DOCUMENT RESUME

ED 467 371 TM 034 297

AUTHOR	Hendrawan, Irene; Glas, Cees A. W.; Meijer, Rob R.
TITLE	The Effect of Person Misfit on Classification Decisions. Research Report.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
REPORT NO	RR-01-05
PUB DATE	2001-00-00
NOTE	34p.
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	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Classification; Decision Making; Estimation (Mathematics); *Goodness of Fit; *Item Response Theory; *Scores; Simulation
IDENTIFIERS	*Mastery Evaluation; *Person Fit Measures

ABSTRACT

The effect of person misfit to an item response theory (IRT) model on a mastery/nonmastery decision was investigated. Also investigated was whether the classification precision can be improved by identifying misfitting respondents using person-fit statistics. A simulation study was conducted to investigate the probability of a correct classification using different estimation methods, person-fit statistics, model violations, test lengths, and sample sizes. In this simulation study, the effect of the presence of misfitting items score patterns on the item parameter estimates was also taken into account. Results show that the effect of the presence of misfitting item score patterns on the classification of nonaberrant simulees was in general small; that is, the classification precision for these simulees hardly suffered. Further, for simulees classified as nonaberrant using a person-fit statistic, the classification decisions were comparable with a priori known nonaberrant simulees. The conclusion is that person-fit statistics can be used for identifying a subsample of respondents where relatively precise mastery/nonmastery decisions can be made. These results were comparable across different person-fit statistics and estimation methods. (Contains 7 tables and 29 references.) (Author/SLD)

TM

The Effect of Person Misfit On Classification Decisions

Research Report 01-05

Irene Hendrawan
Cees A.W. Glas
Rob R. Meijer

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of Educational Measurement and Data Analysis

The Effect of Person Misfit in Classification Decisions

Irene Hendrawan

Cees A.W. Glas

Rob R. Meijer

Abstract

The effect of person misfit to an item response theory (IRT) model on a mastery/non-mastery decision was investigated. Furthermore, it was investigated whether the classification precision can be improved by identifying misfitting respondents using person-fit statistics. A simulation study was conducted to investigate the probability of a correct classification using different estimation methods, person-fit statistics, model violations, test lengths and sample sizes. In this simulation study, the effect of the presence of misfitting item score patterns on the item parameter estimates was also taken into account. Results showed that the effect of the presence of misfitting item score patterns on the classification of non-aberrant simulees was in general small, that is, the classification precision for these simulees hardly suffered. Further, for simulees classified as non-aberrant using a person-fit statistic, the classification decisions were comparable with a priori known non-aberrant simulees. The conclusion is that person-fit statistics can be used for identifying a sub-sample of respondents where relatively precise mastery/non-mastery decisions can be made. These results were comparable across different person-fit statistics and estimation methods.

Key Words: IRT, 3PNO, Guessing, Item disclosure, Person-fit statistics, Estimation methods

Introduction

To determine the fit of an item score pattern to an item response theory (IRT, Lord, 1980; van der Linden & Hambleton, 1997) model, several fit statistics have been proposed. Investigating the fit of an item score pattern to an IRT model may help the researcher to obtain additional information about the response behavior of a person, which may, for instance, be influenced by guessing, or preknowledge of the items. For an overview of person-fit research, see Meijer and Sijtsma (2001). From this overview it can be concluded that many person-fit studies have concentrated on the power of person-fit methods to detect misfitting item score patterns. In simulation studies, the percentage of correctly classified misfitting item score patterns is investigated given a priori knowledge of some type of misfitting response behavior. A general conclusion from these studies is that the power of the person-fit statistics is relatively low (Meijer & Sijtsma, 2001). However, relatively low power of a person-fit statistic should be interpreted in relation to the effect on the estimation of the latent trait θ or the classification of a person in a prespecified category (for example, when taking mastery/non-mastery decisions). Therefore, knowledge about the effect of misfit of an item score pattern on the classification is crucial for the use of person-fit in an applied setting. To know what type of misfit has an effect on classification decisions may help the researcher to test against specific types of aberrant behavior. In this paper, we will first investigate the robustness of the classification decision under different types of misfit using different methods to estimate θ.

In realistic situations, it is unknown which respondents are aberrant. Therefore, it must be expected that the item parameter estimates will be biased by the contamination of the data. The impact of the presence of non-fitting response patterns on the outcome of the person fit tests and on the classification decisions will be the second topic of the investigation. More specifically, it will be investigated whether person-fit statistics can be used for identifying a sub-sample of respondents where relatively precise mastery/non-mastery decisions can be made.

Three estimation methods for θ will be used: a Maximum Likelihood Estimation (MLE) method and two Bayesian methods. The first Bayesian method is based on an Expected A Posterior (EAP) estimate of θ given estimates of the item parameters, the second Bayesian method is based on the complete posterior distribution of θ and the item parameters. The latter method has the advantage that both the uncertainty about the person and the uncertainty about the item parameters are taken into account. The computations for the latter procedure are performed using a Markov Chain Monte Carlo (MCMC) algorithm.

This paper is organized as follows. First, we will introduce the relevant IRT model and some methods to estimate θ. Second, we will discuss person-fit statistics that are often used in practice. Third, we will present the result of a simulation study in which we investigate the effect of person misfit and the performance of person fit tests for different levels of misfit, test lengths, sample sizes, estimation methods and test statistics.

IRT Model

In IRT, the probability of a correct response on item j ($j = 1, 2, ..., k$), $P_j(\theta)$, is a function of the latent trait value θ and a number of item characteristics. Models that are most often used are the one-, two-, and three-parameter logistic model (1-, 2-, and 3-PL; Hambleton and Swaminathan, 1985, pp. 35-48). In this study, however, we will use the 3-parameter normal ogive (3PNO; Lord, 1980, pp. 13-14) model, because in a Bayesian framework, the 3PNO model has some computational advantages over the logistic models (see, for example, Albert, 1992). The 3PNO model and 3PL model are completely equivalent for all practical purposes. In the 3PNO model, the item is characterized by a difficulty parameter β_j, a discrimination parameter α_j, and a (pseudo) guessing probability γ_j. The probability of correctly answering an item is given by

$$P_j(\theta) = \gamma_j + (1 - \gamma_j)\, \Phi(\alpha_j \theta - \beta_j), \tag{1}$$

where Φ is the standard cumulative normal distribution. Since the 3PNO and 3PL models predict nearly identical item response functions (IRFs), few differences in either model fit or parameter estimates are expected (Embretson & Reise, 2000, pp. 78-79).

