DOCUMENT RESUME

ED 467 374 TM 034 300

AUTHOR	van der Linden, Wim J.
TITLE	Computerized Test Construction. Research Report.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
REPORT NO	RR-01-08
PUB DATE	2001-00-00
NOTE	15p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Computer Assisted Testing; *Test Construction; Test Format; Test Items
IDENTIFIERS	Constraints; *Linear Tests; Optimization; *Sequential Testing

ABSTRACT

This report contains a review of procedures for computerized assembly of linear, sequential, and adaptive tests. The common approach to these test assembly problems is to view them as instances of constrained combinatorial optimization. For each testing format, several potentially useful objective functions and types of constraints are discussed. (Contains 14 references.) (Author/SLD)

Computerized Test Construction

Research Report 01-08

Wim J. van der Linden

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

Computerized Test Construction

Wim J. van der Linden

To appear as entry 2.2.121 Psychometrics in: *N. J. Smelser & P. B. Baltes (Eds.) Computerized test construction.* Oxford: Elsevier Science Ltd.

Abstract

A review of procedures for computerized assembly of linear, sequential, and adaptive tests is given. The common approach to these test assembly problems is that they are viewed as instances of constrained combinatorial optimization. For each testing format several potentially useful objective functions and types of constraints are discussed.

Introduction

Like many other areas of psychology, the availability of cheap plentiful computational power has revolutionized the technology of educational and psychological testing. It is no longer necessary to restrict testing to the use of items with a paper-and-pencil format in a group-based session. We now have the possibility to build multimedia testing environments to which test takers respond by manipulating objects on a screen, working with application programs, or manipulating devices with built-in sensors. Moreover, such tests can be assembled from banks with items stored in computer memory and delivered immediately to examinees who walk in when they are ready to take the test test.

Computerized assembly of tests from an item bank is treated as an optimization problem with a solution that has to satisfy a potentially long list of statistical and nonstatistical specifications for the test. The general nature of this optimization problem is outlined, and applications to the problems of assembling tests with a linear, sequential, and adaptive format are reviewed.

Test Assembly as an Optimization Problem

The formal structure of a test assembly problem is known as a constrained combinatorial optimization problem. It is an optimization problem because the test should be assembled to be best in some sense. The problem is combinatorial because the test is a combination of items from the bank and optimization is over the space of admissible combinations. Finally, the problem is constrained because only those combinations of items that satisfy the list of test specifications are admissible.

The quintessential combinatorial optimization problem is the knapsack problem (Nemhauser & Wolsey, 1988). Suppose a knapsack has to be filled from a set of items indexed by i=1,..., I. Each item has utility u_i and weight w_i. The optimal combination of items is required to have maximum utility but should not exceed weight limit W. The combination is found defining decision variables.

$$x_i \equiv \begin{cases} 1 & \text{if item } i \text{ is selected,} \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

and solving the problem

$$\max \sum_{i=1}^{I} u_i x_i, \qquad \text{(maximum utility)}$$

subject to

$$\sum_{i=1}^{I} w_i x_i \leq W, \qquad \text{(weight limit)}$$

$$x_i = 0,\ 1. \qquad \text{(range of variables)}$$

for an optimal set of values for variables x_i.

Problems with this structure are known as 0-1 linear programming (LP) problems (e.g., see *Linear and Nonlinear Programming)*. Several test assembly problems can be formulated as a 0-1 LP problem; others need integer variables or a combination of integer and real variables. In a typical testassembly model, the objective function is used to maximize a statistical attribute of the test whereas the constraints serve to guarantee its content validity.

Objective Functions in Test Assembly Problems

Suppose the items in the bank are calibrated using an itme response theory IRTmodel, for example, the two-parameter logistic (2PL) model:

$$p_i(\theta) \equiv \Pr(U_i = 1 \mid \theta) \equiv \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}, \tag{2}$$

where $\theta \in (-\infty, \infty)$ is the ability of the examinee, and $b_i \in (-\infty, \infty)$ and $a_i \in [0, \infty)$ are parameters for the difficulty and discriminating power of item i, respectively (e.g., see *Factor Analysis and Latent Structure: IRT and Rasch Models)*.

