DOCUMENT RESUME

ED 424 267 TM 029 140

AUTHOR van der Linden, Wim J.; Glas, Cees A. W.
TITLE Capitalization on Item Calibration Error in Adaptive Testing. Research Report 98-07.
INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
PUB DATE 1998-00-00
NOTE 28p.
AVAILABLE FROM Faculty of Educational Science and Technology. University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Research (143)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Ability; *Adaptive Testing; *Computer Assisted Testing; *Error of Measurement; Estimation (Mathematics); Foreign Countries; *Test Items
IDENTIFIERS *Calibration; Optimization

ABSTRACT

In adaptive testing, item selection is sequentially optimized during the test. Since the optimization takes place over a pool of items calibrated with estimation error, capitalization on these errors is likely to occur. How serious the consequences of this phenomenon are depends not only on the distribution of the estimation errors in the pool or the ratio of the test length to the pool size, but also on the structure of the item selection criterion used. A simulation study demonstrated the existence of the phenomenon empirically. It also showed that its effect on the errors in the ability estimates interacts strongly with the distribution of the items in the pool. (Contains 1 table, 7 figures, and 15 references.) (Author)

TM

Capitalization on Item Calibration Error in Adaptive Testing

Research Report 98-07

Wim J. van der Linden
Cees A.W. Glas

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

Department of
Educational Measurement and Data Analysis

University of Twente

Capitalization on Item Calibration Error in Adaptive Testing

Wim J. van der Linden

Cees A.W. Glas

Abstract

In adaptive testing, item selection is sequentially optimized during the test. Since the optimization takes place over a pool of items calibrated with estimation error, capitalization on these errors is likely to occur. How serious the consequences of this phenomenon are depends not only on the distribution of the estimation errors in the pool or the ratio of the test length to the pool size, but also on the structure of the item selection criterion used. A simulation study demonstrated the existence of the phenomenon empirically. It also showed that its effect on the errors in the ability estimates interacts strongly with the distribution of the items in the pool.

Capitalization on Item Calibration Error in Adaptive Testing

The ideal underlying computerized adaptive testing (CAT) is to adapt the properties of the test items optimally to the ability of the examinee. The proper framework to realize this goal is item response theory (IRT). The unique feature of IRT models is that they have separate parameters to represent the properties of the items and the ability of the examinee. As a consequence, these models can be used to select items such that an optimal match is obtained between (a function of) the values of the item parameters and the value of ability parameter. Since the value of the ability parameter is not known, the test begins with an a priori estimate of the value of the ability parameter that is updated after each new item response. The values of the item parameters are estimated in advance; during the test these estimates are usually treated as if they are the true values of the parameters. A more complete description of adaptive testing is given in Wainer (1990).

One of the functions of the item parameters often used in adaptive testing, is Fisher's information function (Hambleton & Swaminathan, 1985, chap. 6; Lord, 1980, chap. 5). This function not only has the advantage of being monotonically related to the (asymptotic) standard error of the ML estimator of the ability parameter but is also additive in the items. Use of the function is generally accompanied by the application of the maximum-information criterion of item selection which selects the next item to have maximum information at the current estimate of the value of the ability parameter. If the value of the ability parameter is estimated in a Bayesian fashion, that is, by its posterior distribution given the responses on the previous items, other functions of the item parameter values are used. A well-known example of these functions is the expected reduction of the posterior variance. In Bayesian adaptive testing, the next item is selected to minimize this function. A more complete description of these item selection criteria is given below.

Application of an item selection criterion over a pool of items for a given examinee always involve optimization, that is, the choice of the next item for which the criterion has a maximum or minimum value. However, since the values for the item parameters are estimated, a process known as capitalization on chance may occur. The process operates on the fact that extreme values of a function of the item parameters can be the result of extreme true values of the parameters as well as large estimation errors. Consequently, if items are selected optimizing the value of this function, large estimation errors tend to be

overrepresented among the items selected. The result is an ability estimator with an accuracy likely to be worse than expected.

In test theory, the phenomenon of capitalization on chance has been well addressed for the problem of choosing a battery of variables with the largest predictive validity for job performance or academic success in a selection problem. The measure usually taken to counter its effects is to split the sample into a screening and a calibration sample. The variables are then selected in the screening sample but their regression parameters are re-estimated in the calibration sample (Lord & Novick, 1968, chap. 13). The effect of this cross validation is a shrinkage of the initial estimates of the regression parameters to more realistic sizes.

The problem of capitalization on chance was not addressed in the literature on test assembly until recently in papers by Hambleton and associates (Hambleton & Jones, 1994; Hambleton, Jones & Rogers, 1993). These authors show that if test forms are assembled to have maximum information over an ability interval and the values of the item parameters are estimated from a sample of N=400, the height of the information function may be overestimated by as much as 25-40%. Samples of this size are not uncommon in educational testing.

Several factors can be expected to have an impact on the process of capitalization on calibration error. The first is the distributions of the errors in the estimated parameter values in the item pool. Obviously, the larger the errors (or the smaller the calibration sample), the larger the effects of the capitalization on the values of the criterion. The second is the ratio of the number of items selected to the number in the pool. The smaller the ratio, the larger the likelihood of selecting items only from those with the larger estimation errors. The roles of both factors were confirmed in the studies by Hambleton et al.

The authors of this paper had no strong prior opinion as to the question whether the effects of capitalization on error in CAT would be more or less serious than in the assembly of test form with a fixed format. The size of the estimation errors and the selection ratio were certainly expected to remain important factors but the role of two new factors was unclear. The first new factor is the structure of the function of the item parameters used in the item selection criterion. As shown in an analysis below, item selection criteria are certainly sensitive to estimation error. On the other, it is known that for CATs of realistic length the ability estimator is quite robust with respect to the choice of the item selection criterion

(Chang & Ying, 1996; Veerkamp & Berger, 1997; van der Linden, 1998; van der Linden & Reese, 1998). The same may thus hold with respect to variation in the criterion values due to estimation error. The second factor deals with the question how the effects of early capitalization on errors in a CAT propagate later on in the test. In another context, it has been found that early bias in the ML ability estimator in a CAT tends to be neutralized by the maximum-information criterion later in the process (van der Linden, 1998). However, not much is known with respect to the effects of errors in the estimated values of the item parameter.

From a practical point of view, errors due to capitalization on chance in CAT are much more serious than in the assembly of forms for paper-and-pencil testing. All items are selected in real time and the estimates of their parameter values are used immediately to find the next "optimal" item. In adaptive testing, cross validation of item selection is impossible.

