DOCUMENT RESUME

ED 467 375 TM 034 301

AUTHOR	Glas, Cees A. W.; Meijer, Rob R.
TITLE	A Bayesian Approach to Person Fit Analysis in Item Response Theory Models. Research Report.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
REPORT NO	RR-01-09
PUB DATE	2001-00-00
NOTE	41p.
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	*Bayesian Statistics; *Item Response Theory; Markov Processes; Models; Monte Carlo Methods; Simulation
IDENTIFIERS	Normal Ogive Models; *Person Fit Measures; Three Parameter Model; Type I Errors

ABSTRACT

A Bayesian approach to the evaluation of person fit in item response theory (IRT) models is presented. In a posterior predictive check, the observed value on a discrepancy variable is positioned in its posterior distribution. In a Bayesian framework, a Markov Chain Monte Carlo procedure can be used to generate samples of the posterior distribution of the parameters of interest. These draws can also be used to compute the posterior predictive distribution of the discrepancy variable. The procedure is worked out in detail for the three-parameter normal ogive model, but it is also shown that the procedure can be directly generalized to many other IRT models. Type I error rate and the power against some specific model violations are evaluated using a number of simulation studies. (Contains 8 figures and 55 references.) (Author/SLD)

A Bayesian Approach to Person Fit Analysis in Item Response Theory Models

Research Report 01-09

Cees A.W. Glas
Rob R. Meijer

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

A Bayesian Approach to Person Fit Analysis In Item Response Theory Models

Cees A.W. Glas

Rob R. Meijer

Abstract

A Bayesian approach to the evaluation of person fit in item response theory (IRT) models is presented. In a posterior predictive check, the observed value on a discrepancy variable is positioned in its posterior distribution. In a Bayesian framework, a Markov chain Monte Carlo procedure can be used to generate samples of the posterior distribution of the parameters of interest. These draws can also be used to compute the posterior predictive distribution of the discrepancy variable. The procedure is worked out in detail for the 3-parameter normal ogive model, but it is also shown that the procedure can be directly generalized to many other IRT models. Type I error rate and the power against some specific model violations are evaluated using a number of simulation studies. *Index terms: Bayesian statistics, item response theory, person fit, model fit, 3-parameter normal ogive model, posterior predictive check, power studies, type I error.*

Applications of item response theory (IRT) models to the analysis of test items, tests, and item score patterns are only valid if the IRT model holds. Fit of items can be investigated across persons and fit of persons can be investigated across items. Item fit is important because in psychological and educational measurement, instruments are developed that are used in a population of persons; item-fit then can help the test constructor to develop an instrument that fits an IRT model in that particular population. Item-fit statistics have been proposed by, for example, Mokken (1971), Andersen (1973), Yen (1981, 1984), Molenaar (1983), Glas (1988, 1999), and Orlando and Thissen (2000). As a next step, the fit of an individual's item score pattern can be investigated. Although a test may fit an IRT model, persons may produce patterns that are unlikely given the model, for example, because they have preknowledge of the correct answers to some of the most difficult items. Investigation of person fit may help the researcher to obtain additional information about the answering behavior of a person. By means of a person-fit statistic, the fit of a score pattern can be determined given that the IRT model holds. Some statistics can be used to obtain information at a subtest level and a more diagnostic approach can be followed. Meijer and Sijtsma (1995; in press) give an overview of person-fit statistics proposed for various IRT models.

To decide whether an item score pattern fits an IRT model, a sampling distribution under the null model, that is, the IRT model, is needed. Let t be the observed value of a person-fit statistic T. Then the significance probability or probability of exceedance is defined as the probability under the sampling distribution that the value of the test statistic is equal or smaller than the observed value, that is, $p = P(T \leqslant t)$, or equal or larger than the observed value, that is, $p = P(T \geq t)$, depending on whether low or high values of the statistic indicate aberrant item score patterns. As will be discussed below, for some statistics theoretical asymptotic or exact distributions are known which can be used to classify an item score pattern as fitting or nonfitting. An alternative is to simulate data according to an IRT model based on the estimated item parameters and then determine p empirically (e.g., Reise, 1995, 1999; Reise & Widaman, 1999; Meijer & Nering, 1997). However, the true values of both the item and person parameters are unknown, and the uncertainty about these parameters is often not taking into account when simulating an empirical distribution.

In this article, we will explore an alternative approach based on a Bayesian framework and posterior predictive checks using Markov chain Monte Carlo (MCMC) methods. For more information on posterior predictive checks, refer to Meng (1994), Gelman, Carlin, Stern, and Rubin (1995), and Gelman, Meng, and Stern (1996). In principle, this approach applies to any IRT model, but in this study we will focus on the 3-parameter normal ogive (3PNO) model.

Compared to the traditional frequentist approach, this Bayesian approach has several advantages. First, there is no need to derive the theoretical sampling distribution of the statistic, which sometimes may be very difficult, if not impossible. Second, the person-fit statistic may depend on unknown quantities as the item and person parameters which uncertainty is explicitly taken into account. The third advantage pertains to generality of the procedure. Simulation studies have show that a fully Bayesian approach to estimation of the parameters in simple IRT models (say 1- or 2-parameter models) are generally not superior to estimates obtained by a maximum marginal likelihood (MML) procedure or a Bayes modal procedure (see, for instance, Baker, 1998, or Kim, 2001). However, the Bayesian approach also applies to complicated IRT models, where MML or Bayes modal approaches pose important problems. Recently, the fully Bayesian approach has been adopted to the estimation of IRT models with multiple raters, multiple item types, missing data (Patz & Junker, 1997, 1999), testlet structures (Bradlow, Wainer & Wang, 1999, Wainer, Bradlow & Du, 2000), latent classes (Hoijtink & Molenaar, 1997), models with a multi-level structure on the ability parameters (Fox & Glas, 2001) and the item parameters (Janssen, Tuerlinckx, Meulders & de Boeck, 2000), and multidimensional IRT models (Béguin & Glas, 2001). The motivation for the recent interest in Bayesian inference and MCMC estimation procedures is that the complex dependency structures in the mentioned models require the evaluation of multiple integrals to solve the estimation equations in an MML or Bayes modal framework (Patz & Junker, 1999). These problems are easily avoided in an MCMC framework. Procedures for the evaluation of model fit, such as the procedures for the evaluation of person fit presented here, can be directly generalized. This point will be returned to in the discussion. In this article, several well-known person-fit statistics are generalized to the Bayesian framework. Note that Reise (2000) used empirical Bayes estimation methods in a logistic regression framework to determine the fit of an item score pattern.