A common objective function in IRT-based test assemblyis based on the test information function , which is Fisher's measure of information on the unknown ability θ in the response vector $U_1, ..., U_n$, where n is the number of items in the test. For the 2-PL model the test information function is given by

$$I(\theta) \equiv I_{U_1 \ldots U_n}(\theta) = \sum_{i=1}^{n} \frac{[p_i'(\theta)]^2}{p_i(\theta)[1 - p_i(\theta)]},$$

with $p_i'(\theta) \equiv \frac{\partial}{\partial \theta} p_i(\theta)$. Test information functions are additive in the contributions by the

individual items, which are denoted by $I_i(\theta)$.

The first step in IRT-based test assembly is to formulate a target for the test information function . The next step is to assemble the test to have its actual information function as closely as possible to the target. Examples of popular targets are a uniform function over an ability interval in diagnostic testing and a peaked function at a cutoff score θ_c in admission testing. An objective function to realize the former is presented in (4)-(5) below. The latter can be realized using objective function

$$\max \sum_{i=1}^{n} I_i(\theta_c) x_i, \tag{3}$$

under the condition of an appropriate set of constraints on the test. Other possible objective functions are maximization of classical test reliability and minimization of the length of the test. For a review of these and other examples, see van der Linden (1998).

Constraints in Test Assembly Problems

Formally, test specifications can be viewed as a series of upper and/or lower bounds on numbers of item attributes in the tests or on functions thereof. An important distinction is between constraints on (1) categortical item attributes, (2) quantitative item attribtues, and (3) logicaL relations between the items in the test. Categorical attributes are attributes such as item content, cognitive level, format, and use of graphics. Examples of quantitative attributes are statistical item parameters, expected response times, word counts, and readability indices. Logical (or Boolean) constraints deal with such issues in test assembly as items that can not figure in the same test because they have clues to each other's solution or items that are organized as sets around common stimuli.

Let V_j be a set of items in the bank with a common value for an attribute. The general shape of a constraint on a categorical item attribute in a test assembly model with 0-1 decision variables is

$$\sum_{i \in V_j} x_i \gtreqless n_j,$$

where n_j is a bound on the number of items from V_j. This type of constraint can also be formulated on intersections or unions of sets of items.

Constraints on quantitative attributes are typically on a function of their values for a set

of items. For example, if a typical test taker has response time t_i on item i and the total testing time available is T (both in seconds), a useful constraint on the test is:

$$\sum_{i=1}^{I} t_i x_i \leq T.$$

As an example of a logical constraint, suppose that W_s represents a set of items in the bank with a common stimulus s, and that n_s items have to be selected if and only if stimulus s has. This requirement leads to the following constraint

$$\sum_{i \in V_s} x_i = n_s z_s,$$

with z_s being an auxiliary 0-1 decision variable for the selection of stimulus s.

In a full-fledged test assembly problem, constraints may also be needed to deal, for instance, with stimulus attributes or with relations between different test forms if a set of forms is to be assembled simultaneously. For a review of these and other types of constraints, see van der Linden (1998; 2000a).

Linear-Test Assembly

Linear tests have a fixed number of items presented in a fixed order. For measurement over a larger ability interval, it is customary to choose a discrete set of target values for the information function, $T(\theta_k)$, k=1,...,K. In practice, because information functions are well-behaved continuous functions, target values at three to five equally-spaced θ values suffice. The need to match more than one target value simultaneously creates a multiobjective decision problem .

An effective way to deal with multiple target values is to apply a maximin criterion. This criterion leads to the following core of a test assembly model

$$\max y \quad \text{(common factor)} \tag{4}$$

subject to

$$\sum_{i=1}^{I} I_i(\theta_k) x_i \geq T(\theta_k) y \quad k = 1, ..., K, \quad \text{(minimum information at } \theta_k) \tag{5}$$

where y is a common factor in the right-hand side bounds in (5) that is maximized and coefficients $T(\theta_k)$ control the shape of the information function (van der Linden & Boekkooi-Timminga, 1989). Constraints to deal with the remaining content specifications should be added to this model. For a large-scale testing program in education, it is not unusual to have hundreds of those constraints.