Methods for Estimating θ

MLE

In IRT, the probability of a correct response on an item depends on θ, and the parameters that characterize the item. Both θ and the item parameters are unknown. Almost all person-fit studies have used the MLE method to estimate θ (Meijer & Sijtsma, 1995). The MLE, estimator $\widehat{\theta}$ is the value of θ that maximizes the likelihood function for a particular response pattern. It is usually assumed that the items fit the IRT model and that the item parameters are known. Advantages of MLE is that $\widehat{\theta}$ values tend to be consistent and efficient (Hambleton & Swaminathan, 1985). The main disadvantages of MLE are that $\widehat{\theta}$ does not exist for a perfect item score pattern (all items correct) and pattern with all items incorrect, and that $\widehat{\theta}$ is biased, that is, it is overestimated for positive values of θ and underestimated for negative values of θ (Lord, 1983; see Warm, 1989, for improvements). Though it is assumed that the item parameters are known, in realistic situations they are unknown and have to be estimated. In the simulation studies reported below, the item parameters will be estimated using the a maximum marginal likelihood (MML) procedure implemented in BILOG-MG. For more information about MLE procedure for the normal ogive model, refer to Baker (1992).

EAP

As Bayesian alternatives to MLE in person-fit research, Reise (1995) used EAP estimation and Glas and Meijer (2001) used MCMC estimation. In EAP estimation, both the response vector and information about the examinees are combined. The posterior distribution is proportional to the product of the likelihood of the item score pattern given θ and a, usually, normal prior for θ. The EAP estimate is simply the mean of the posterior

distribution. An advantage of EAP estimation is that the extra information obtained using the prior can improve the estimation of θ, and unreasonable values of $\widehat{\theta}$ can be avoided. This, of course, has the side effect that the resulting $\widehat{\theta}$ will regress to the mean of the prior (shrinkage). Also in EAP estimation, the item parameters are imputed as fixed known constants. Below, values for these constants are estimated using the Bayes modal procedure implemented in BILOG-MG (Mislevy & Bock, 1990).

MCMC

Whereas procedures for conventional frequentist statistical inference focus attention on point estimates and their standard errors, Bayesian methods often seek to characterize the posterior distribution of the parameters. This can done by an MCMC method, which produces samples from the joint posterior density of model parameters that may be then summarized to estimate θ (Jackman, 2000). Because this technique is somewhat less well known as MLE and EAP methods we will discuss it in more detail.

Below, the MCMC method used will be the Gibbs sampler (Geman & Geman, 1984, Gelfand & Smith, 1990). To implement the Gibbs sampler, the parameter vector is divided into a number of components, and each successive component is sampled from its conditional distribution given sampled values for all other components. This sampling scheme is repeated until the sampled values form stable estimates of the posterior distributions. Albert (1992) applies Gibbs sampling to estimate the parameters of the well known 2PNO model; a generalization to the 3PNO model is given by Béguin and Glas (in press). The latter generalization entails a data augmentation scheme defined as follows. Let the binary variable W_{ij} be defined as:

$$W_{ij} = \begin{cases} 1 & \text{if person } i \text{ knows the correct answer to item } j \\ 0 & \text{if person } i \text{ doesn't know the correct answer to item } j \end{cases} . \qquad (2)$$

The relation between $W_{ij} = 1$ and observed response variable Y_{ij} is given by a model where $\Phi\left(\eta_{ij}\right)$, with $\eta_{ij} = \alpha_j\theta_i - \beta_j$, is the probability that the respondent knows the item and gives a correct response with probability one, and a probability $\left(1 - \Phi\left(\eta_{ij}\right)\right)$ that

the respondent does not know the item and guesses with γ_j as the probability of a correct response. The data are also augmented with latent data Z_{ij} which are independent and normally distributed with mean η_{ij} and a standard deviation equal one. These variables are related to W by $Z_{ij} > 0$ if $W_{ij} = 1$ and $Z_{ij} \leq 0$ if $W_{ij} = 0$. The aim of the procedure is to simulate samples from the joint posterior distribution, $p(\alpha, \beta, \gamma, \theta, z, w|y)$, where the data Y are responses of n test simulees to k items. The procedure to calculate the posterior distribution, consists of the following steps:

1. Draw from the posterior $p(z, w|y; \alpha, \beta, \gamma, \theta)$ via the data augmentation model;
2. Draw from the conditional distribution of θ given Z and α, β via a normal regression model;
3. Draw from the conditional distribution of the parameters of item j, α_j and β_j via a normal regression model;
4. Draw from conditional distribution of γ_j, which is a beta distribution when the conjugate Beta prior is used.

Convergence is evaluated by comparing the between and within sequence variance. Bayes modal estimates obtained using a standard software package as BILOG-MG can be used as starting points. The estimate $\widehat{\theta}$ is taken as the average of the draws in Step 2. So also in this case, the point estimate of θ is an expected a posterior estimate. For more information on this algorithm refer to Albert (1992) and Béguin and Glas (in press).

Person-Fit Statistics

Several person-fit statistics for investigating the goodness of fit of item score patterns have been proposed. In this paper, we will use a number of statistics that have been most often used in the literature (Meijer & Sijtsma, 2001). Glas and Meijer (2001) found that these statistics had an acceptable type I error rate when simulating the distribution for these statistics using an MCMC method. Type I error rate indicates the number of incorrectly rejected null hypotheses based on the statistical tests. The following statistics were used.

The W statistic (Wright and Stone, 1979) is defined by

$$W = \frac{\sum_{j=1}^{k} [Y_j - P_j(\theta)]^2}{\sum_{j=1}^{k} P_j(\theta)[1 - P_j(\theta)]}, \tag{3}$$

where the difference between the item score Y_j and the expected item score $P_j(\theta)$ is weighted by the variance of the item score.

A related statistic was proposed by Smith (1985,1986) where the set of test items is divided into S non-overlapping subtests denoted A_s ($s = 1, ..., S$). Then the unweighted between-sets fit statistics UB is defined as

$$UB = \frac{1}{S-1} \sum_{s=1}^{S} \frac{\left\{\sum_{j \in A_s} [Y_j - P_j(\theta)]\right\}^2}{\sum_{j \in A_s} P_j(\theta)[1 - P_j(\theta)]}. \tag{4}$$

The UB statistic is a weighted W statistic computed at the subtest level.