This paper is organized as follows. First, we will introduce some relevant IRT models and some person-fit statistics that are often used. Second, we will discuss the principles of MCMC methods to sample the posterior distribution of a person-fit statistic. Third, we will conduct a simulation study in which we will investigate how many persons and how many items are needed in the sample to apply this method in practice. Finally, we will conduct a simulation study to determine the effectiveness of several person-fit statistics.

IRT models and Person Fit

In IRT (Rasch, 1960; Birnbaum, 1968; Mokken, 1971; Lord, 1980; Hambleton & Swaminathan, 1985; van der Linden & Hambleton, 1997) the probability of a correct response on item j ($j = 1, ..., k$), $P_j(\theta)$, is a function of the latent trait value θ and a number of item characteristics. Often used models are the one, two, and three parameter logistic (1, 2, and 3PL) models (Hambleton & Swaminathan, 1985). For example, in the 3PL model, the item is characterized by a difficulty parameter β_j, a discrimination parameter α_j and a (pseudo-)guessing probability γ_j, which is the lower asymptote of $P_j(\theta)$ when $\theta \rightarrow -\infty$. Most person-fit studies have been conducted in the context of the logistic IRT models (Meijer & Sijtsma, in press). In a Bayesian framework, however, the 3PNO model (e.g., Lord, 1980, pp. 13-14) has some computational advantages, although the 3PNO model and the 3PL model are completely equivalent for all practical purposes. In the 3PNO model, the probability of correctly answering an item is given by

$$P_j(\theta) = \gamma_j + (1 - \gamma_j)\Phi(\alpha_j\theta - \beta_j), \qquad (1)$$

where Φ denotes the standard normal cumulative distribution.

To investigate the goodness-of-fit of item score patterns, several IRT-based person-fit statistics have been proposed. Most person-fit statistics have the form

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

where Y_j is the response to item j, where the weight $v_j(\theta)$ is often defined as an increasing function of the likelihood of the observed item scores. So the test is based on the discrepancy between the observed scores Y_j and the expected scores under the model, $P_j(\theta)$. A straightforward example of a member of the class defined by (2) is the W-statistic by Wright and Stone (1979), which is defined as

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

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 statistic UB is defined as

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

Other obvious members of the class defined by (2) are two statistics proposed by Tatsuoka (1984): ζ_1 and ζ_2. The ζ_1-statistic is the standardization with a mean of 0 and unit variance of

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

where n_j denotes the number of correct answers to item j and $\overline{n}_j$ denotes the mean number of correctly answered items in the test. The index will be positive indicating misfitting response behavior 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 ways, the index will obtain a large positive value. 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. This index is sensitive to item score patterns with correct answers to difficult items and incorrect answers to easy items; 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

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

first proposed by Levine 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 was purported to be asymptotically standard normally distributed; l_z is defined as

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

where $E(l)$ and $Var(l)$ denote the expectation and the variance of l, respectively. These quantities are given by

$$E(l) = \sum_{j=1}^{k} \{P_j(\theta) \log [P_j(\theta)] + [1 - P_j(\theta)] \log [1 - P_j(\theta)]\}, \tag{9}$$

and

$$Var(l) = \sum_{j=1}^{k} P_j(\theta)\left[1 - P_j(\theta)\right]\left[\log \tfrac{P_j(\theta)}{1-P_j(\theta)}\right]^2. \tag{10}$$

It can easily be shown that $l - E(l)$ can be written in the form of Equation (2) by choosing

$$v_j(\theta) = \log\left(\frac{P_j(\theta)}{1 - P_j(\theta)}\right). \tag{11}$$

The assessment of person fit is usually contaminated with the estimation of θ. If $\hat{\theta}$ is an estimate of θ then the distributions of a person-fit statistic using $\hat{\theta}$ instead of θ will differ. For example, Molenaar and Hoijtink (1990) showed that the distribution of l_z differs substantially from the standard normal distribution for short tests. Snijders (in press) derived expressions for the first two moments of the distribution: $E\left[V(\hat{\theta})\right]$ and $Var\left[V(\hat{\theta})\right]$ and performed a simulation study for relatively small tests consisting of 8 and 15 items and for large tests consisting of 50 and 100 items, fitting the 2PL model, and estimating θ by maximum likelihood. The results showed that the approximation was satisfactory at Type I error levels of $\alpha = 0.05$ and $\alpha = 0.10$, but that the empirical Type I error was smaller than the nominal Type I error for smaller values of α. In fact, both the distribution of l_z and the version of l_z corrected for $\hat{\theta}$, denoted l_z^*, are negatively skewed (Snijders, in press; van Krimpen-Stoop & Meijer, 1999). This skewness influences the difference between nominal and empirical Type I error rates for small Type I error values. For example, Snijders (in press; see also Krimpen-Stoop and Meijer, 1999) found that for a 50-items test at $\alpha = .05$ the discrepancy between the nominal and the empirical Type I error for l_z and l_z^* at $\theta = 0$ was small (.001), whereas for $\alpha = .001$ for both statistics it was larger (approximately .005). Van Krimpen-Stoop and Meijer (1999) found that increasing the item discrimination resulted in a distribution that was more negatively skewed. An alternative may be to use a χ^2-distribution; statistical theory that incorporates the skewness of the distribution is not yet available, however, for the 2PL and 3PL models.

Examples of person-fit tests outside the class defined by (2) are the uniformly most powerful (UMP) tests by Klauer (1991, 1995; see also Levine & Drasgow, 1988). Klauer's approach entails an UMP test for testing whether a person's item score pattern complies with the Rasch (1960) model against a specific alternative model that can be viewed as a generalization of the Rasch model. The statistic that forms the basis of an UMP test is the sufficient statistic for the parameters that have to be added to the null model (in this case the Rasch model) to define the alternative model.

For example, consider the test of the Rasch model against an alternative model where the ability parameter differs between subtest A_1 and subtest A_2. Let θ_1 be the individual's ability on subtest A_1 and let θ_2 be the individual's ability on subtest A_2. Furthermore, consider the number-correct score on the first and second subtest, respectively and let $\delta = \theta_1 - \theta_2$. Then H_0: $\delta = 0$ can be tested against H_1: $\delta \neq 0$ using the number-correct score on either one of the subtests. Note that in an IRT model it is assumed that for each person the latent trait is invariant across items, if this is not the case this may point at aberrant response behavior. In contrast to, for example, calculating the log-likelihood as given in (7) or (8) we now explicitly test against an alternative hypothesis. So when the null hypothesis is rejected for a particular person this person can be classified as aberrant. We will denote the statistical test where we test if the total score on the first subtest is too high compared to what we expect based on the model as T_1, and we will denote the statistical test where we test if the test score on the second subtest is too high compared to what we expect on the basis of the model as T_2.