Methods to solve test assembly models can be distinguished into algorithms that have been proven to lead to optimality and intuitively plausible heuristics, which typically select one item at a time. Well-known heuristics are those that pick the items with the largest impact on the test information function (Luecht 1998) or with the smallest weighted average deviation from all bounds in the model (Swanson & Stocking, 1993). Optimal solutions can be found using a branch-and-bound algorithm (Nemhauser & Wolsey, 1988), or, if the structure of the models boils down to a network-flow problem, a simplified version of the simplex algorithm (Armstrong et al. 1995). Several algorithms and heuristics are implemented in the test assembly package ConTest (Timminga, van der Linden & Schweizer, 1996).

Sometimes it is necessary to build a set of linear test forms, for instance, parallel forms to support different testing sessions or forms of different difficulties for use in an evaluation study with a pretest-posttest design. Sequential application of a model of linear-test assembly is bound to show a decreasing quality of the solutions. A simultaneous approach balancing the quality of the individual test forms can be obtained by replacing the decision variables in (1) by

$$x_{if} \equiv \begin{cases} 1 & \text{if item } i \text{ is selected for } f, \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

using these variables to model the test specifications for all forms f=1,...,F, and solving the model for all variables simultaneously. If no overlap between forms is allowed, logical constraints must be added to the model to prevent the variables from taking the value of one more than once. Efficient combinations of sequential and simultaneous approaches are presented in van der Linden and Adema (1998).

Sequential Test Assembly

Sequential test assembly is used in testing for selection or mastery with a cutoff score on the test that represents the level beyond which the test taker is accepted or considered to master the domain of knowledge tested, respectively. An obvious linear approach is to assemble a

test using the objective function in (3), but a more efficient procedure is to assemble the test sequentially, sampling one item from the bank at a time and stopping when the test taker is classified with enough precision.

If the items are dichotomous, the number-correct score of a test taker follows a bonomial distibutions

$$\Pr(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x},$$

with π being the success parameter and n the (random) number of items sampled. In a sequential probability ratio test (SPRT) for the decision to reject test takers with $\pi < \pi_0$ and accept those with $\pi \geq \pi_1$, with (π_0, π_1) being a small interval around cutoff score π_c, the decision rule is based on the likelihood ratio

$$\lambda_n \equiv \prod_{i=1}^{n} f(x_i \mid \pi_1)/f(x_i \mid \pi_0).$$

If $\lambda_n < A$ or $\lambda_n \geq B$ the decision is to reject or accept, respectively, whereas sampling of items is continued otherwise. The constants A and B are known to satisfy

$$\begin{aligned} A &\geq \beta(\pi_1)/\{1 - \alpha(\pi_0)\}, \\ B &\leq \{1 - \beta(\pi_0)\}/\alpha(\pi_0), \end{aligned} \tag{7}$$

with $\alpha(\pi_0)$ and $\beta(\pi_1)$ being the probabilities of a false positive and false negative decision for test takers with $\pi = \pi_0$ and $\pi = \pi_1$, respectively. For more on sequential methods, see *Sequential Statistical Methods.*

Alternative sequential approaches to test assembly follow an IRT-based SPRT (Reckase, 1983) or a Bayesian framework (Kingsbury & Weiss, 1983). Sequential Bayesian methods are further explained in *Bayesian Decision Theory*.

Adaptive Test Assembly

If the items in the pool are calibrated using a model as in (2), adaptive test assembly becomes possible. In adaptive test assembly, the items are selected to be optimal at the ability estimate of the test taker, updated by the computer after each new response. Adaptive testing leads to much shorter tests; savings are typically over 50% relative to a linear test with the same

precision.