The remainder of this paper is organized as follows: First, the item selection criteria used in this study are introduced and analyzed for their liability to errors in item parameter estimation. Then, the design of the simulation study is discussed. The last section of the paper presents the results from the simulation study and draws some practical conclusions.

Item Selection Criteria and Estimation Error

As already indicated, the effects of capitalization on calibration error in CAT depend not only on the size of the calibration errors but also on the function defined on the item parameters optimized. One of the functions in use for CAT is Fisher's information function. For dichotomously scored items, the function has the following form:

$$I_i(\theta) \equiv \frac{P'(\theta)_i^2}{P_i(\theta) Q_i(\theta)}, \qquad (1)$$

$P_i(\theta)$ being the response function for item i, $P'(\theta)$ its first derivative with respect to θ, and $Q_i(\theta) \equiv 1 - P_i(\theta)$ (Lord, 1980, sect. 5.4). In CAT, the function is used to find the item in the pool that yields the largest value at $\theta = \hat{\theta}$, where $\hat{\theta}$ is the current estimate of the ability of the examinee.

For the two-parameter logistic (2-PL) model

$$P_i(\theta) \equiv \{1 + \exp[-a_i(\theta - b_i]\}^{-1}, \quad (2)$$

with a_i and b_i being the discrimination and difficulty parameter of item i, respectively, the information function is equal to

$$I_i(\theta) = a_i^2 P_i(\theta) Q_i(\theta). \quad (3)$$

Analytically, for a fixed value of a_i the function in (3) reaches a maximum for $\theta = b_i$, that is, for the θ value that gives $P_i(\theta)=.50$. At this point the maximum is equal to $.25\,a_i^2$. Thus, a CAT algorithm based on the maximum-information criterion will have a tendency to select items from the pool with values of b_i close to $\hat{\theta}$ and large values for a_i.

The critical factor in (3) is the size of the discrimination parameter a_i rather than the factor $P_i(\theta)Q_i(\theta)$. Because the parameter is squared in (1), the effect of estimation errors is enlarged. On the other hand, the factor $P_i(\theta)Q_i(\theta)$ in (1) is quite robust with respect to values for b_i in the neighborhood of the θ value of the examinee, even for larger values of a_i. If the value of $P_i(\theta)$ is in the range of [.40,.60], the maximal difference between the product $P_i(\theta)Q_i(\theta)$ and its maximum value is .01. If the range is enlarged to [.30,.70], the difference is still not larger than .04. Thus, a CAT algorithm based on the maximum-information criterion can be expected to capitalize on large errors in a_i but to be relatively robust with respect to errors in b_i.

If the three-parameter logistic (3-PL) model

$$P_i(\theta) \equiv c_i + (1 - c_i)\{1 + \exp[-a_i(\theta - b_i]\}^{-1} \quad (4)$$

with guessing parameter c_i is chosen, the structure of the information function is remains identical to the one in (3). The only change is the replacement of the factor $P_i(\theta)Q_i(\theta)$ in (3) by

$$[\{\frac{P_i(\theta)-c_i}{1-c_i}\}^2 / P_i(\theta)]Q_i(\theta). \quad (5)$$

(In all the expressions, $P_i(\theta)$ and $Q_i(\theta)$ still denote values obtained under the 2-PL model.) Note that (5) is generally smaller than the factor $P_i(\theta)Q_i(\theta)$ in (3) but that equality is obtained if $c_i \rightarrow 0$. It can therefore be concluded that (5) has a smaller effect on the value of the information function than the factor $P_i(\theta)Q_i(\theta)$ in (3) and that the value of the discrimination parameter a_i remains the critical factor.

A Bayesian criterion for item selection in CAT is the criterion of minimum expected posterior variance. An approximate version of the criterion for use in CAT was introduced by Owen (1975). In the criterion it is assumed that the ability estimation starts with a prior distribution for θ which is updated after each item response using Bayes theorem. The next item is selected to have a predicted posterior distribution with minimum variance among all items. For a more detailed description of this criterion, see van der Linden (1998).

To present the criterion more formally, let $(u_1,...,u_{k-1})$ be the responses obtained on the first k-1 items in the CAT. If item i is selected, the expected posterior variance is

$$\sum_{j=1}^{2} P_i(U_i = j \mid u_1,...,u_{k-1}) \mathrm{Var}(\theta \mid u_1,...,u_{k-1}, U_i = j), \quad (6)$$

where $\mathrm{Var}(\theta|u_1,\ldots,u_{k-1})$ is the posterior variance of θ and

$$P_i(U_i = j \mid u_1,...,u_{k-1}) = \int P_i(U_i = j \mid \theta) g(\theta \mid u_1,...,u_{k-1}) d\theta \quad (7)$$

is the posterior predictive probability of response u_i on item i given the responses $u_1,...,u_{k-1}$ to the previous items. The next item is selected to have a minimal value for (6) among the items in the pool.

A variation of the criterion in (6) is the maximum expected posterior-weighted information criterion. The criterion also predicts the probabilities of responses $U_i=1$ and $U_i=0$ for each item i in the pool but uses these probabilities to calculate the expected posterior-weighted information:

$$\sum_{j=1}^{2} P_i(U_i = j \mid u_1,\ldots,u_{k-1}) \int I_{U_1,\ldots,U_{k-1},U_i}(\theta)\, g(\theta \mid u_1,\ldots,u_{k-1},U_i = j)\, d\theta \qquad (8)$$

where $g(\theta \mid .)$ is the posterior density of θ after k-1 items have been selected.

The critical difference between the maximum-information criterion in (3) and the maximum expected posterior-weighted information in (8) is the role of the posterior distribution of θ. In (3) the information function is evaluated close to the center of the posterior distribution of θ whereas in (8) the information function is integrated over the full posterior. It is expected that the two criteria show different behavior at the beginning of the test where (3) has a preference for information functions that peak at the center of the posterior but that differences disappear as the posterior itself becomes peaked later in the test.

Simulation Study

To further explore the role of capitalization on error in CAT a simulation study was conducted. The effects of the following factors were studied:

1. The size of the calibration sample (N=500, 1500, 2500, ∞);
2. The length of the test (n=10, 20, 40);
3. The size of the item pool (k=40, 80, 400, 1200);
4. The nature of the item selection criterion (maximum information; minimum expected posterior variance; maximum expected posterior-weighted information).

In all cases, ability was estimated using the expected a posteriori (EAP) estimator with a N(0,1) prior. For the maximum information criterion, ability was also estimated using the weighted maximum likelihood (WML) estimator derived in Warm (1989). The latter is attractive because of its negligible bias.