As another example, Klauer (1991, 1995) proposed a person-fit test for violation of the assumption of local independence using an alternative model proposed by Kelderman (1984, also see, Jannarone, 1986) where the probability of a response pattern $(y_1, ..., y_j, ..., y_k)$ is given by

$$P(y_1, ..., y_j, ..., y_k|\theta) \propto \exp\left[\sum_{j=1}^{k} y_j(\theta - \beta_j) + \sum_{j=1}^{k-1} y_j y_{j+1}\delta\right]. \qquad (12)$$

Note that δ models the dependency between y_j and y_{j+1}. If $\delta = 0$ the model equals the Rasch model. An UMP test denoted as T_{lag} of the null hypothesis $\delta = 0$ can be based on a the sufficient statistic with realizations $\sum_{j=1}^{k-1} y_j y_{j+1}$.

When the item discrimination parameters α_j are considered known, the principle of the UMP test can also be applied to the 2PL model. Analogous UMP tests for the 3PL model and the normal ogive model cannot be derived because these models have no sufficient statistic for θ. Even though UMP tests do not exist for these models, the notion of using statistics related to the parameters of an alternative model as a basis of a test is intuitively appealing. Therefore, the generalizations of these tests to the 3PNO model in a Bayesian framework will also be studied below.

Bayesian estimation of the 3PNO model

In this study, an MCMC procedure will be used to generate the posterior distributions of interest. The MCMC chains will be constructed using the Gibbs sampler (Gelfand & Smiths, 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 posterior distributions.

Albert (1992; see also Baker, 1998) applies Gibbs sampling to estimate the parameters of the well known 2PNO model (e.g., Lord & Novick, 1968). Johnson and Albert (1999, Section 6.9) generalized the procedure to the 3PNO. For application of the Gibbs sampler, it is important to create a set of partial posterior distributions that are easy to sample from. This often involves the data augmentation, that is, the introduction of additional latent variables that lead to a simple set of posterior distributions. In the Gibbs sampling algorithm, these latent variables are sampled along with the variables of interest. The present procedure is based on two data augmentation steps. The first step entails the introduction of binary variables W_{ij} such that

$$W_{ij} = \begin{cases} 1 & \text{if person } i \text{ knows the correct answer to item } j \\ 0 & \text{if person } i \text{ does not know the correct answer to item } j. \end{cases} \tag{13}$$

So if $W_{ij} = 0$, person i guessed the response to item j, if $W_{ij} = 1$, person i knows the right answer and gives a correct response. The relation between W_{ij} and the observed response variable Y_{ij} is given by a model where $\Phi(\eta_{ij})$, 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 $(1 - \Phi(\eta_{ij}))$ that the respondent does not know the item and guesses with γ_j as the probability of a correct response. So the probability of a correct response is a sum of a term $\Phi(\eta_{ij})$ and a term $\gamma_j(1 - \Phi(\eta_{ij}))$. Summing up we have

$$\begin{aligned} P(W_{ij} = 1 \,|Y_{ij} = 1, \eta_{ij}, \gamma_j) &\propto \Phi(\eta_{ij}) \\ P(W_{ij} = 0 \,|Y_{ij} = 1, \eta_{ij}, \gamma_j) &\propto \gamma_j(1 - \Phi(\eta_{ij})) \\ P(W_{ij} = 1 \,|Y_{ij} = 0, \eta_{ij}, \gamma_j) &= 0 \\ P(W_{ij} = 0 \,|Y_{ij} = 0, \eta_{ij}, \gamma_j) &= 1. \end{aligned} \tag{14}$$

The second data augmentation step is derived using a rationale which is analogous to a rationale often used as a justification of the 2PNO (see, for instance, Lord, 1980, Section 3.2). In that rationale, it is assumed that, if person i is presented item j, a latent variable Z_{ij} is drawn from a normal distribution with mean η_{ij} and a variance equal to one. A correct response $y_{ij} = 1$ is given when the drawn value is positive. Analogously, in the present case, a variable Z_{ij} is introduced with a distribution defined by

$$Z_{ij} \,|W_{ij} = w_{ij} \sim \begin{cases} N(\eta_{ij}, 1) \text{ truncated at the left by } 0 & \text{if } W_{ij} = 1 \\ N(\eta_{ij}, 1) \text{ truncated at the right by } 0 & \text{if } W_{ij} = 0. \end{cases} \tag{15}$$

The item parameters $\boldsymbol{\alpha}$ have a prior $p(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{j=1}^{k} \mathrm{I}(\alpha_j > 0)$,which insures that the discrimination parameters are positive. Note that this prior is uninformative with respect to $\boldsymbol{\beta}$. The guessing parameter γ_j has the conjugate prior Beta(a, b). The ability parameters θ have a standard normal distribution, that is, $\mu = 0$ and $\sigma = 1$.

The procedure described below is based on the Gibbs sampler. The aim of the procedure is to simulate samples from the joint posterior distribution of $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}$, $\mathbf{z}$ and $\mathbf{w}$,given the data $\mathbf{y}$, which are the responses of n test takers to k items. This posterior distribution is given by

$$\begin{aligned} p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w} \,|\mathbf{y}) &= p(\mathbf{z}, \mathbf{w} \,|\mathbf{y}\ ; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta},) p(\boldsymbol{\gamma}) p(\boldsymbol{\alpha}, \boldsymbol{\beta}) p(\boldsymbol{\theta}) \\ &= C \prod_{i=1}^{n} \prod_{j=1}^{k} p(z_{ij} \,|w_{ij}, \eta_{ij}) p(w_{ij} \,|y_{ij}, \eta_{ij}, \gamma_j) p(\gamma_j) \mathbf{I}(\alpha_j > 0) \\ &\quad \phi(\theta_i; \mu = 0, \sigma = 1) \end{aligned} \tag{16}$$

where $p(w_{ij} \,|y_{ij}, \eta_{ij}, \gamma_j)$ is given by (14) and $p(z_{ij} \,|w_{ij}, \eta_{ij})$ follows from (15).

Although the distribution given by (16) has an intractable form, as a result of the two data augmentation steps, the conditional distributions of $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}$, $\mathbf{z}$ and $\mathbf{w}$ are now each tractable and easy to sample from. A draw from the full conditional distribution can be obtained in the following steps.