Two other statistics, ζ_1 and ζ_2, were proposed by Tatsuoka (1984). The ζ_1 statistic is the standardization with a mean of 0 and unit variance of

$$\zeta_1^* = \sum_{j=1}^{k} [P_j(\theta) - Y_j](n_j - \overline{n}), \tag{5}$$

where n_j denotes the number of correct answers to item j and $\overline{n}$ denotes the mean number of correctly answered items in the test. The index will be positive when easy items are incorrectly answered and difficult items are correctly answered, and it will also be positive if the number of correctly answered items deviates from the overall mean score of the respondents. If a response pattern is misfitting in both senses, the magnitude of the index will be largely positive. The ζ_2 statistic is a standardization of

$$\zeta_2^* = \sum_{j=1}^{k} [P_j(\theta) - Y_j](P_j(\theta) - R/k), \tag{6}$$

where R is the person's number-correct score on the test. The index will be positive if the response pattern is misfitting in the sense that easy items are incorrectly answered and difficulty items are correctly answered; the overall response tendencies of the total sample of persons is not important here.

Another well-known person-fit statistic is the log-likelihood statistic

$$l = \sum_{j=1}^{k} \{Y_j \ln P_j(\theta) + (1 - Y_j) \ln [1 - P_j(\theta)]\}, \tag{7}$$

which was first proposed by Levin and Rubin (1979). It was further developed in Drasgow, Levine, and Williams (1985), and Drasgow, Levine, and McLaughlin (1991).

Drasgow et al. (1985) proposed a standardized version l_z of l which is asymptotically standard normally distributed; l_z is defined as

$$l_z = \frac{l - E(l)}{(Var(l))^{1/2}}, \tag{8}$$

where $E(l)$ and $Var(l)$ denote the expectation and the variance of l, respectively. The person-fit statistic l_z is often used, but Molenaar and Hoijtink (1990), and Van Krimpen-Stoop and Meijer (1999) showed that the distribution of l_z is negatively skewed. This skewness influences the differences between nominal and empirical Type I error rates for small Type I error values. They found that increasing the item discrimination resulted in a distribution that was more negatively skewed. In an MCMC framework, the distribution of a statistic is simulated, so its skewness is of minor importance. Therefore, we will only consider the person-fit statistic l instead of l_z.

Evaluating the fit of an item score pattern.

To evaluate the fit of an item score pattern a norm distribution is needed for classifying an item score pattern as fitting or misfitting. This norm distribution can be obtained using a theoretical distribution (e.g., a normal distribution) or it can be simulated. In this paper, we will simulate the norm distribution because often the

theoretical distributions proposed in the literature are not in the agreement with the empirical distributions (Meijer & Sijtsma, 2001). Also, the error in the item and person parameters can be taken into account when we simulate the norm distribution. Recently, Glas and Meijer (2001) used a Bayesian approach. Their approach has the advantage that they take into account the uncertainty of the parameters in the IRT model. In this Bayesian method, the posterior distribution of the parameters of the 3PNO model, say $p(\xi|y)$ where ξ are the item and person parameters in the model and y is the observed data, is simulated using the MCMC method given above. Person fit is then evaluated using a posterior predictive check based on an index $T(y,\xi)$ where T refers to the person-fit statistics given above. When the Markov chain has converged, draws from the posterior distribution can be used to generate model-conform data y^{rep} and to compute the Bayes p-value defined by

$$\text{Bayes } p\text{-value} = \Pr\left(T\left(y^{rep},\xi\right) \geq T\left(y,\xi\right)|y\right). \tag{9}$$

Thus, the Bayes p-value is defined as the probability that the replicated data are more extreme than the observed data. Posterior predictive checks are performed by inserting the person-fit statistics: l, W, UB, ζ_1, and ζ_2 into equation (9). After the burn-in period, when the Markov Chain has converged, in every n-th iteration ($n \geq 1$), using the current draw of the item and person parameters, a person-fit statistic $T(y,\xi)$ is computed, a new model conform response pattern is generated, and a value $T(y^{rep},\xi)$ is computed. Finally, a Bayes p-value is computed as the proportion of iterations where $T(y^{rep},\xi) \geq T(y,\xi)$.

Simulation Studies

The objective of these studies was to assess the probability of correctly classifying a person according to his or her θ value. A simulee was classified as a master if $\theta > 0$ and as a non-master if $\theta \leqslant 0$. The cut-off score was chosen equal to zero. A classification error arises when the sign of the generating value of θ is not equal to the sign of the estimate.

The reason for choosing a cut-off score equal to zero is to minimize the effect of the bias of $\widehat{\theta}$. For example, when an EAP estimator is used to estimate θ, the estimate shrinks towards the mean of θ, which is assumed to be equal to zero (Mislevy, 1986). Thus, the bias is smallest at the mean of zero.

Model violations

Guessing

To investigate the robustness of the classification decision, random guessing was simulated for a subset of the simulees in the data on part of the items on the test. For that part of the test, these simulees were randomly responding with the probability of a correct score equal to 0.20. Random response behavior may result from disinterestedness in the test in situations where, for example, a test is used to evaluate educational achievement without consequences for individual student. We simulated guessing on the easiest items, because there the effect of guessing to the estimate of ability is most detrimental (Meijer & Nering, 1997). Note that when guessing occurs only on the easy items, the actual score will be lower than expected. This implies that the resulting ability estimate will also be lower than the 'true' ability predicted by the model.

Item Disclosure

It is possible that a person has preknowledge of some of the items in the test, either about the type of test questions or about the correct answers, for example, as a result of repeated test taking. Item disclosure may result in a larger percentage of the correct answers than expected.

Note, that in general it is unknown on which of the items and on how many items a person has knowledge of the correct answers. Item preknowledge on a few items will only have a minor effect on the number-correct score (Meijer & Nering, 1997). Also, item preknowledge of the correct answers on the easiest items in the test will only slightly improve the number-correct score. This suggests that preknowledge on the items of median and high difficulty may have the most profound effect on the total score. Therefore, in this paper, we will always assume that item disclosure will affect only the

difficult items. For these items, simulees will get a correct score with a probability equal to 0.80.