Method

A calibrated pool of items was simulated as follows. A data matrix with 1,000 examinees by 100 items was available from a Dutch national school leaving exam of English as a foreign language. The items were calibrated under the 2-PL model in (2) using the method of marginal maximum likelihood estimation with a N(0,1) distribution for the ability parameter in the model. In addition, the information matrix for the item parameters was estimated from the data. To simulate calibration samples of different sizes, the required numbers of examinees were drawn from the data matrix at random and with replacement. As the information matrix is additive in the examinees, it could easily be adapted to the various samples of examinees.

The true parameter values were equated to the values estimated from the data matrix; their distributions are displayed in Table 1. The distribution of the values for the item

[Table 1 about here]

difficulty parameter had a mean of .970, for an ability distribution with mean and standard deviation normed at .00 and 1.0, respectively. Thus, the item pool was relatively difficult for the examinees.

Item calibration errors were drawn from normal distributions using the information matrix to calculate their variances. To simulate calibrated pools with larger numbers of items, the set of true values of the item parameters were duplicated and independent draws for the error distributions were made.

Each of the item pools in this study had 1,200 simulated items. In one part of the study the item pool consisted of a mixture of items calibrated using different sample sizes; one third of the items was simulated to be calibrated on a sample of 500 examinees, one third on a sample of 1500 examinees, and one third on a sample of 2500 items. These sections of the pool thus had identical distributions of their true parameter values but differed in the size of their calibration errors. The presence of capitalization on calibration errors was examined by counting the numbers of times items from the three sections were used in the adaptive tests.

In the second part of the study, the item pools were homogeneous with respect to the size of the calibration sample. These pools were used to assess the effect of item calibration error on the final ability estimates in the adaptive procedures.

The adaptive testing procedure was replicated 100 times for θ = 2.0, -1.0, 0.0, 1.0, 2.0. to obtain stable estimates of the counts and mean absolute errors.

Results

Figure 1-3 display the counts of the numbers selected in the adaptive procedure from

[Figures 1-3 about here]

the sections in the item pool calibrated on samples of N=500, 1500, and 2500 examinees as a function of θ. In each panel, the curves always sum to 100n (that is, the number of replications times test length). The dominant impression from the figures is that the smaller the size of the calibration sample, the larger the number of items selected. A surprisingly strong effect was present for the maximum posterior-weighted expected information criterion in combination with tests of n=10 items. However, an exception was obtained for maximum-information criterion and WML ability estimation for n=10; an explanation for this anomaly could not be found. The effect showed a tendency to decline for tests with 40 items but was still present at this test length, in particular at the high end of the ability scale.

Though not reported in these figures, the values of the discrimination parameters, a_i, for the items selected were broken down into sets of items with $a_i<.7$ and $a_i\geq.7$. This distinction roughly corresponds to items with discrimination values below and above the average value for the items in the pool (see Table 1). However, for nearly all θ values and item-selection criteria, items with values for a_i in the lower category were never chosen. The only exception were a few cases with low θ values for the maximum-information criterion. These results remind us of a experience well known in the practice of adaptive testing: Due to the presence of low discriminating items, the effective size of the item pool is generally much smaller than the number of items present in the pool.

In Figures 4-6, each curve represents the mean absolute error in the ability estimates

[Figures 4-6 about here]

estimates as a function of θ for the item pools calibrated on samples with sizes of N=500, 1500 and 2500 examinees, the mixture of these samples sizes used above, and the true parameter values ($N=\infty$). For n=10, the U-shaped curves typical of a short adaptive test with a prior for the ability parameter located at at $\theta=0$ were obtained. For n=20 and 40 the curves became flatter, where the curves for the Bayesian item-selection criteria tended to be lower and flatter than those for the maximum-information criterion. Though the four criteria showed different degrees of capitalization on calibration error in Figures 1-3, the curves in Figures 3-6 were more homogeneous. The most conspicuous property of the latter, however, was much larger variation in the mean absolute error between the different calibration samples at the

higher part of the ability scale. Also, at this part of the scale, the size of the mean absolute errors was inversely related to the size of the calibration sample. This result is due to the larger supply difficult items in the pool (see Table 1). As a consequence, the item-selection ratio at this part of the scale is considerably smaller, and the tendency to capitalize on item parameter estimation errors much stronger.

The effect of the item-selection ratio is also shown in Figure 7. For an item pool with

[Figure 7 about here]

with size k=40, that is, a large item-selection ratio, capitalization on calibration errors was not expected to occur. For this pool size, the curves in Figure 7 showed a mean absolute error in the ability estimates that was high at the lower end of the scale but smaller at the higher end. This shape reflects the fact that the majority of the items were relatively difficult. When the size of the item pool increased, and thus the item-selection ratio decreased, the curves for the smaller calibration samples deterioriated at the higher end of the scale whereas the curve for the true parameter values further improved. This increase in differences between the curves for the various samples sizes at the high end of the scale across the four panels in this figure is therefore expected to be due to capitalization on calibration error.

Conclusion

The general picture emerging from this example is that capitalization on calibration does occurs in adaptive testing and that its most important determinant is the item-selection ratio. Item pools and test lengths of various sizes were used to study the effects of this ratio on the ability estimates. However, due to the fact that the item pools were generated from an empirical data set, difficult items were overrepresented, the result being an actual item-selection ratio smaller than expected at the higher end of the ability scale.

This unexpected result showed that the composition of the item pool is an important factor interacting with the effect of capitalization on errors in the item parameters on the errors in the ability estimates. Large numbers of items for certain θ values - intuitively an attractive feature of an item pool - is <u>not</u> a desideratum if the calibration sample is small.

References

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. Applied Psychological Measurement, 20, 213-229.

Hambleton, R.K., & Jones, R.W. (1994). Item parameter estimation errors and their influence on test information functions. Applied Measurement in Education, 7, 171-186.

Hambleton, R.K., Jones, R.W., & Rogers, H.J. (1993). Influence of item parameter estimation errors in test development. Journal of Educational Measurement, 30, 143-155.

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer-Nijhoff.

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

Lord, F.M. (1983). Small N justifies Rasch model. In D.J. Weiss (Ed.), New horizons in testing: Latent trait theory and computerized adaptive testing. New York: Academic Press.

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

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive testing. Journal of the American Statistical Association, 70, 351-356.

Tang, K.L., Way, W.D., & Carey, P.A. The effect of small calibration sample sizes on TEOFL IRT-based equating (TOEFL Technical Report TR-7). Princeton, NJ: Educational Testing Service.