Step 1 The posterior $p(\mathbf{z}, \mathbf{w} \,|\mathbf{y}\ ; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta})$ is factored as $p(\mathbf{z} \,|\mathbf{y}\ ; \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta})$ $p(\mathbf{w} \,|\mathbf{y}\ ; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta})$, and values of $\mathbf{w}$ and $\mathbf{z}$ are drawn in two substeps:

- Draw w_{ij} from the distribution of W_{ij} conditional on the data $\mathbf{y}$ and $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}$, given by (14).
- Draw z_{ij} from the conditional distribution of Z_{ij} given all other variables using (15),

Step 2 Draw from the conditional distribution of $\boldsymbol{\theta}$ given the values $\mathbf{z}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$,and $\mathbf{y}$. Since $p(\boldsymbol{\theta} \,|\mathbf{y}\ ; \mathbf{z}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta})$ is proportional to $p(\mathbf{z} \,|\boldsymbol{\theta}\ \boldsymbol{\alpha}, \boldsymbol{\beta}) p(\boldsymbol{\theta}) p(\boldsymbol{\gamma}, \mathbf{w}, \mathbf{y} \,|\mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, and the last term also does not depend on $\boldsymbol{\theta}$, it follows from the definition of Z_{ij} given above that the error term ε_{ij} in $Z_{ij} - \beta_j = \alpha_j \theta_i + \varepsilon_{ij}$ is a normally distributed. So the full-conditional distribution of $\boldsymbol{\theta}$ entails a normal model for the regression of $Z_{ij} - \beta_j$ on α_j,with θ_i as a regression coefficient which has a normal prior with parameters $\mu = 0$ and $\sigma = 1$. (see, for instance, Gelman, et al., 1995, p.45 and p.78)

Step 3 Draw from the conditional distribution of the parameters of item j, α_j, and β_j. Analogous to the previous step, also this step entails sampling from a regular normal linear model. Defining $\mathbf{Z}_j = (Z_{1j}, ..., Z_{nj})^T$, and $\mathbf{X} = (\boldsymbol{\theta}, -\mathbf{1})$, with $-\mathbf{1}$ being the n dimensional column vector with elements -1, the two items parameters can be viewed as regression coefficients in $\mathbf{Z}_j = \mathbf{X}(\alpha_j, \beta_\mathrm{j})^T + \varepsilon$, where ε is a vector of random errors. So also this step boils down to sampling the regression coefficients in a regular Bayesian linear regression problem.

Step 4 Sample from the conditional distribution of γ_j. The likelihood of $w_{1j}, ..., w_{nj}$ is a binomial with parameter γ_j. With the noninformative conjugate Beta prior introduced above, the posterior distribution of γ_j also follows a beta distribution (see, for instance, Gelman, et al., 1995, Section 2.1).

So the procedure boils down to iteratively generating a number of sequences of parameter values using these four steps. Convergence can be evaluated by comparing the between- and within-sequence variance (see, for instance, Gelman, et al., 1995). Starting points of the sequences can be provided by the Bayes modal estimates of BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). For more information on this algorithm refer to Albert (1992), Baker (1998), and Johnson and Albert (1999).

In the Bayesian approach, the posterior distribution of the parameters of the 3PNO model, say $p(\xi|y)$, is simulated using a Markov chain Monte Carlo (MCMC) method proposed by Johnson and Albert (1999). Person fit will be evaluated using a posterior predictive check based on an index $T(y, \xi)$. When the Markov chain has converged, draws from the posterior distribution can be used to generate model-conform data y^{rep} and to compute a so-called Bayes p-value defined by

$$\Pr(T(y^{rep}, \xi) \geq T(y, \xi) \mid y). \tag{17}$$

So person-fit is evaluated by computing the relative proportion of replications, that is, draws of ξ from $p(\xi|y)$, where the person-fit index computed using the data, $T(y, \xi)$, has a smaller value than the analogous index computed using data generated to conform the IRT model, that is $T(y^{rep}, \xi)$.Posterior predictive checks are performed by inserting the person-fit statistics given in the previous section into Equation (17). 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 index $T(y, \xi)$ is computed, a new model-conform response pattern is generated, and a value $T(y^{rep}, \xi)$ is computed. Finally, a Bayesian p-value is computed as the proportion of iterations were $T(y^{rep}, \xi) \geq T(y, \xi)$.

Simulation studies

This simulation study consists of two parts. In the first part we will investigate the Type I error rate as a function of test length and sample size. In the second part we will investigate the detection rates of the different statistics for different model violations, test lengths, and sample sizes. Furthermore, we will investigate the impact of nonfitting item scores on the bias in θ as a function of the number of test items affected by lack of model fit. In all three simulation studies we will use the statistics l, W, UB, ζ_1, ζ_2, T_{lag}, T_1 and T_2 as defined above.

Study 1: Type I Error Rate

Method

The simulation studies with respect to the Type I error rate were performed in two conditions: one with random and one with fixed item parameters. In both conditions, the ability parameters were drawn from a standard normal distribution. In the first condition, for every replication the item parameters were drawn from the default prior distributions used in BILOG-MG. The guessing parameter γ was drawn from a Beta(a, b) distribution with a and b equal to 5 and 17, respectively. This results in a mean γ of 0.20. Further, the item discrimination parameters were drawn from a lognormal distribution with mean zero and a variance equal to 0.5 and the item difficulty parameters β were drawn from a normal distribution, also with mean zero and variance 0.50. In the second condition, the item parameters were fixed. The γ was fixed to 0.20 for all items. Item difficulty and discrimination parameters were chosen as follows:

- 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 $\beta_i = -2.00 + 0.20(i - 1)$, $i = 1, ..., 20$.

Three samples sizes were used: $n = 100$, $n = 400$, and $n = 1000$. The true values of the parameters were used as starting values for the MCMC procedure. The procedure had a run length of 4000 iterations with a burn-in period of 1000 iterations. That is, the first 1000 iterations were discarded. In the remaining 3000 iterations, $T(y^{rep}, \xi)$ and $T(y, \xi)$ were computed every 5 iterations. So the posterior predictive checks were based on 600 draws. For the statistics that uses 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. Finally, for every condition, 100 replications were simulated and the proportion of replications with a Bayesian p-value less than 0.05 was determined.

Results

Insert Table 1 and 2 about here

The results for the condition with random item parameters are shown in Table 1; the results for the condition with fixed item parameters are shown in Table 2. It can be seen that, in general, the significance probabilities converge to their nominal value of 0.05 as a function of sample size and test length, and the nominal significance probability is best approximated by the combination of a test length $k = 60$ and a sample size $n = 400$ or $n = 1000$. Note that for $n = 100$ the significance probabilities are much too large. There are no clear effects for specific person-fit statistics, except that the UB and the T_{lag} seem to be quite conservative for $n = 400$ and $n = 1000$ and random item parameter selection. Finally, at the bottom of the two tables, the mean over replications and simulees of the absolute difference between the true and the estimated ability parameters is given. This mean absolute error (MAE) will be used to interpret the bias in the ability parameters of simulees with nonfitting response vectors in the simulation study discussed below.

