Clib-toetsen.* Arnhem: Cito.

Staphorsius, G., & Krom, R.S.H. (1985a). *Leesbaarheidsindex voor het basisonderwijs.* (Bulletin nr. 36). Arnhem: Cito.

Staphorsius, G., & Krom, R.S.H. (1985b). Predictie van leesbaarheid. *Tijdschrift voor Taal- beheersing, 7,* 192-211.

Stine, W.W. (1989). Interobserver relational agreement. *Psychological Bulletin, 106,* 341 -347.

Suen, H.K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data.* Hillsdale: Lawrence Erlbaum.

Tatsuoka, K.K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 95-110.

Theunissen, T.J.J.M. (1985). Binary programming and test design. *Psychometrika, 50,* 411-420.

Theunissen, T.J.J.M. (1986). Some applications of optimization algorithms in test design and adaptive testing. *Applied Psychological Measurement, 10,* 381-389.

Theunissen, T.J.J.M. (1987). Text banking and test design. *Language Testing, 4,* 1-8.

Thissen, D. (1988). *MULTILOG: Multiple categorical item analysis and test scoring using item response theory*. Mooresville: Scientific Software.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567-577.

Thorndike, R.L. (1951). Reliability. In: E.F. Lindquist (red.). *Educational Measurement* (pp. 560-620). Washington, DC: American Council on Education.

Thorndike, R.L. (1982). *Applied psychometrics.* Boston: Houghton Mifflin Company.

Tinsley, H.E.A., & Weiss, D.J. (1975). Interrater reliability and agreement of subjective judgements. *Journal of Counseling Psychology, 23,* 358-376.

Uebersax, J.S. (1984). *Reliablility, validity and the kappa coefficient.* (Technical Report No. 12). Austin: University of Texas.

Uebersax, J.S. (1991). *Quantitative methods for the analysis of observer agreement: Towards a unifying model.* Santa Monica: RAND Corporation.

Uiterwijk, J.H. (1990). Verschillen tussen autochtonen en allochtonen bij de overgang van basisonderwijs naar voortgezet onderwijs. In: C.A.C. Klaassen, & P.L.M.

Jungbluth (red.). *Onderwijs researchdagen 1990, onderwijs en samenleving.* Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.

Uiterwijk, J.H., & Engelen, R.J.H. (1993). *Verantwoording eindtoets basisonderwijs 1990.* Arnhem: Cito.

Umesh, U.N., Peterson, R.A., & Sauber, M.H. (1989). Interjudge agreement and the maximum value of kappa. *Educational and Psychological Measurement, 49,* 835-850.

Vale, C.D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10,* 333-344.

Verhelst, N.D. (1989). Informatiewinst bij vertakt toetsen. In: W.J. van der Linden, & L.J.Th. van der Kamp (red.). *Meetmethoden en data-analyse* (pp. 89-96). Lisse: Swets en Zeitlinger.

Verhelst, N.D. (1993). *On the standard errors of parameter estimates in the Rasch model.* (Measurement and Research Department Reports 93-1). Arnhem: Cito.

Verhelst, N.D., Glas, C.A.W., & van der Sluis, A. (1984). Estimation problems in the Rasch model: The basic symmetric functions. *Computational Statistics Quarterly, 1,* 245-262.

Verhelst, N.D., & Eggen, T.J.H.M. (1989). *Psychometrische en statistische aspecten van peilingsonderzoek*. (PPON-rapport, nr. 4). Arnhem: Cito.

Verhelst, N.D., & Kamphuis, F.H. (1989). *Statistiek met* $\hat{\theta}$.(Bulletinreeks nr. 77). Arnhem: Cito.

Verhelst, N.D., Verstralen, H.H.F.M., & Eggen, T.J.H.M. (1991). *Finding starting values for the item parameters and suitable discrimination indices in the one-parameter logistic model.* (Measurement and Research Department Reports 91-10). Arnhem: Cito.