To show the principle of adaptive testing, let $i = 1, ..., I$ denote the items in the pool and $k = 1, ..., K$ the items the test. It follows that i_k is the index of the item in the pool administered as the kth item in the test. The set $S_k \equiv \{i_1, ..., i_{k-1}\}$ contains the first $k-1$ items in the test. These items involve responses variables $\mathbf{U}_{k-1} \equiv (U_{i_1}, ..., U_{i_{k-1}})$. The update of the ability estimate after $k-1$ responses is denoted as $\widehat{\theta}_{\mathbf{u}_{k-1}}$. Item k in the test is selected to be optimal at $\widehat{\theta}_{\mathbf{u}_{k-1}}$ among the items in the set $R_k \equiv \{1, ..., I\} \backslash S_{k-1}$.

A popular criterion of optimality in adaptive testing is maximization of information at the current ability estimate, that is, selection of item i_k according to objective function

$$\max_{i \in R_k} I_i(\widehat{\theta}_{\mathbf{u}_{k-1}}). \tag{8}$$

Alternative objective functions are based on Kullback-Leibler information or on Bayesian criteria that use the posterior distribution of θ after $k-1$ items. These functions are reviewed in van der Linden and Pashley (2000).

Several procedures have been suggested to realize content constraints on adaptive tests. The four major approaches are: (1) partitioning the bank according to the main item attributes and spiraling item selection among the classes in the partition to realize a desired content distribution; (2) building deviations from content constraints into the objective function (Swanson & Stocking, 1993); (3) testing from a pool with small sets of items built according to content specifications; and (4) using a shadow test approach in which prior to each item a full linear tests is assembled that contains all previous items, meets all content constraints, and is optimal at the ability estimate, and from which the most informative item is selected for administration (van der Linden, 2000b).

Adaptive testing is currently one of the dominant modes of computerized testing. Several aspects of computerized adaptive testing not addressed in this entry are reviewed in Wainer (1990).

Bibliography

Armstrong, R. D., Jones, D. H., & Wang, Z. (1995). Network optimization in constrained standardized test construction. In K. D. Lawrence (Ed.), *Applications of management science: Network optimization applications* (Volume 8; pp. 189-212). Greenwich, CT: JAI Press.

Kingsbury, G. G., & Weiss D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 257-283). New York: Academic Press.

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement,* **22,** 224-236.

Nemhauser, G., & Wolsey, L. (1988). *Integer and combinatorial optimization*. New York: Wiley.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237-255). New York: Academic Press.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement,* **17,** 151-166.

Timminga, E., van der Linden, W. J., & Schweizer, D. A. (1996). *ConTEST 2.0: A decision support system for item banking and optimal test assembly* [Computer Program and Manual]. Groningen, The Netherlands: iec ProGAMMA.

van der Linden, W. J. (1998) Optimal assembly of educational and psychological tests, with a bibliography. *Applied Psychological Measurement,* **22,** 195-211. [With a bibliography]

van der Linden, W. J. (2000). Optimal assembling of tests with item sets. *Applied Psychological Measurement,* **24,** 225-240.

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.) *Computerized adaptive testing: Theory and practice* (pp. 27-52). Norwell MA: Kluwer Academic Publishers.

van der Linden, W. J., & Adema, J. J. (1998) Simultaneous assembly of multiple test forms. *Journal of Educational Measurement,* **35,** 185-198.

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

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.) *Computerized adaptive testing: Theory and practice* (pp. 1-25). Norwell, MA: Kluwer Academic Publishers.

Wainer, H. (Ed.) (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Erlbaum

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-01-08 W.J. van der Linden, *Computerized Test Construction*

RR-01-07 L.S. Sotaridona & R.R. Meijer, *Two New Statistics to Detect Answer Copying*

RR-01-06 L.S. Sotaridona & R.R. Meijer, *Statistical Properties of the K-index for Detecting Answer Copying*

RR-01-05 I. Hendrawan, C.A.W. Glas, & R.R. Meijer, *The Effect of Person Misfit on Classification Decisions*