Study 2: Detection rates

Method

Guessing. In several studies problems are discussed of unmotivated persons that take a test which they do not have personal interest in. For example, Schmitt, Cortina, and Whitney (1993) noted the potential for suspicion and disdain among employees in a concurrent validation study, the authors predicted that factors including poor motivation and cheating may lead to inaccurate assessment of abilities for some employees. In such testing conditions persons may guess the correct answers to groups of items, or they may produce typical item score patterns like repeated patterns of item responses. Identifying these examinees prior to item calibration, equating, and score reporting may help to improve the usefulness of results from a large scale testing program. It has been suggested that person-fit statistics may be useful to detect such behavior (Haladyna, 1994, p.165).

To evaluate the detection rate of guessing, a number of simulation studies were carried out. These studies generally had the same set-up as the Type I error rate studies (Study 1) under the condition with fixed item parameters, with the following alterations. The condition with sample size of $n = 100$ was not used because of its inflated Type I error rate. The data were generated in such a way that guessing occurred for 10% of the simulees, so data matrices with $n = 400$ simulees had 40 aberrant simulees, and data matrices with $n = 1000$ simulees had 100 aberrant simulees. For these aberrant simulees, guessing was imposed in three conditions, where $1/6$, $1/3$, or $1/2$ of the test was corrupted by guessing. So for the test with $k = 30$ items, the number of corrupted items was either 5, 10, or 15, and for the test with $k = 60$ items, the number of corrupted items was either 10, 20, or 30. Guessing was always imposed on the items with the lowest item difficulty. This was done because guessing on the easiest items has the most detrimental effect on the estimation of θ (Meijer & Nering, 1997) and thus detection of these item score patterns is important. The probability of a correct response to these items by aberrant simulees was 0.20.

Test statistics were computed in the same way as in Study 1. So, again, for statistics based on a partitioning of the test, two subtests of equal size were formed: a difficult and an easy one. As a result, the corrupted items were in the easiest test, although the partitioning did not completely conform to the pattern of corrupted and uncorrupted items. So in this sense, the partition was not optimal. However, in real situations, there is usually no prior knowledge of which items are corrupted, so the setup was considered realistic.

A final remark concerns the computation of T_1 and T_2. The latter was computed as described above, that is, its Bayesian p-value indicates how often the observed score was lower than the score replicated under the model. So a low p-value for T_2 indicates that the score on the second subtest was too high. However, in the simulation study, the item parameters were ordered from difficult to easy, and guessing was imposed on the easy items. Therefore, it is expected that the score on the easiest subtest will be too low, so for T_1 the orientation of the test is changed from right-tailed (too high scores) to left-tailed (too low scores). That is, T_1 should detect too low scores. 50 replications were simulated in every condition.

Item disclosure. In high-stakes testing, persons may be tempted to obtain knowledge about the type of test questions or even about the correct answers to the items in the test. In computerized adaptive testing this is one of the major threats to the validity of test scores. But also in standardized paper-and-pencil tests this is a realistic problem. For example, in personnel selection commercial available tests are often used by different companies. This makes the threat of item disclosure realistic due to repeated test taking. Item disclosure may result in a larger percentage of correct answers than expected on the basis of the trait that is being measured.

Note, that in general it is unknown on which 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 in particular item preknowledge on the items of median difficulty and on the most difficult items may have an effect on the total score. Thus, the effect of item preknowledge will be important in particular for persons with a low ability level that answer many difficult items correctly.

The setup of the simulation study to the detection rate of the tests for item disclosure was analogous to the study to the detection rate for guessing. So 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 prominent for 10% of the simulees, and for these simulees, $1/6$, $1/3$ or $1/2$ of the difficult items in the test were corrupted. 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, except for T_1 and T_2, which are now right-tailed as all other statistics in the study. That is, both statistics are designed to detect scores that are too high. Again, 50 replications were simulated in every condition.

Violations of local independence. When previous items provide new insights useful for answering the next item or when the process of answering items is exhausting, the assumption of local independence may be violated. This may result, for example, due to speeded testing situations or in situations were there is exposure to material among students (Yen, 1993; see also Embretson & Reise, 2000, pp. 231-233).

The setup of the simulation study to the detection rate of the tests for violation of local independence was analogous to the studies to the detection rate of guessing and item disclosure. Again, data were generated for sample sizes of $n = 400$ and $n = 1000$ simulees, and test lengths of $k = 30$ and $k = 60$ items, the model violation was imposed on 10% of the simulees, and for these simulees, $1/6$, $1/3$ or $1/2$ of the test was corrupted. Responses to corrupted items were generated with the model defined by (12), with $\delta = 1.0$. In these simulations, the items were ordered such that the affected items succeeded each other. For the condition were $1/3$ of the test was corrupted, the model violation was imposed on the items with $\alpha = 1.0$. For the condition were $1/6$ of the test was corrupted, the model violation was imposed on the items with $\alpha = 1.0$ and the lowest item difficulties. For the condition were $1/2$ of the test was corrupted, the model violation was imposed on the items with $\alpha = 1.0$ and, in the case when there were too few items with $\alpha = 1$, on the items with $\alpha = 0.5$ and the lowest item difficulties. The impact of a violation with $\delta = 1.0$ was an average increase in lag, $\sum_{j=1}^{k-1} y_j y_{j+1}$, of 1.6, 4.0 and 5.9 for a test of 60 items with 1/6, 1/3 or 1/2 of the items corrupted, respectively, and of 0.7, 1.5 and 2.4 for a test of 30 items with 1/6, 1/3 or 1/2 of the items corrupted, respectively.

Test statistics were computed in the same way as in the study of item disclosure reported above, with the exception that for T_1 the focus was on higher-than-expected outcomes. Again, 50 replications were made in every condition.

Results

Guessing. The proportions of "hits", that is, the proportion of correctly identified aberrant simulees are shown in Table 3. The proportions of "false alarms", that is, the proportion of normal simulees incorrectly identified as aberrant, are shown in Table 4.

Insert Table 3 and 4 about here

The optimal condition for the detection of guessing is a large sample size and a large test length. Therefore, the results of the condition with $n = 1000$ simulees and $k = 60$ items will be discussed first. The main overall trend for all tests is that the detection rate decreases as the number of affected items increases. This can be explained by the inflated MAE of θ for the misfitting simulees (bottom Table 1). It can be seen that the MAE for the misfitting simulees is grossly inflated, where the MAE is larger for $p = 1/2$ and $p = 1/3$ than for $p = 1/6$. Comparing these results with the results in Table 1, it can also be concluded that the presence of 10% misfitting simulees in the calibration sample affected the MAE for the fitting simulees to some degree. As the number of affected items increases, the MAE also increases, and since the fit statistics are computed conditionally on $\hat{\theta}$, the detection rate decreases. Inspection of the results in the condition with $n = 400$ simulees and $k = 60$ shows that the detection rate is little affected by the smaller calibration sample. Furthermore note that the detection rate of T_1 is lower than the detection rate of T_2. So the bias in $\hat{\theta}$ is such that the low scores on the first part of the test are less unexpected than the relatively high scores on the second part of the test.