Verhelst, N.D., & Veldhuijzen, N.H. (1991). *A new algorithm for computing elementary symmetric functions and their first and second derivatives.* (Measurement and Research Department Reports 91-1). Arnhem: Cito.

Verhelst, N.D., & Verstralen, H.H.F.M. (1991). *The partial credit model with non-sequential solution strategies.* (Measurement and Research Department Reports 91-5). Arnhem: Cito.

Verhelst, N.D., & Glas, C.A.W. (in druk). A dynamic generalization of the Rasch model. *Psychometrika, 58.*

Verhelst, N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1993). *OPLM: One parameter logistic model.* Computer program and manual. Arnhem: Cito.

Verhelst, N.D., Verstralen.H.H.F.M., & Jansen, M.G.H. (1993) *A logistic model for time limit tests.* (Measurement and Research Department Reports 92-1). Arnhem: Cito.

Verschoor, A.J. (1991). *Optimal test design.* (Computer progamm and manual). Arnhem: Cito.

Verschoor. A.J., & Sanders, P.F. (1993). *Parallel test construction using the framework of classical test theory*. (Measurement and Research Department Reports 93-2). Arnhem: Cito.

Verstralen, H.H.F.M., & Verhelst, N.D. (1992). *The sample strategy of a test information function in computerized test design.* (Measurement and Research Department Reports 91-6). Arnhem: Cito.

Vogel, M., & Washburne, C. (1928). An objective method of determining grade placement of children's reading material. *Elementary School Journal, 28,* 373-381.

Wainer, H., & Mislevy, R.J. (1990). Item response theory, item calibration and proficiency estimation. In: H. Wainer (red.). *Computerized adaptive testing: A primer* (pp. 65-101). Hillsdale: Lawrence Erlbaum.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society, 54,* 426-482.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427-450.

Weiss, D.J. (red.). (1983). *New horizons in testing.* New York: Academic Press.

Wijnstra, J.M. (1988). *Balans van het rekenonderwijs in de basisschool.* Arnhem: Cito.

Wilson, D.T., Wood, R., & Gibbons, R.T. (1991). *TESTFACT.* Chicago: Scientific Software.

Wilson, M., & G.N. Masters, (1993). The partial credit model and null categories. *Psycho- metrika, 58,* 87-99.

Witkin, H.A. (1950). Individual differences in ease of perception of embedded figures. *Jour- nal of Personality, 19,* 1-15.

Witkin, H.A., & Goodenough, D.R. (1981). Cognitive styles: Essence and origins. *Psychological Issues* (Monograph 51). New York: International Universities Press.

Wollenberg, A.L. van den (1979). *The Rasch model and time limit tests.* Nijmegen: Studentenpers.

Wollenberg, A.L. van den (1982). Two new test statistics for the Rasch model. *Psychometrika, 47,* 123-140.

Wright, B.D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23-48.

Wright, B.D., & Mead, R.J. (1977). *BICAL: Calibrating items and scales with the Rasch model.* (Research Memorandum 23). Chicago: University of Chicago, Department of Education, Statistical Laboratory.

Wright, B.D., & Stone, M.H. (1979). *Best test design.* Chicago: MESA Press.

Yen, W.M. (1981). Using simultaneous results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245-262.

Yen, W.M. (1984). Tau-equivalence and equipercentile equating. *Psychometrika, 48,* 353- 369.

Zegers, F.E. (1989). Het meten van overeenstemming. *Nederlands Tijdschrift voor de Psychologie, 44,* 145-156.

Zegers, F.E. (1991). Coefficients for interrater agreement. *Applied Psychological Measurement, 15,* 321-333.

Zieky, M.J. (1987). *Methods of setting standards of performance on criterion referenced tests.* Paper presented at the 13th International Conference of the IAEA, Bangkok.

Zwinderman, A.H. (1991). *Studies of estimating and testing Rasch models.* (NICI Technical Report 91-02). Nijmegen: NICI.