When item disclosure affects only the difficult items, the actual score will be higher than expected. Thus, the effect of item preknowledge is the largest for persons with low θ that answer many difficult items correctly. For these persons, the resulting $\widehat{\theta}$ will be higher than the 'true' θ predicted by the model.

Simulation Methods

Data Generation

The item parameters were chosen as follows. The γ-parameter was fixed to 0.20 for all items. Item difficulty and discrimination parameters were chosen as

- for a test length $k = 30$, three values of the discrimination parameter, 0.5, 1.0, and 1.5, were crossed with ten item difficulties $\beta_i = -2.00 + 0.40(i - 1), i = 1, ..., 10$.
- for a test length $k = 60$, three values of discrimination parameters, 0.5, 1.0, and 1.5, were crossed with twenty item difficulties $\beta_i = -2.00 + 0.20(i - 1), i = 1, ..., 20$.

The ability parameters were drawn from a standard normal distribution. Using these item and ability parameters, data were generated, the parameters were both estimated using BILOG-MG followed by MLE and EAP and by the MCMC method. Then, the item score patterns were classified as normal or aberrant using the person-fit statistics discussed above. Because we were interested in the robustness of the classification decision of an individual person under model violations, we computed the proportion of correctly identified simulees $\widehat{\theta} > 0$ with $\theta > 0$ [i.e., $P(\widehat{\theta} > 0 \,|\, \theta > 0)$] and the proportion of correctly identified simulees $\widehat{\theta} \leqslant 0$ with $\theta \leqslant 0$ [i.e., $P(\widehat{\theta} \leqslant 0 \,|\, \theta \leqslant 0)$].

Guessing

Two sample sizes were used $n = 400$ with 40 misfitting simulees, and $n = 1000$ with 100 misfitting simulees. *Guessing* was simulated on $1/6$, $1/3$, and $1/2$ of the easy items in the test. The probability of a correct response to these items was chosen to be 0.20. For every condition, 100 replications were made with a nominal significance probability of 0.05 for every person-fit statistic.

To investigate the robustness of the classification decisions, we first determined the proportion of correct mastery decisions in a group with only fitting item response patterns. Then, we determined the proportion correct mastery decisions in a group with fitting and misfitting simulees where the item parameters were estimated using both the fitting and misfitting simulees. Finally, we determined the proportions of correct mastery decisions in the groups of simulees classified as fitting and misfitting on the basis of a person-fit statistic.

Item Disclosure

The setup of the simulation study for *item disclosure* was analogous to study for the guessing. Data were generated for sample sizes of $n = 400$ and $n = 1000$ simulees, and test lengths of $k = 30$ and $k = 60$ items. Item disclosure was simulated for 10% of the simulees, and for these simulees, $1/6$, $1/3$, and $1/2$ of the difficult items in the test were affected. The probability of a correct response to these items was chosen to be 0.80. Test statistics were computed in the same way as in the guessing study. Again, in every condition, 100 replications were made with a nominal significance probability of 0.05 for every person-fit statistic.

Again, the proportion of correct mastery decisions without the presence of misfitting simulees was used as a base rate. Furthermore, we determined the proportion of correct mastery decisions for fitting or misfitting simulees, and for fitting and non-fitting simulees as identified using person fit statistics, respectively.

The MCMC procedure

For the MCMC procedure, a run length of 4000 iterations with a burn-in period of 1000 iterations was chosen (see Albert, 1992). That is, the first 1000 iterations were discarded. In the remaining 3000 iterations, $T(y^{rep}, \xi)$ and $T(y, \xi)$ were computed every fifth iteration. So the posterior predictive checks were based on 600 draws. For the statistics that use a partitioning of the items into subtests, the items were ordered according to their item difficulty β and then two subtests of equal size were formed, one with the difficult and one with the easy items.

Results

Guessing

For a sample with all item score patterns fitting, the proportion of correct mastery decisions for $n = 400$ and $n = 1000$ simulees with $k = 30$ and $k = 60$ items is given in Table 1. We divided the simulees into two groups based on θ: groups with $\theta \leqslant 0$ and $\theta > 0$. Recall that in this setup, the proportion of correct mastery decisions is defined as the conditional probability $P(\widehat{\theta} \leqslant 0 \,|\, \theta \leqslant 0)$ for $\theta \leqslant 0$ and defined as $P(\widehat{\theta} > 0 \,|\, \theta > 0)$ for $\theta > 0$, respectively. For the example, in Table 1 for the combination $n = 400$ and $k = 30$, the proportion of correct mastery decisions using the MCMC method is 0.89 for $\theta \leqslant 0$. This means that using the MCMC method 89% of the simulees have $\widehat{\theta}$ estimates less than or equal to zero in the group with $\theta \leqslant 0$.

Insert Table 1 about here

Comparing the proportions of correct mastery decisions in Table 1, it can be seen that there are no main effects of mastery ($\theta \leqslant 0$ versus $\theta > 0$), and estimation method. Furthermore, the proportion of correct mastery decisions is little affected by the sample size, that is, by the precision of the item parameter estimates. There is, however, a main effect of test length, that is, all proportions for $k = 30$ are less than those for $k = 60$. This result is as expected because using longer tests will result in a higher proportion of correct mastery decisions for the normal simulees.

Insert Table 2 about here

Table 2 gives the proportions of correct mastery decisions for data sets with 10% guessing simulees. The item parameters were estimated using the data of fitting and misfitting simulees simultaneously. Comparing the proportion of correct mastery decisions for $n = 400$ in the normal (non-aberrant) group with $\theta \leqslant 0$ and with $\theta > 0$ (Table 2, upper panel), the proportion of correct mastery decisions in the normal group

with $\theta > 0$ is higher across all estimation methods for $p = 1/6$, $p = 1/3$, and $p = 1/2$, than those in the normal group with $\theta \leqslant 0$. There is one exception for the MCMC method, where for $\theta > 0$, the proportion is 0.71 and for $\theta \leqslant 0$ is 0.85. However, for the guessing simulees, the proportion of correct mastery decisions in the group with $\theta > 0$ is always smaller than in the group with $\theta \leqslant 0$ for $p = 1/6$, $p = 1/3$, and $p = 1/2$. In general, for $p = 1/2$ the proportion of correct mastery decisions in the group $\theta \leqslant 0$ is almost equal to one. This can be explained by noting that guessing is imposed on the easy items resulting in lower $\widehat{\theta}$ than true θ.