For a test length of $k = 30$ items, the detection rate is slightly less than for $k = 60$ items. This is as expected, because the statistics are computed on an individual level and on this level the test length is the number of observations on which the test is based. Note that the relatively low detection rates of T_{lag} and T_1 found for $k = 60$ also applies for $k = 30$. Finally, at the bottom of the table it can be seen that the MAE was less inflated than for the study with $k = 60$. The explanation is that the absolute numbers of affected items that was responded to is lower here.

Item disclosure. The proportions of hits and false alarms are shown in Table 5 and 6, respectively. It can be concluded that the effects of test length and proportion of affected items are also found here. Furthermore, the absence of an effect of calibration sample size is replicated here. The detection rates of T_{lag} are relatively low.

Insert Table 5 and 6 about here

Remember that now the items in the second part of the test, that is, the easy items, were affected by the model violations. Therefore, it was expected that T_2 would be sensitive to the increase in the total score on the second half of the test. Table 5 shows that this expectation was confirmed (e.g., detection rates between 0.24 and 0.50 for $n = 1000$ and $k = 30$). However, note that for $k = 30$ the detection rate of T_1 was also relatively large (between 0.27 and 0.30) with, contrary to T_2, a high false alarm rate (between 0.27 and 0.29). Thus, the bias of the ability estimate caused by the model violation was large enough to affect the simulation of the predictive distribution of T_1. In practice this is undesirable, because one does not know a priori which part of the test is affected and the interpretation of the outcome of T_1 and T_2 is problematic.

Violations of local independence. The proportions of hits and false alarms are shown in Table 7 and 8, respectively.

Insert Table 7 and 8 about here

Although for the aberrant simulees the increase in lag reported above was considerable, in the two bottom lines of Table 7 it can be seen that the resulting bias in their ability estimates was not impressive. Also the detection rate of the tests was negligible. Even the power of the T_{lag}-test, which is specially targeted at this model violation was negligible. The only exception was the T_1-test. The reason is that the affected items were placed at the beginning of the test, and the increase in lag also resulted in an increase in the total score on the first part of the test. Note, however, that in Table 8 it can be seen that the false alarm rate of this test also increased.

Discussion

Aberrant response behavior in psychological and educational testing may result in inadequate measurement of some persons. Therefore, misfitting item scores should be detected and removed from the sample. To classify an item score pattern as nonfitting, the researcher can simply take the top 1 or top 5 percent of aberrant cases or he/she use a theoretical sampling distribution or can simulate datasets based on the estimated item parameters in the sample. In the first case person-fit statistics are used as descriptive statistics. In this study, we followed the approach in which we used person-fit statistics to test the hypothesis that an item score pattern is not in agreement with the underlying test model. Simulation methods thus far applied in the literature did not take into account the uncertainty of parameters of the IRT model. In this study, we used Bayesian methods that take into account this uncertainty to classify an item score pattern as fitting or nonfitting. Although Bayesian methods are statistically superior to other simulation methods, a drawback is that they are relatively complex and computational intensive.

Depending on the type of data and the problems envisaged, a researcher may choose a particular person-fit statistic, although not all statistics have equally favorable properties in a statistical sense. In general, sound statistical methods have been derived for the Rasch model, but because this model is rather restrictive to empirical data, the use of these statistics is also restricted. For the 2PL model and the 3PL model and for short tests and tests of moderate length (say, 10-60 items) due to the use of $\hat{\theta}$ rather than θ, for most statistics the nominal Type I error rate under the standard normal distribution is not in agreement with the empirical Type I error rate (van Krimpen-Stoop & Meijer, 1998). As an alternative one may use the correction proposed by Snijders (in press) or one may use Bayesian simulation procedures discussed in this paper.

From the results it can be concluded that even for a test as short as 30 items and for 400 simulees the type I error is well under control (approximately 0.03 at an nominal for most statistics studied). In particular it is interesting to compare these results with the results obtained using the theoretical distribution. For example, Snijders (in press) found using the 2PL model and the log-likelihood statistic corrected for $\hat{\theta}$ in a simulation study with 100,000 replications for nominal type I error rates $\alpha = .05$ (resulting in standard errors between 0.001 and 0.005) for a 15-items test empirical type I errors between 0.053 and 0.061. Note, however, that he considered the item parameters known. This is only realistic when the item parameters can be estimated very accurately, that is, for very large sample sizes. For small sample sizes the method proposed in this study may be more suitable.

Detection rates differed for different statistics and different types of model violations simulated. In general, it can be concluded that the detection rates for guessing and item disclosure were higher than for violations against local independence. Note, however, that also the MAE was relatively small in the latter case, in contrast to the MAE for the guessing condition. Also for item disclosure, the MAE was often slightly larger for misfitting score patterns compared to the MAE for fitting score patterns, although the power of some person-fit statistics was high (Table 5). Aggregated over all conditions, the ζ_2-test had the highest power. The expectation that the UMP tests for person fit in the Rasch model (T_1, T_2, and T_{lag}) may also be superior in the framework of a 3PL model in an Bayesian framework was not corroborated. Traditional discrepancy tests do better here. Not reported above, but also included in the study were versions of T_1, T_2, and T_{lag} where the item scores were weighted by the discrimination parameters. The detection rates of these tests were consistently lower than those of the tests based on the unweighted scores. Interesting was that the detection rates decreased when the number of items affected by guessing increased. This is contrary to findings in earlier studies (Meijer & Sijtsma, 2001). This may be explained as follows. In the present study a larger amount of guessing resulted in a lower $\hat{\theta}$ than the original $\hat{\theta}$. As a result of using this lower $\hat{\theta}$, item score patterns are less aberrant than using the original $\hat{\theta}$. In other studies, the $\hat{\theta}$ is often fixed, and as a result item score patterns are more often classified as misfitting.

A final remark concerns the generalization of the procedure presented here to a general IRT framework incorporating models with multiple raters, testlet structures, latent classes, and multi-level structures (references given above). The common theme in these models is their complex dependency structure and the fact that these complex models can be estimated using the Gibbs sampler. In all cases, the structure of the estimation procedure is analogous: draws from the posterior distribution are made by partitioning the complete parameter vector into a number of components and sampling each component conditionally on the draws for the other component. Usually, the partition of the complete parameter vector is in the item parameters, the person parameters, augmented data (such as Z and W above) and hyperparameters which may be related to some restrictions on the parameters (as in testlet and other multilevel IRT models) and some of the priors. In all these models, the statistics described above can be computed given the current draw of the item and person parameters, both for the observed data y and replicated data y^{rep} drawn from the posterior predictive distribution.