Tsutakawa, R.K., & Johnson, J.C. (1990). The effect of uncertainty on item parameter estimation on ability estimates. Psychometrika, 55, 371-390.

van der Linden, W.J. (1998). Bayesian item selection criteria for adaptive testing. Psychometrika, 63.

van der Linden, W.J., & Reese, L.M. (1998). A model for optimal constrained adaptive testing. Applied Psychological Measurement, 22.

Veerkamp, W.J.J., & Berger, M.P.F. (1997). Some new item selection criteria for adaptive testing. Journal of Educational and Behavioral Statistics, 22, 203-226.

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

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

Authors' Note

This study received funding from the Law School Admission Council (LSAC). The opinions and conclusions contained in this paper are those of the authors and do not necessarily reflect the position or policy of LSAC.

Table 1

Distribution of true parameter values in simulated item pool

	Mean	Minimum	Maximum	Standard Deviation
a_i	.777	.222	1.841	.288
b_i	.970	-1.262	3.590	.885

Figure Captions

Figure 1. Numbers of items in the adaptive tests selected from the sections in the pool calibrated on N=500, 1500, and 2500 examinees for the various item-selection criteria (n=10).

Figure 2. Numbers of items in the adaptive tests selected from the sections in the pool calibrated on N=500, 1500, and 2500 examinees for the various item-selection criteria (n=20).

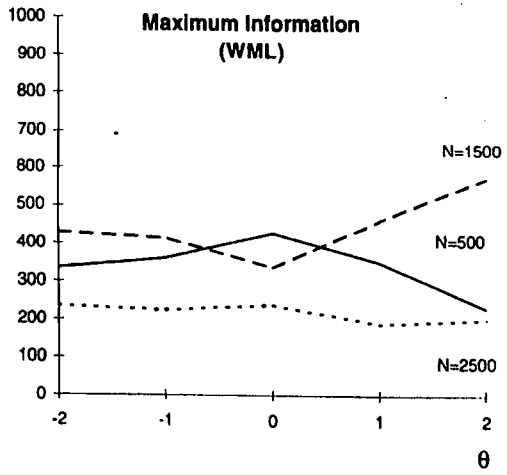
Figure 3. Numbers of items in the adaptive tests selected from the sections in the pool calibrated on N=500, 1500, and 2500 examinees for the various item-selection criteria (n=40).

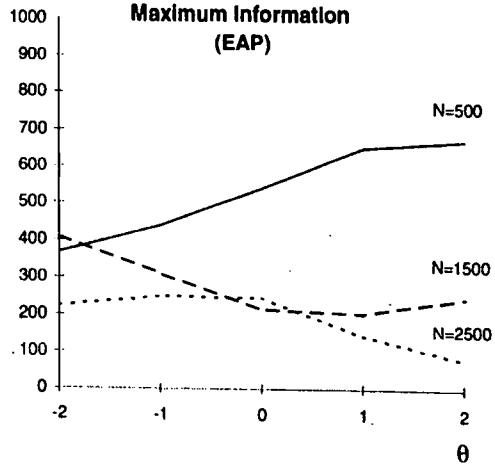
Figure 4. Mean absolute error in the ability estimates for item pools calibrated on N=500 (solid curve), 1500 (dashed curve), 2500 (dotted curve), a mixture of these sample sizes (bold curve), and N=∞ examinees (grey curve) for the four item-selection criteria (n=10).

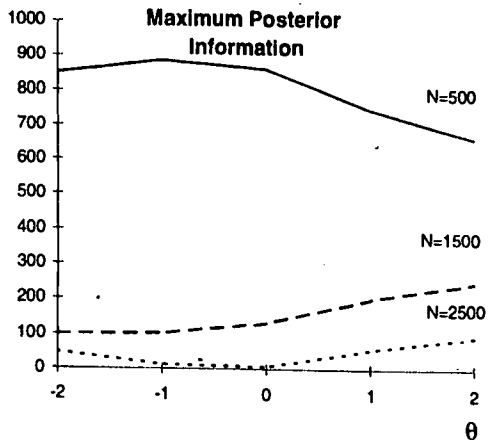
Figure 5. Mean absolute error in the ability estimates for item pools calibrated on N=500 (solid curve), 1500 (dashed curve), 2500 (dotted curve), a mixture of these sample sizes (bold curve), and N=∞ examinees (grey curve) for the four item-selection criteria (n=20).

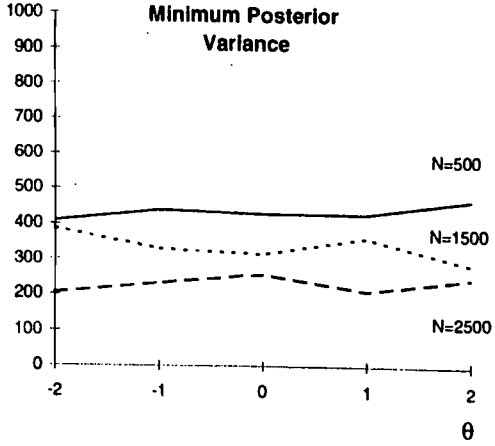
Figure 6. Mean absolute error in the ability estimates for item pools calibrated on N=500 (solid curve), 1500 (dashed curve), 2500 (dotted curve), a mixture of these sample sizes (bold curve), and N=∞ examinees (grey curve) for the four item-selection criteria (n=40).

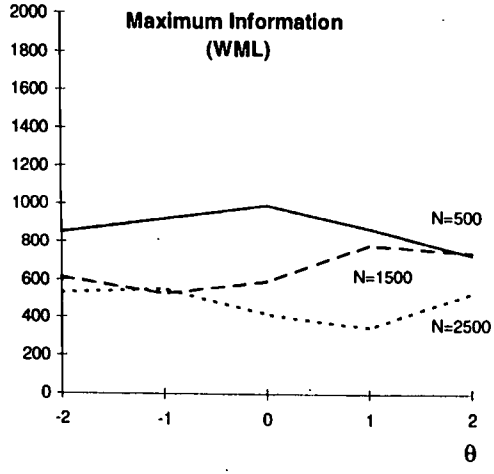
Figure 7. Mean absolute error in the ability estimates for item pools calibrated on N=500 (solid curve), 1500 (dashed curve), 2500 (dotted curve), a mixture of these sample sizes (bold curve), and N=∞ examinees (grey curve) for pool sizes of k=40, 80, 400, and 1200 items (maximum-information criterion with weighted maximum likelihood estimation of ability; n=20).