For both test lengths, for $k = 30$ and $k = 60$, it can be seen that if p increases, the proportion of correct mastery decisions decreases in the normal group with $\theta \leqslant 0$ and it increases for the normal group with $\theta > 0$. In contrast, for the guessing simulees the proportion of correct mastery decisions increases for $\theta \leqslant 0$ and it decreases for $\theta > 0$. This is due to the fact that when p increases and guessing is imposed on easy items, given a fixed score s, the probability to get a score higher than s decreases. Thus, when $\theta > 0$ it will result in a lower $\widehat{\theta}$ which implies that the number of the guessing simulees that are being misclassified increases as p increases.

The results across the different estimation methods for the normal simulees are similar across test lengths and proportion of simulated guessing p. For guessing simulees and $\theta \leqslant 0$ results are also comparable across estimation methods. However, for guessing simulees and $\theta > 0$ and $p = 1/3$ and $p = 1/2$ the proportions correct classifications differ substantially, with MCMC as the least effective estimation method. In the latter case, the estimates of θ are distorted to such a degree that the proportion of correct classifications approaches zero.

Inspection of the results in the condition with $n = 1000$ (Table 2, lower panel) and comparing these results with $n = 400$ (Table 2, upper panel) shows that the proportion of correct mastery decisions is little affected by the smaller calibration sample.

Insert Tables 3 and 4 about here

In realistic situations, it is unknown which respondents are aberrant. The objective of the following simulation study is to assess whether the same precision for mastery classifications can be attained when persons are classified as aberrant or non-aberrant using person fit statistics. The results are given in Table 3 ($k = 30$, $n = 400$) and Table 4 ($k = 60$, $n = 400$). Comparing classification rates for both $k = 30$ and $k = 60$ with the classification rates in Table 2, it can be concluded that the classification rates for the normal simulees based on a priori knowledge (Table 2) and based on the classification on the basis of a person-fit statistic are similar. In the group of guessing simulees, it can be seen that the proportion of correct mastery decisions using the person-fit statistics is less than in Table 2 for $\theta \leq 0$, whereas it is generally higher for $\theta > 0$ across p, that is, the extent to which the test is affected by guessing. This last result may be explained by noting that the group of simulees classified as misfitting consists of both simulated guessing simulees and normal simulees. Because the normal simulees have in general a higher θ value than the guessing simulees, the presence of some normal simulees may increase the average $\widehat{\theta}$ value in the guessing group. Furthermore, for $p = 1/6$, the proportion of correct mastery decisions is higher than those obtained for $p = 1/3$ and $p = 1/2$.

In Table 4, the results for $k = 60$ are depicted. In general, the proportion of correct mastery decisions is higher for $k = 60$ than for $k = 30$. For $n = 1000$ (not tabulated), analogous trends were found.

Item Disclosure

The proportion of correct mastery decisions for $n = 400$ with $k = 30$ and $k = 60$ based on a priori knowledge, that is, without using person-fit statistics, is given in Table 5 (upper panel) and $n = 1000$ (lower panel).

Insert Table 5 about here

Comparing the proportion of correct mastery decisions in the normal (non-aberrant) group with $\theta \leqslant 0$ and with $\theta > 0$ for $n = 400$ (Table 5, upper panel), the proportion of correct mastery decisions for $\theta \leqslant 0$ is a little higher than for $\theta > 0$ across all estimation

methods and proportions of item disclosure. Comparing the results with Table 1, it can be seen that the proportions are almost similar and that there is no effect of bias in the item parameter estimates as a result of the presence of 10% aberrant simulees. For the item disclosure simulees, the proportion of correct mastery decisions in the group with $\theta \leqslant 0$ is always smaller than in the group with $\theta > 0$. Because item disclosure will lead, in general, to a higher $\widehat{\theta}$ value, the proportion of correct mastery decisions for $\theta \leqslant 0$ is reduced. In general, the proportion of correct mastery decisions in the group $\theta > 0$ is almost equal to one.

Across the different p values, for $k = 30$ and $k = 60$, it can be seen that the proportion of correct mastery decisions for $\theta \leqslant 0$ and for $\theta > 0$ are similar. In contrast, for the item disclosure simulees, the proportion of correct mastery decisions decreases for $\theta \leqslant 0$ and it increases for $\theta > 0$. This is due to the fact that when p increases and item disclosure is imposed on difficult items, given a fixed score s, the probability to get a score higher than s increases. Thus, when $\theta \leqslant 0$ it will result in a higher $\widehat{\theta}$, which implies that the number of the item disclosure simulees that are being misclassified increases as p increases.

With respect to the estimation methods, it can be seen that for normal simulees, there are only small (inconsistent) differences. For item disclosure simulees and $\theta > 0$ results are also comparable across estimation methods. However, for item disclosure simulees and $\theta \leqslant 0$ and $p = 1/3$ and $p = 1/2$, the MCMC method performed better than the EAP and MLE.

Inspection of the results in the condition with $n = 1000$ (Table 5, lower panel) and comparing them to the results for $n = 400$ (Table 5, upper panel) shows again that the proportion of correct mastery decisions is little affected by the smaller calibration sample.

Insert Tables 6 and 7 about here

In the Tables 6 and 7, the classification percentages using the person-fit statistics for $n = 400$ and $k = 30$ (Table 6) and $k = 60$ (Table 7) are given. From Table 6 it is clear that for normal simulees the different estimation methods result in almost the same percentages

of correct classifications across the different p values for $\theta \leqslant 0$ and $\theta > 0$. Also, similar percentages were found as in Table 5 where the normal simulees were identified based on a priori knowledge. For the item disclosure simulees, proportion correct classifications differ. Comparing the classification rates in Table 5 and Table 6 we note that in the group of item disclosure simulees the proportion of correct mastery decisions using the person-fit statistics is, in general, higher than in Table 5 for $\theta \leq 0$, whereas it is generally lower for $\theta > 0$ across different p values. This last result may be explained by noting that the group of simulees classified as misfitting consists of both simulated item disclosure and normal simulees. Because the normal simulees have in general a lower $\widehat{\theta}$ value than the item disclosure simulees, the presence of some normal simulees may reduce the average $\widehat{\theta}$ value in the item disclosure group. No large differences were found across person-fit statistics.