References

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

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

Baker, F.B. (1998). An investigation of item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement 22,* 153-169.

Béguin, A.A. and Glas, C.A.W. (2001). MCMC estimation of multidimensional IRT models. *Psychometrika, 66, December issue.*

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

Bradlow, E.T., Wainer, H., and Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153-168.

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.

Fox, J.P. & Glas, C.A.W. (2001). Bayesian Estimation of a Multilevel IRT Model using Gibbs Sampling. *Psychometrika 66, 271-288.*

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

Gelman, A, Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis.* London: Chapman and Hall.

Gelman, A., Meng, X-L., & Stern, H.S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica 6,* 733-807.

Glas, C.A.W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53,* 525-546.

Glas, C. A. W. (1999). Modification indices for the 2-pl and the nominal response model. *Psychometrika, 64,* 273-294.

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

Hoijtink, H. and Molenaar, I.W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs Sampler and posterior predictive checks. *Psychometrika, 62,* 171-189.

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

Janssen, R., Tuerlinckx, F., Meulders, M. & de Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal od Educational and Behavioral Statistics, 25,* 285-306.

Johnson, V.E., & Albert, J.H. (1999) *Ordinal data modeling.* New York, NJ: Springer.

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

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

Klauer, K. C. (1995). The assessment of person fit. In G. H. Fischer &. I. W. Molenaar (Eds.), *Rasch models, foundations, recent developments, and applications* (pp. 97-110). New York, NJ: Springer-Verlag.

Kim, S.-H. (2001). An evaluation of a Markov Chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement, 25,* 163-176.

Levine, M.V., & Drasgow, F (1988). Optimal appropriateness measurement. *Psychometrika, 53,* 161-176.

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

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

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

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.

Meng, X.L. (1994). Posterior predictive p-values. *Ann. Statist., 22,* 1142-1160.

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

Molenaar, I.W. (1983). Some improved diagnostics for failure in the Rasch model. *Psychometrika, 48,* 49-72.

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

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50-64.

Patz, R.J., and Junker, B.W. (1997). *Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses.* (Technical Report No.670). Pittsburgh: Carnegie Mellon University, Department of Statistics.

Patz, R.J., and Junker, B.W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response theory models. *Journal of Educational and Behavioral Statistics, 24,* 146-178.

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

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

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research, 35,* 543-570.

Reise, S. P. & Widaman, K.F. (1999). Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure models. *Psychological methods*, 4, 3-21.

Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and psychological measurement*, *45*, 433-444.

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

Snijders, T. (1998). *Asymptotic Distribution of Person-Fit Statistics with Estimated Person Parameter*. Accepted by *Psychometrika.*

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

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.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2000). Detecting person-misfit in adaptive testing using statistical process control techniques. In W.J. van der Linden &, C.A.W. Glas (Eds.), *Computerized Adaptive Testing: Theory and practice* (pp.201-220). Boston, MA: Kluwer Academic Publishers.

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

Wainer, H., Bradlow, E.T., and Du, Z. (1999). Testlet Response Theory: an Analogue for the 3-PL Useful in Testlet-Based Adaptive Testing. In W.J. van der Linden and C.A.W.Glas (Eds.). *Computer Adaptive Testing: Theory and Practice.* Boston: Kluwer-Nijhoff Publishing. To Appear.

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

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

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125-145.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Mesurement, 30,* 187-213.

Zimowski, M.F., Muraki, E., Mislevy, R.J., and Bock, R.D. (1996). *Bilog MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items.* Chicago: Scientific Software International, Inc.

Table 1
Actual Type I Error Rates for a Nominal $\alpha = .05$ Test
Random Item Parameters

	$k = 30$			$k = 60$		
	$n = 100$	$n = 400$	$n = 1000$	$n = 100$	$n = 400$	$n = 1000$
l	0.10	0.03	0.03	0.19	0.03	0.03
W	0.10	0.03	0.03	0.20	0.03	0.03
UB	0.07	0.01	0.02	0.16	0.02	0.01
ζ_1	0.12	0.04	0.04	0.20	0.04	0.04
ζ_2	0.19	0.07	0.06	0.26	0.06	0.06
T_{lag}	0.16	0.03	0.02	0.22	0.01	0.01
T_1	0.19	0.08	0.06	0.24	0.04	0.04
T_2	0.18	0.08	0.06	0.25	0.05	0.05
MAE	0.33	0.28	0.28	0.33	0.22	0.20

Table 2
Actual Type I Error Rates for a Nominal $\alpha = .05$ Test
Fixed Item Parameters

	$k = 30$			$k = 60$		
	$n = 100$	$n = 400$	$n = 1000$	$n = 100$	$n = 400$	$n = 1000$
l	0.10	0.02	0.02	0.18	0.04	0.04
W	0.10	0.02	0.02	0.19	0.04	0.04
UB	0.08	0.00	0.01	0.15	0.03	0.04
ζ_1	0.14	0.04	0.05	0.20	0.04	0.04
ζ_2	0.21	0.09	0.06	0.26	0.05	0.05
T_{lag}	0.22	0.07	0.06	0.24	0.03	0.03
T_1	0.24	0.04	0.04	0.25	0.03	0.04
T_2	0.24	0.05	0.04	0.25	0.05	0.05
MAE	0.46	0.30	0.30	0.32	0.23	0.22

Table 3
Detection Rate for Guessing Simulees

	$k = 30$						$k = 60$					
	$n = 400$			$n = 1000$			$n = 400$			$n = 1000$		
	$p = 1/6$	$p = 1/3$	$p = 1/2$	$p = 1/6$	$p = 1/3$	$p = 1/2$	$p = 1/6$	$p = 1/3$	$p = 1/2$	$p = 1/6$	$p = 1/3$	$p = 1/2$
l	.47	.25	.19	.50	.23	.18	.57	.33	.26	.57	.32	.24
W	.45	.26	.20	.47	.24	.19	.57	.34	.26	.56	.33	.25
UB	.49	.24	.16	.52	.22	.15	.52	.34	.20	.61	.33	.23
ζ_1	.33	.22	.30	.36	.20	.27	.35	.25	.32	.37	.23	.30
ζ_2	.52	.33	.25	.55	.32	.25	.63	.38	.29	.62	.37	.29
T_{lag}	.09	.39	.25	.07	.38	.24	.23	.49	.34	.16	.47	.33
T_1	.16	.02	.04	.18	.02	.04	.23	.03	.07	.19	.02	.04
T_2	.23	.44	.40	.23	.43	.39	.34	.48	.46	.38	.47	.45
MAE normal	.31	.36	.36	.30	.36	.36	.28	.34	.35	.25	.33	.33
MAE abberant	.75	1.77	1.82	.73	1.77	1.82	1.00	1.88	1.90	1.01	1.95	1.95