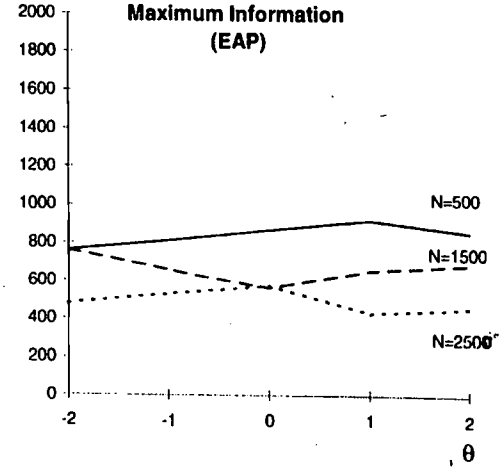


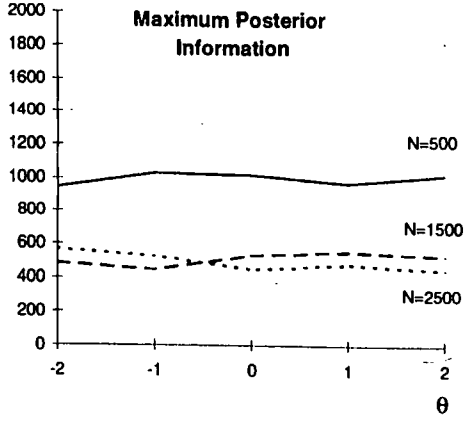


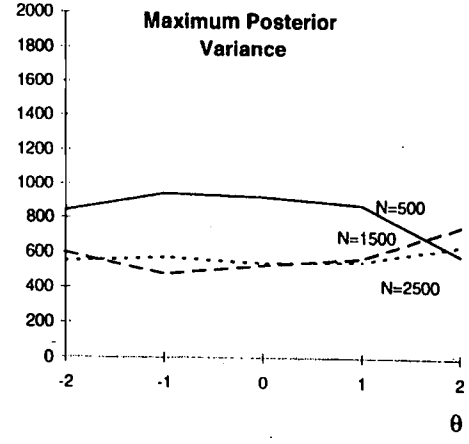


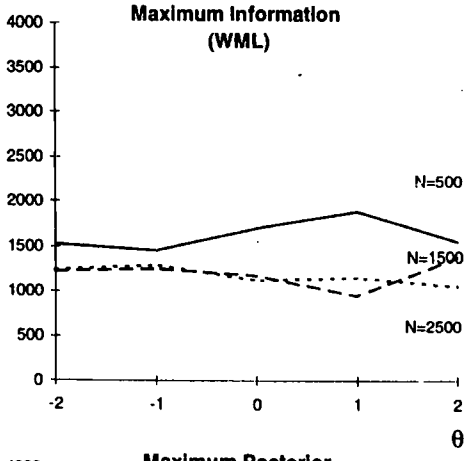


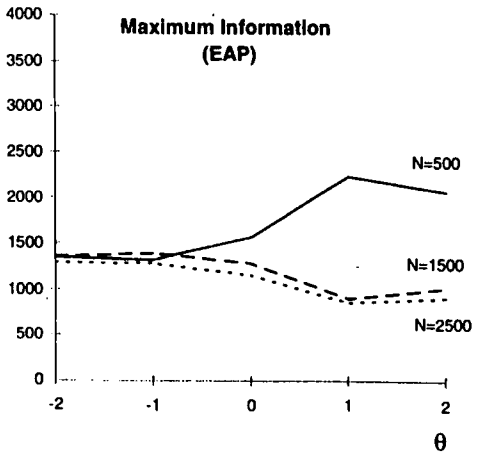


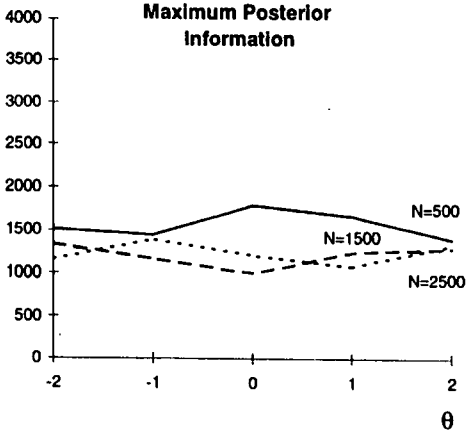


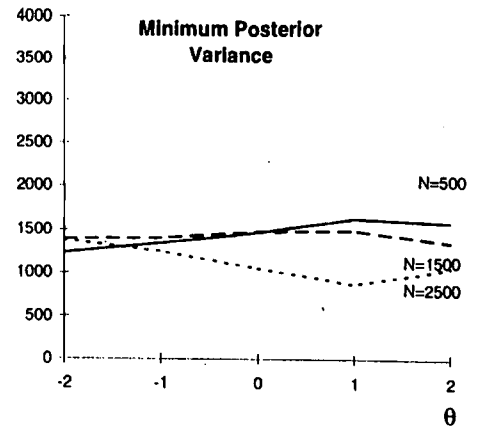




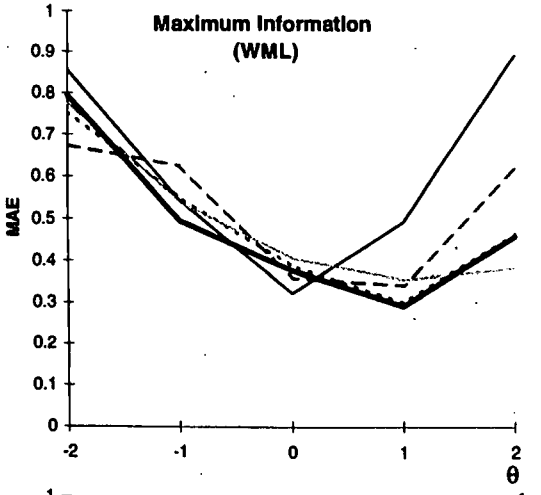


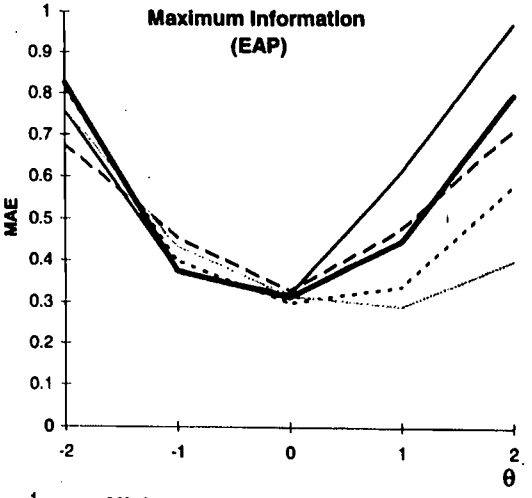


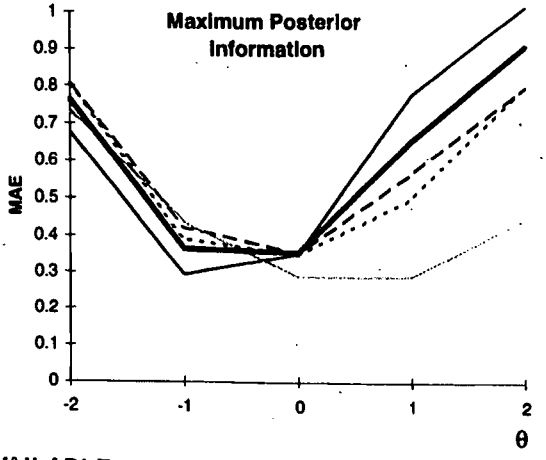


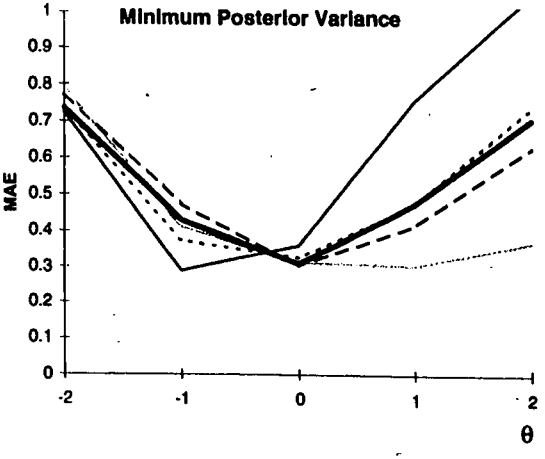


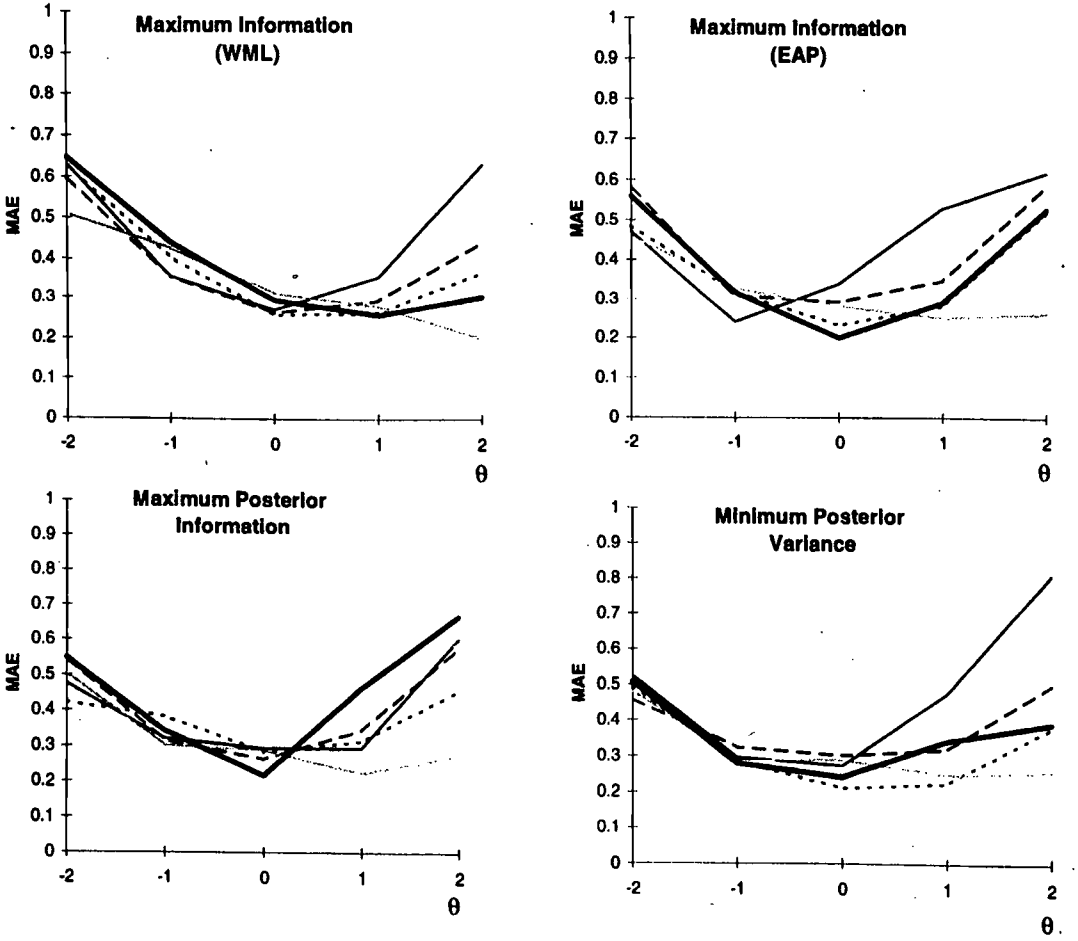
n=10



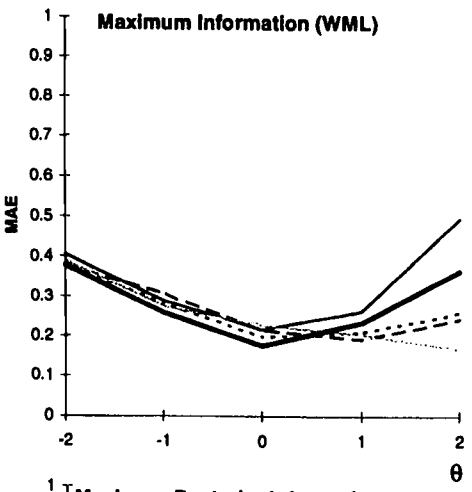


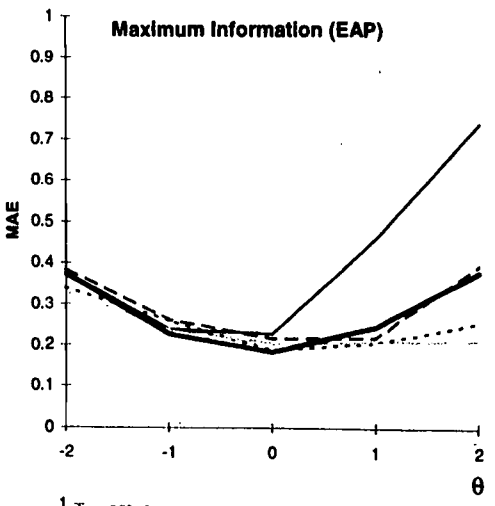


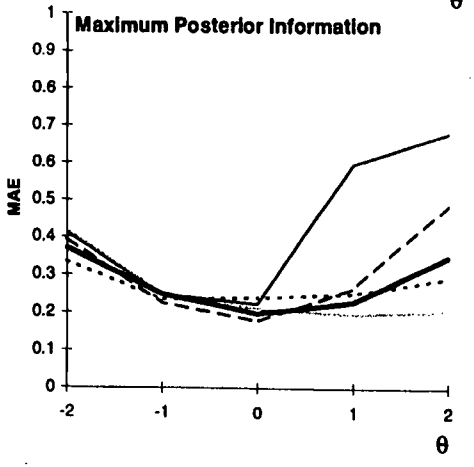


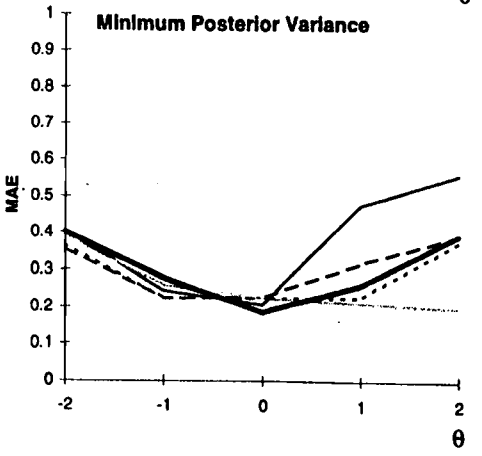


n=40

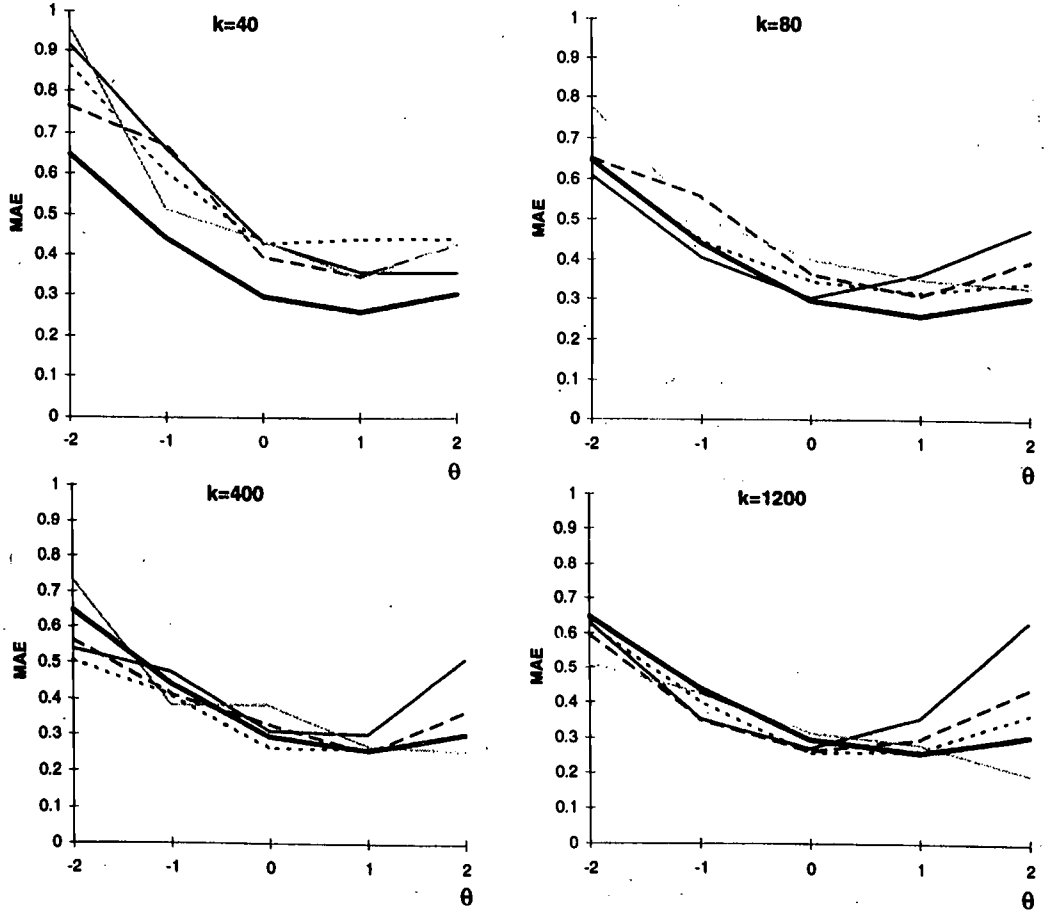








n=20