In Table 7, the results for $k = 60$ are given. In general, the proportion of correct mastery decisions is higher for $k = 60$ than for $k = 30$. For $n = 1000$ (not tabulated) the same trends were found and the classification percentages were similar as in Table 6 and Table 7.

Discussion

The effect of person misfit to an IRT model on a mastery/non-mastery decision was investigated and it was investigated whether using person-fit statistics can be helpful in judging the acceptability of such a decision. Results showed the following.

(1) The effect of the presence of 10% misfitting simulees had little effect on the item parameter estimates in the sense that the mastery classification of normal simulees was little affected.

(2) The classification precision of aberrant simulees was greatly affected: the precision for guessing simulees with $\theta > 0$ became virtually zero, especially when the MCMC method was applied. For item disclosure simulees with $\theta \leq 0$, the effects were less dramatic than for the guessing, although for $p = 1/2$ the classification rates were substantially lower than in the case for normal simulees.

(3) For simulees classified as *normal* by means of a person-fit statistic the classification rates were comparable with the situation where there was a priori knowledge about normal/aberrant behavior.

In general the effect of the type of estimation method and the type of person-fit statistic was small, though the ζ_2-statistic generally gave the best results. The MCMC method seemed to result in greater classification precision in the case of item disclosure. In the case of guessing, and in cases were relatively small numbers of items were misfitting ($p = 1/6$ and $p = 1/3$), the EAP method performed better. However there were no trends where one estimation method outperformed the other estimation methods.

The main conclusion is that the classification precision in the sub-sample identified as normal (non-aberrant) by a person fit statistic is comparable to the classification precision that can be attained if aberrant and non-aberrant respondents were known in advance. Classification precision for aberrant persons is erratic. This suggests that further research might be done to methods for identifying a subset of items where an aberrant person gives model-conforming responses and using this subset to estimate θ. An additional research question will then be to what extent thus shortened test length decreases the precision of the estimate of θ and the ensuing classification precision.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response functions using Gibbs sampling. *Journal of Education Statistics, 17*, 251-269.

Baker, F. B. (1992). *Item response theory: parameter estimation techniques.* New York: Marcel Dekker.

Béguin, A. A., & Glas, C. A. W. (in press). MCMC estimation of multidimensional IRT models. *Psychometrika.*

Drasgow, F., Levine, M .V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15*, 171-191.

Drasgow, F., Levine, M .V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721-741.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85,* 398-409.

Glas, C. A. W., & Meijer, R. R. (2001). A Bayesian approach to person fit analysis in item response theory models. *Research Report*, University of Twente, Enschede, The Netherlands.

Hambleton, R. K., & Swaminatan, H. (1985). *Item response theory: Principles and applications* (2rd ed.). Boston: Kluwer-Nijhoff Publishing.

Jackman, S. (2000). Estimation and inference via bayesian simulation: an introduction to markov chain monte carlo. *American Journal of Political Science, 44*, 369-398.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test score. *Journal of Education Statistics*, *4*, 269-290.

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

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

Meijer, R. R., & Nering, M. L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement*, *21*, 321-336.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: a review and new developments, *Applied Measurement in Education*, *8*, 261-272.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: evaluating person fit. *Applied Psychological Measurement*, *25*, 107-135.

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

Mislevy, R.J., & Bock, R.D. (1990). BILOG user's Gide [software manual]. Chicago: Scientific Software.

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

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement*, *19*, 213-229.

Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Education and Psychological Measurement*, *45*, 433-444.

Smith, R. M. (1986). Person fit in the Rasch model. *Education and Psychological Measurement*, *46*, 359-372.

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

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of Modern Item Response Theory*. NY:Springer Verlag.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23*, 327-345.

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

Wright, B. D., & Stone, M. H. (1979). *Best Test Design*. Chicago, IL: MESA Press University of Chicago.

Table 1
The Proportion of Correct Mastery Decisions.
All Item Score Patterns Fitting.

		n = 400		n = 1000	
	Method	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$
k = 30	MCMC	0.89	0.89	0.90	0.90
	EAP	0.89	0.89	0.90	0.89
	MLE	0.88	0.89	0.89	0.90
k = 60	MCMC	0.93	0.93	0.93	0.93
	EAP	0.93	0.92	0.93	0.93
	MLE	0.93	0.92	0.93	0.93

Table 2
The Proportion of Correct Mastery Decisions in a Sample with 10% of the Simulees Guessing.
Based on a Priori Knowledge whether a Simulee is Fitting or Misfitting.

	Method	Normal		Guessing ($p = 1/6$)		Normal		Guessing ($p = 1/3$)		Normal		Guessing ($p = 1/2$)	
		$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$
n = 400													
	MCMC	0.85	0.71	0.97	0.42	0.81	0.93	1.00	0.00	0.77	0.97	1.00	0.00
k = 30	EAP	0.87	0.92	0.97	0.58	0.80	0.95	0.94	0.39	0.78	0.97	0.96	0.24
	MLE	0.86	0.93	0.98	0.52	0.82	0.94	1.00	0.17	0.80	0.97	1.00	0.40
	MCMC	0.86	0.97	0.98	0.37	0.80	0.98	1.00	0.00	0.81	0.97	1.00	0.00
k = 60	EAP	0.89	0.96	0.99	0.48	0.85	0.96	0.98	0.37	0.82	0.98	0.99	0.17
	MLE	0.89	0.96	0.99	0.42	0.84	0.96	0.99	0.06	0.84	0.98	0.99	0.31
n = 1000													
	MCMC	0.86	0.91	0.98	0.45	0.81	0.95	1.00	0.00	0.78	0.96	1.00	0.00
k = 30	EAP	0.86	0.91	0.99	0.56	0.80	0.96	0.96	0.39	0.78	0.96	0.96	0.18
	MLE	0.86	0.92	0.98	0.57	0.80	0.96	0.99	0.25	0.80	0.95	1.00	0.33
	MCMC	0.85	0.96	0.99	0.36	0.80	0.98	1.00	0.00	0.77	0.99	1.00	0.00
k = 60	EAP	0.90	0.95	0.99	0.50	0.84	0.97	0.99	0.33	0.87	0.97	1.00	0.13
	MLE	0.90	0.95	0.99	0.47	0.84	0.97	0.99	0.12	0.86	0.97	1.00	0.38

Table 3

The Proportion of Correct Mastery Decisions in a Sample with 10% of the Simulees Guessing for n = 400 and k = 30. Person-Fit Statistics are Used to Classify whether a Simulee is Fitting or Misfitting.