Table 4
False Alarm Rate for Guessing Simulees

	$k = 30$						$k = 60$					
	$n = 400$			$n = 1000$			$n = 400$			$n = 1000$		
	$p = 1/6$	$p = 1/3$	$p = 1/2$	$p = 1/6$	$p = 1/3$	$p = 1/2$	$p = 1/6$	$p = 1/3$	$p = 1/2$	$p = 1/6$	$p = 1/3$	$p = 1/2$
l	.01	.02	.02	.01	.01	.01	.01	.02	.02	.01	.01	.02
W	.01	.02	.02	.01	.01	.01	.01	.02	.02	.01	.01	.01
UB	.01	.02	.01	.01	.01	.01	.02	.03	.01	.02	.03	.01
ζ_1	.03	.05	.04	.02	.03	.03	.02	.03	.03	.02	.02	.02
ζ_2	.06	.08	.06	.04	.05	.05	.05	.06	.05	.04	.05	.05
T_{lag}	.03	.05	.03	.01	.01	.01	.02	.01	.01	.00	.00	.00
T_1	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01	.00	.00
T_2	.04	.08	.05	.03	.04	.03	.03	.03	.03	.02	.02	.02

Table 5
Detection Rate for Item Disclosure

	$k=30$						$k=60$					
	$n=400$			$n=1000$			$n=400$			$n=1000$		
	$p=1/6$	$p=1/3$	$p=1/2$	$p=1/6$	$p=1/3$	$p=1/2$	$p=1/6$	$p=1/3$	$p=1/2$	$p=1/6$	$p=1/3$	$p=1/2$
l	.13	.32	.38	.13	.34	.38	.28	.58	.54	.26	.54	.51
W	.15	.36	.40	.15	.38	.39	.33	.62	.54	.32	.58	.51
UB	.10	.25	.35	.10	.28	.35	.21	.49	.51	.23	.41	.58
ζ_1	.25	.55	.59	.25	.54	.58	.52	.76	.77	.50	.74	.74
ζ_2	.32	.64	.68	.32	.63	.67	.61	.84	.87	.56	.84	.85
T_{lag}	.06	.15	.17	.06	.10	.10	.08	.29	.22	.07	.22	.15
T_1	.30	.31	.29	.30	.29	.27	.01	.03	.14	.01	.03	.19
T_2	.24	.47	.53	.24	.43	.50	.43	.65	.63	.38	.60	.61
MAE normal	.29	.29	.30	.29	.29	.29	.23	.23	.24	.22	.22	.22
MAE abberant	.34	.48	.62	.34	.47	.64	.26	.39	.57	.26	.40	.64

Table 6
False Alarm Rate for Item Disclosure

	$k=30$						$k=60$					
	$n=400$			$n=1000$			$n=400$			$n=1000$		
	$p=1/6$	$p=1/3$	$p=1/2$	$p=1/6$	$p=1/3$	$p=1/2$	$p=1/6$	$p=1/3$	$p=1/2$	$p=1/6$	$p=1/3$	$p=1/2$
l	.02	.02	.02	.02	.02	.01	.02	.01	.01	.02	.01	.01
W	.02	.02	.01	.02	.01	.01	.02	.01	.01	.02	.01	.01
UB	.02	.02	.01	.02	.02	.01	.02	.01	.01	.01	.01	.02
ζ_1	.03	.02	.02	.02	.02	.02	.02	.01	.01	.01	.01	.01
ζ_2	.05	.04	.04	.04	.04	.03	.04	.03	.03	.04	.03	.02
T_{lag}	.02	.01	.02	.01	.01	.01	.01	.00	.00	.00	.00	.00
T_1	.29	.28	.29	.29	.27	.27	.00	.00	.09	.00	.00	.10
T_2	.04	.03	.03	.03	.02	.02	.02	.01	.01	.02	.01	.01

Table 7
Detection Rate for Violation of Local Independence

	$k = 30$						$k = 60$					
	$n = 400$			$n = 1000$			$n = 400$			$n = 1000$		
	$p = 1/6$	$p = 1/3$	$p = 1/2$	$p = 1/6$	$p = 1/3$	$p = 1/2$	$p = 1/6$	$p = 1/3$	$p = 1/2$	$p = 1/6$	$p = 1/3$	$p = 1/2$
l	.02	.02	.01	.01	.02	.01	.02	.03	.01	.02	.03	.01
W	.02	.02	.01	.01	.02	.01	.02	.03	.01	.02	.03	.01
UB	.02	.03	.01	.01	.02	.01	.02	.03	.01	.02	.02	.02
ζ_1	.02	.03	.02	.01	.02	.01	.02	.04	.02	.02	.04	.01
ζ_2	.04	.06	.05	.03	.04	.03	.05	.07	.04	.03	.06	.04
T_{lag}	.02	.04	.05	.01	.01	.02	.01	.02	.03	.00	.01	.01
T_1	.14	.33	.37	.13	.32	.37	.07	.27	.35	.07	.27	.34
T_2	.02	.02	.02	.01	.00	.01	.03	.01	.01	.01	.01	.01
MAE normal	.30	.30	.29	.29	.28	.29	.23	.23	.23	.22	.22	.22
MAE abberant	.33	.33	.35	.31	.33	.34	.25	.26	.29	.23	.26	.30

Table 8
False Alarm Rate for Violation of Local Independence

	$k = 30$						$k = 60$					
	$n = 400$			$n = 1000$			$n = 400$			$n = 1000$		
	$p = 1/6$	$p = 1/3$	$p = 1/2$	$p = 1/6$	$p = 1/3$	$p = 1/2$	$p = 1/6$	$p = 1/3$	$p = 1/2$	$p = 1/6$	$p = 1/3$	$p = 1/2$
l	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03
W	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03
UB	.03	.03	.03	.03	.03	.03	.01	.01	.01	.01	.02	.02
ζ_1	.04	.04	.04	.03	.03	.03	.03	.03	.03	.03	.03	.03
ζ_2	.06	.07	.07	.06	.06	.06	.06	.06	.06	.06	.05	.06
T_{lag}	.02	.02	.02	.01	.01	.01	.01	.01	.01	.00	.00	.00
T_1	.10	.10	.09	.09	.09	.08	.06	.05	.04	.05	.04	.04
T_2	.06	.07	.06	.04	.05	.05	.03	.04	.04	.03	.03	.03

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

RR-01-09 C.A.W. Glas & R.R. Meijer, *A Bayesian Approach to Person Fit Analysis in Item Response Theory Models*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

...

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

faculty of

EDUCATIONAL SCIENCE AND TECHNOLOGY

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