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

RR-98-07 W.J. van der Linden & C.A.W. Glas, *Capitalization on Item Calibration Error in Adaptive Testing*

RR-98-06 W.J. van der Linden, D.J. Scrams & D.L.Schnipke, *Using Response-Time Constraints in Item Selection to Control for Differential Speededness in Computerized Adaptive Testing*

RR-98-05 W.J. van der Linden, *Optimal Assembly of Educational and Psychological Tests, with a Bibliography*

RR-98-04 C.A.W. Glas, *Modification Indices for the 2-PL and the Nominal Response Model*

RR-98-03 C.A.W. Glas, *Quality Control of On-line Calibration in Computerized Assessment*

RR-98-02 R.R. Meijer & E.M.L.A. van Krimpen-Stoop, *Simulating the Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests*

RR-98-01 C.A.W. Glas, R.R. Meijer, E.M.L.A. van Krimpen-Stoop, *Statistical Tests for Person Misfit in Computerized Adaptive Testing*

RR-97-07 H.J. Vos, *A Minimax Sequential Procedure in the Context of Computerized Adaptive Mastery Testing*

RR-97-06 H.J. Vos, *Applications of Bayesian Decision Theory to Sequential Mastery Testing*

RR-97-05 W.J. van der Linden & Richard M. Luecht, *Observed-Score Equating as a Test Assembly Problem*

RR-97-04 W.J. van der Linden & J.J. Adema, *Simultaneous Assembly of Multiple Test Forms*

RR-97-03 W.J. van der Linden, *Multidimensional Adaptive Testing with a Minimum Error-Variance Criterion*

RR-97-02 W.J. van der Linden, *A Procedure for Empirical Initialization of Adaptive Testing Algorithms*