RR-01-04 R. Ben-Yashar, S. Nitzan & H.J. Vos, *Optimal Cutoff Points in Single and Multiple Tests for Psychological and Educational Decision Making*

RR-01-03 R.R. Meijer, *Outlier Detection in High-Stakes Certification Testing*

RR-01-02 R.R. Meijer, *Diagnosing Item Score Patterns using IRT Based Person-Fit Statistics*

RR-01-01 W.J. van der Linden & H. Chang, *Implementing Content Constraints in Alpha-Stratified Adaptive testing Using a Shadow test Approach*

RR-00-11 B.P. Veldkamp & W.J. van der Linden, *Multidimensional Adaptive Testing with Constraints on Test Content*

RR-00-10 W.J. van der Linden, *A Test-Theoretic Approach to Observed-Score equating*

RR-00-09 W.J. van der Linden & E.M.L.A. van Krimpen-Stoop, *Using Response Times to Detect Aberrant Responses in Computerized Adaptive Testing*

RR-00-08 L. Chang & W.J. van der Linden & H.J. Vos, *A New Test-Centered Standard-Setting Method Based on Interdependent Evaluation of Item Alternatives*

RR-00-07 W.J. van der linden, *Optimal Stratification of Item Pools in a-Stratified Computerized Adaptive Testing*

RR-00-06 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using a Multidimensional IRT Model and Bayesian Sequential Decision Theory*

RR-00-05 B.P. Veldkamp, *Modifications of the Branch-and-Bound Algorithm for Application in Constrained Adaptive Testing*

RR-00-04 B.P. Veldkamp, *Constrained Multidimensional Test Assembly*

RR-00-03 J.P. Fox & C.A.W. Glas, *Bayesian Modeling of Measurement Error in Predictor Variables using Item Response Theory*

RR-00-02 J.P. Fox, *Stochastic EM for Estimating the Parameters of a Multilevel IRT Model*

RR-00-01 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items*

RR-99-08 W.J. van der Linden & J.E. Carlson, *Calculating Balanced Incomplete Block Designs for Educational Assessments*

RR-99-07 N.D. Verhelst & F. Kaftandjieva, *A Rational Method to Determine Cutoff Scores*

RR-99-06 G. van Engelenburg, *Statistical Analysis for the Solomon Four-Group Design*

RR-99-05 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *CUSUM-Based Person-Fit Statistics for Adaptive Testing*

RR-99-04 H.J. Vos, *A Minimax Procedure in the Context of Sequential Mastery Testing*

RR-99-03 B.P. Veldkamp & W.J. van der Linden, *Designing Item Pools for Computerized Adaptive Testing*

RR-99-02 W.J. van der Linden, *Adaptive Testing with Equated Number-Correct Scoring*

RR-99-01 R.R. Meijer & K. Sijtsma, *A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test*

RR-98-16 J.P. Fox & C.A.W. Glas, *Multi-level IRT with Measurement Error in the Predictor Variables*

RR-98-15 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory*

RR-98-14 A.A. Béguin & C.A.W. Glas, *MCMC Estimation of Multidimensional IRT Models*

RR-98-13 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Person Fit based on Statistical Process Control in an AdaptiveTesting Environment*

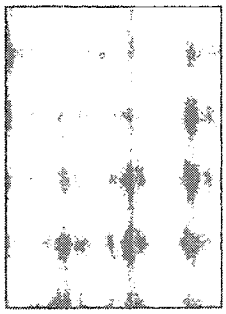
RR-98-12 W.J. van der Linden, *Optimal Assembly of Tests with Item Sets*

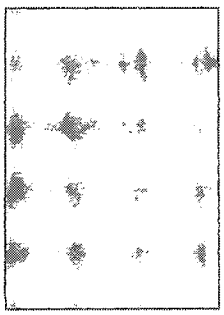
RR-98-11 W.J. van der Linden, B.P. Veldkamp & L.M. Reese, *An Integer Programming Approach to Item Pool Design*

RR-98-10 W.J. van der Linden, *A Discussion of Some Methodological Issues in International Assessments*

...

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.





faculty of

EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

X	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").