Method	Test	Normal		Guessing ($p = 1/6$)		Normal		Guessing ($p = 1/3$)		Normal		Guessing ($p = 1/2$)	
		$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$
	l	0.85	0.71	0.77	0.46	0.80	0.94	0.50	0.00	0.77	0.97	0.38	0.00
	W	0.85	0.92	0.78	0.45	0.80	0.93	0.49	0.00	0.77	0.97	0.40	0.00
MCMC	UB	0.85	0.92	0.76	0.44	0.80	0.93	0.50	0.00	0.77	0.97	0.38	0.00
	ζ_1	0.85	0.92	0.67	0.49	0.81	0.93	0.58	0.00	0.77	0.97	0.69	0.00
	ζ_2	0.85	0.92	0.90	0.44	0.80	0.93	0.86	0.00	0.77	0.97	0.93	0.00
	l	0.87	0.92	0.75	0.69	0.81	0.95	0.43	0.71	0.78	0.97	0.33	0.52
	W	0.87	0.92	0.77	0.69	0.80	0.95	0.39	0.71	0.78	0.97	0.34	0.52
EAP	UB	0.87	0.92	0.75	0.69	0.80	0.95	0.41	0.73	0.78	0.98	0.33	0.55
	ζ_1	0.87	0.91	0.65	0.78	0.81	0.94	0.53	0.75	0.79	0.97	0.66	0.40
	ζ_2	0.87	0.91	0.84	0.64	0.81	0.94	0.78	0.70	0.79	0.97	0.84	0.52
	l	0.87	0.92	0.77	0.60	0.81	0.95	0.50	0.14	0.80	0.97	0.38	0.27
	W	0.87	0.92	0.79	0.58	0.81	0.95	0.49	0.14	0.80	0.97	0.40	0.29
MLE	UB	0.86	0.93	0.77	0.58	0.82	0.94	0.50	0.14	0.80	0.97	0.38	0.24
	ζ_1	0.87	0.92	0.67	0.67	0.81	0.94	0.57	0.20	0.80	0.96	0.69	0.47
	ζ_2	0.87	0.92	0.87	0.55	0.81	0.94	0.85	0.14	0.80	0.96	0.93	0.27

Table 4

The Proportion of Correct Mastery Decisions in a Sample with 10% of the Simulees Guessing for n = 400 and k = 60. Person-Fit Statistics are Used to Classify whether a Simulee is Fitting or Misfitting.

Method	Test	Normal		Guessing ($p = 1/6$)		Normal		Guessing ($p = 1/3$)		Normal		Guessing ($p = 1/2$)	
		$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$
MCMC	l	0.86	0.97	0.70	0.39	0.80	0.98	0.49	0.00	0.82	0.98	0.48	0.00
	W	0.86	0.97	0.73	0.39	0.80	0.98	0.49	0.00	0.82	0.98	0.47	0.00
	UB	0.86	0.97	0.72	0.38	0.80	0.98	0.49	0.00	0.81	0.98	0.46	0.00
	ζ_1	0.86	0.97	0.63	0.43	0.80	0.98	0.64	0.00	0.82	0.97	0.65	0.00
	ζ_2	0.85	0.97	0.89	0.40	0.80	0.98	0.93	0.00	0.82	0.97	0.96	0.00
EAP	l	0.89	0.96	0.69	0.54	0.85	0.96	0.49	0.54	0.82	0.98	0.47	0.31
	W	0.89	0.96	0.72	0.54	0.85	0.96	0.49	0.54	0.82	0.98	0.46	0.31
	UB	0.89	0.96	0.71	0.54	0.85	0.96	0.48	0.54	0.82	0.98	0.44	0.34
	ζ_1	0.89	0.96	0.62	0.76	0.86	0.97	0.62	0.67	0.82	0.98	0.62	0.30
	ζ_2	0.89	0.96	0.87	0.53	0.86	0.97	0.91	0.54	0.83	0.98	0.92	0.35
MLE	l	0.89	0.96	0.70	0.46	0.85	0.96	0.49	0.06	0.85	0.98	0.48	0.24
	W	0.89	0.96	0.73	0.46	0.85	0.96	0.49	0.06	0.85	0.98	0.47	0.24
	UB	0.89	0.96	0.72	0.46	0.85	0.96	0.49	0.06	0.84	0.98	0.46	0.24
	ζ_1	0.89	0.96	0.63	0.61	0.85	0.97	0.64	0.09	0.85	0.98	0.65	0.33
	ζ_2	0.89	0.96	0.89	0.45	0.85	0.97	0.93	0.06	0.84	0.98	0.96	0.24

Table 5
The Proportion of Correct Mastery Decisions in a Sample with 10% of the Simulees Item Disclosure. Based on a Priori Knowledge whether a Simulee is Fitting or Misfitting.

	Method	Normal		Guessing ($p = 1/6$)		Normal		Guessing ($p = 1/3$)		Normal		Guessing ($p = 1/2$)	
		$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$
n = 400													
	MCMC	0.90	0.88	0.85	0.89	0.90	0.87	0.71	0.94	0.92	0.90	0.50	0.96
k = 30	EAP	0.91	0.88	0.80	0.93	0.91	0.87	0.51	0.96	0.92	0.88	0.22	0.99
	MLE	0.91	0.89	0.82	0.91	0.90	0.87	0.66	0.96	0.91	0.88	0.43	0.98
	MCMC	0.93	0.92	0.86	0.96	0.92	0.93	0.75	0.96	0.92	0.90	0.49	0.99
k = 60	EAP	0.94	0.91	0.82	0.96	0.94	0.92	0.55	0.97	0.92	0.89	0.19	1.00
	MLE	0.94	0.91	0.84	0.96	0.93	0.92	0.71	0.96	0.92	0.89	0.46	0.99
n = 1000													
	MCMC	0.89	0.89	0.83	0.94	0.91	0.89	0.69	0.98	0.92	0.88	0.51	0.99
k = 30	EAP	0.89	0.89	0.79	0.95	0.91	0.88	0.55	0.99	0.93	0.88	0.26	0.99
	MLE	0.89	0.89	0.82	0.96	0.91	0.88	0.66	0.98	0.92	0.89	0.49	0.99
	MCMC	0.93	0.93	0.88	0.96	0.93	0.91	0.73	0.98	0.94	0.91	0.48	0.99
k = 60	EAP	0.94	0.92	0.86	0.96	0.94	0.89	0.56	0.99	0.94	0.90	0.22	0.99
	MLE	0.93	0.92	0.88	0.96	0.94	0.90	0.70	0.98	0.94	0.90	0.47	0.99

Table 6

The Proportion of Correct Mastery Decisions in a Sample with 10% of the Simulees Item Disclosure for n = 400 and k = 30. Person-Fit Statistics are Used to Classify whether a Simulee is Fitting or Misfitting.