RR-97-01 W.J. van der Linden & Lynda M. Reese, *A Model for Optimal Constrained Adaptive Testing*

RR-96-04 C.A.W. Glas & A.A. Béguin, *Appropriateness of IRT Observed Score Equating*

RR-96-03 C.A.W. Glas, *Testing the Generalized Partial Credit Model*

RR-96-02 C.A.W. Glas, *Detection of Differential Item Functioning using Lagrange Multiplier Tests*

RR-96-01 W.J. van der Linden, *Bayesian Item Selection Criteria for Adaptive Testing*

RR-95-03 W.J. van der Linden, *Assembling Tests for the Measurement of Multiple Abilities*

RR-95-02 W.J. van der Linden, *Stochastic Order in Dichotomous Item Response Models for Fixed Tests, Adaptive Tests, or Multiple Abilities*

RR-95-01 W.J. van der Linden, *Some decision theory for course placement*

RR-94-17 H.J. Vos, *A compensatory model for simultaneously setting cutting scores for selection-placement-mastery decisions*

RR-94-16 H.J. Vos, *Applications of Bayesian decision theory to intelligent tutoring systems*

RR-94-15 H.J. Vos, *An intelligent tutoring system for classifying students into Instructional treatments with mastery scores*

RR-94-13 W.J.J. Veerkamp & M.P.F. Berger, *A simple and fast item selection procedure for adaptive testing*

RR-94-12 R.R. Meijer, *Nonparametric and group-based person-fit statistics: A validity study and an empirical example*

RR-94-10 W.J. van der Linden & M.A. Zwarts, *Robustness of judgments in evaluation research*

RR-94-9 L.M.W. Akkermans, *Monte Carlo estimation of the conditional Rasch model*

RR-94-8 R.R. Meijer & K. Sijtsma, *Detection of aberrant item score patterns: A review of recent developments*

RR-94-7 W.J. van der Linden & R.M. Luecht, *An optimization model for test assembly to match observed-score.distributions*

RR-94-6 W.J.J. Veerkamp & M.P.F. Berger, *Some new item selection criteria for adaptive testing*

RR-94-5 R.R. Meijer, K. Sijtsma & I.W. Molenaar, *Reliability estimation for single dichotomous items*

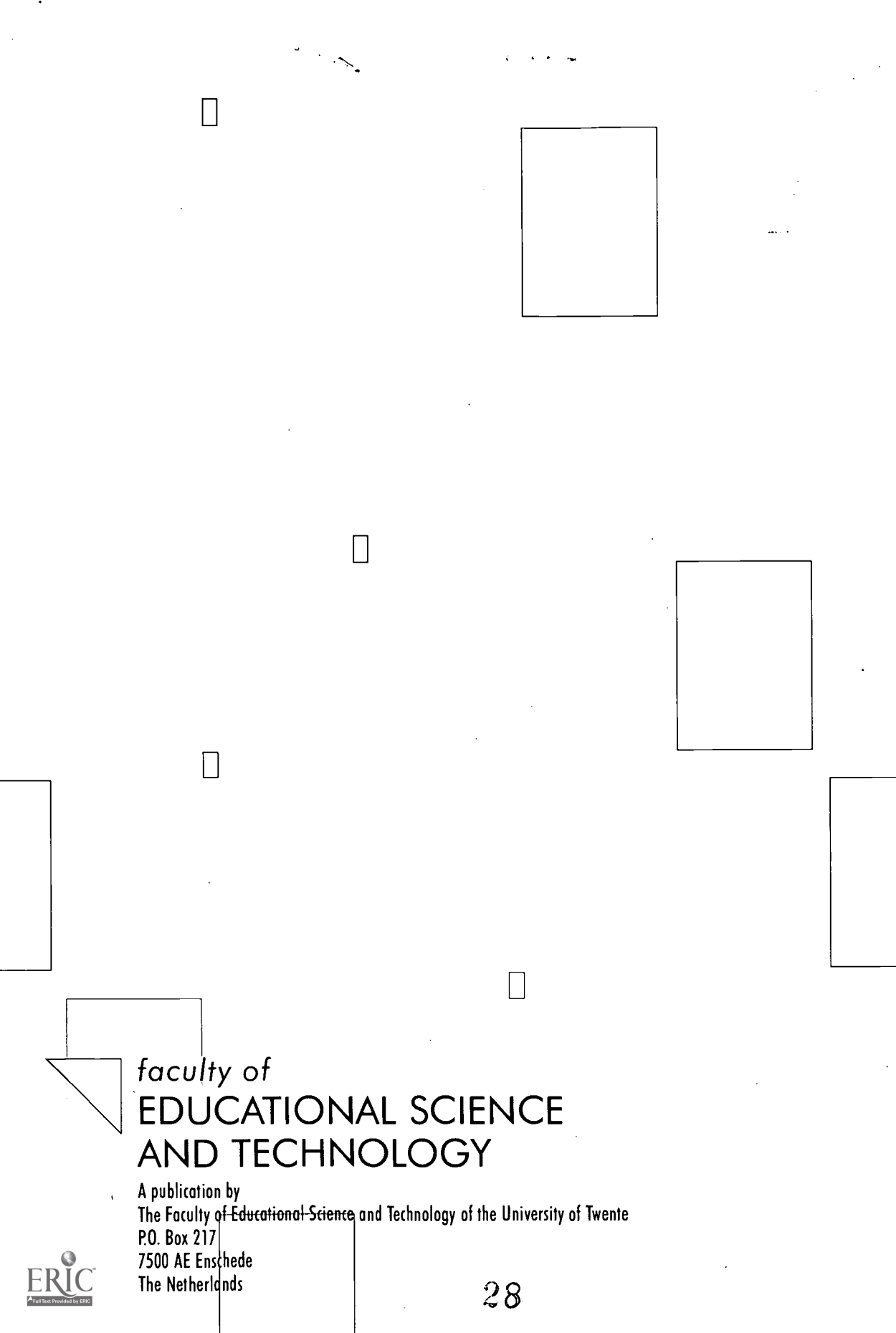
RR-94-4 M.P.F. Berger & W.J.J. Veerkamp, *A review of selection methods for optimal design*

RR-94-3 W.J. van der Linden, *A conceptual analysis of standard setting in large-scale assessments*

RR-94-2 W.J. van der Linden & H.J. Vos, *A compensatory approach to optimal selection with mastery scores*

RR-94-1 R.R. Meijer, *The influence of the presence of deviant item score patterns on the power of a person-fit statistic*

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, Mr. J.M.J. Nelissen, 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