Method	Test	Normal		Guessing ($p = 1/6$)		Normal		Guessing ($p = 1/3$)		Normal		Guessing ($p = 1/2$)	
		$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$
	l	0.90	0.89	0.87	0.46	0.90	0.87	0.84	0.63	0.91	0.90	0.66	0.60
	W	0.90	0.88	0.87	0.52	0.90	0.87	0.79	0.67	0.91	0.90	0.67	0.60
MCMC	UB	0.90	0.89	0.91	0.43	0.90	0.87	0.86	0.55	0.91	0.90	0.69	0.59
	ζ_1	0.90	0.88	0.86	0.74	0.90	0.87	0.73	0.87	0.91	0.90	0.56	0.91
	ζ_2	0.90	0.88	0.83	0.79	0.89	0.87	0.69	0.87	0.91	0.90	0.57	0.94
	l	0.91	0.88	0.78	0.49	0.91	0.87	0.57	0.65	0.91	0.88	0.32	0.62
	W	0.91	0.88	0.74	0.54	0.90	0.87	0.54	0.69	0.91	0.88	0.32	0.64
EAP	UB	0.91	0.88	0.79	0.46	0.90	0.87	0.62	0.57	0.91	0.88	0.32	0.59
	ζ_1	0.91	0.88	0.74	0.75	0.91	0.87	0.50	0.90	0.92	0.88	0.24	0.96
	ζ_2	0.91	0.88	0.68	0.86	0.91	0.87	0.49	0.90	0.92	0.88	0.23	0.97
	l	0.91	0.89	0.84	0.47	0.90	0.87	0.78	0.67	0.91	0.88	0.58	0.62
	W	0.90	0.89	0.85	0.53	0.90	0.87	0.71	0.74	0.91	0.88	0.58	0.63
MLE	UB	0.90	0.89	0.88	0.43	0.90	0.87	0.83	0.56	0.91	0.88	0.61	0.59
	ζ_1	0.91	0.89	0.84	0.75	0.91	0.87	0.68	0.91	0.91	0.88	0.48	0.95
	ζ_2	0.91	0.88	0.79	0.85	0.91	0.87	0.66	0.91	0.92	0.88	0.47	0.97

30

Table 7

The Proportion of Correct Mastery Decisions in a Sample with 10% of the Simulees Item Disclosure for n = 400 and k = 60. Person-Fit Statistics are Used to Classify whether a Simulee is Fitting or Misfitting.

Method	Test	Normal		Guessing ($p = 1/6$)		Normal		Guessing ($p = 1/3$)		Normal		Guessing ($p = 1/2$)	
		$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$	$\theta \leqslant 0$	$\theta > 0$
	l	0.93	0.91	0.85	0.67	0.92	0.94	0.78	0.86	0.93	0.90	0.55	0.93
	W	0.93	0.92	0.84	0.71	0.92	0.94	0.76	0.90	0.92	0.90	0.56	0.94
MCMC	UB	0.93	0.92	0.88	0.51	0.92	0.94	0.83	0.82	0.92	0.90	0.57	0.90
	ζ_1	0.93	0.91	0.83	0.88	0.92	0.94	0.76	0.95	0.92	0.90	0.52	0.97
	ζ_2	0.93	0.91	0.80	0.96	0.92	0.94	0.76	0.95	0.92	0.90	0.51	0.98
	l	0.94	0.91	0.81	0.69	0.94	0.92	0.56	0.88	0.92	0.89	0.20	0.97
	W	0.94	0.91	0.80	0.76	0.94	0.92	0.55	0.91	0.92	0.89	0.21	0.98
EAP	UB	0.94	0.91	0.83	0.54	0.94	0.92	0.58	0.88	0.92	0.89	0.21	0.96
	ζ_1	0.94	0.91	0.81	0.92	0.94	0.92	0.55	0.96	0.92	0.89	0.19	1.00
	ζ_2	0.94	0.92	0.76	0.97	0.94	0.92	0.55	0.97	0.92	0.89	0.19	1.00
	l	0.93	0.92	0.84	0.68	0.93	0.92	0.73	0.86	0.92	0.89	0.49	0.93
	W	0.93	0.92	0.84	0.75	0.93	0.92	0.71	0.90	0.92	0.89	0.49	0.94
MLE	UB	0.93	0.92	0.87	0.54	0.93	0.92	0.80	0.84	0.92	0.89	0.49	0.91
	ζ_1	0.94	0.91	0.83	0.90	0.94	0.92	0.73	0.95	0.92	0.89	0.47	0.97
	ζ_2	0.94	0.92	0.80	0.96	0.94	0.92	0.72	0.95	0.92	0.89	0.46	0.98

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

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*

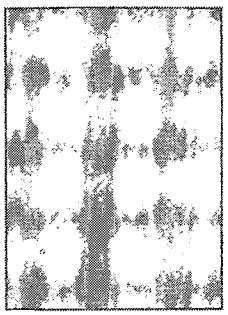
RR-98-12 W.J. van der Linden, *Optimal Assembly of Tests with Item Sets*

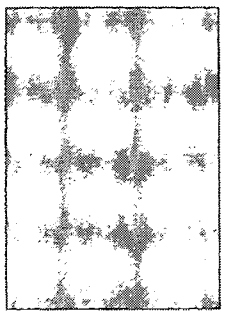
RR-98-11 W.J. van der Linden, B.P. Veldkamp & L.M. Reese, *An Integer Programming Approach to Item Pool Design*

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").