Person fit in order-restricted latent class models

Emons, W.H.M.; Glas, C.A.W.; Meijer, R.R.; Sijtsma, K.

Published in:
Applied Psychological Measurement

Publication date:
2003

Document Version
Publisher's PDF, also known as Version of record

Person Fit in Order-Restricted Latent Class Models

Wilco H. M. Emons, Cees A. W. Glas, Rob R. Meijer and Klaas Sijtsma

Applied Psychological Measurement 2003; 27; 459
DOI: 10.1177/0146621603259270

Person Fit in Order-Restricted Latent Class Models

Wilco H. M. Emons,[1] Cees A. W. Glas,[2] Rob R. Meijer,[2] and Klaas Sijtsma[1]
[1]Tilburg University, The Netherlands, and [2]University of Twente, The Netherlands

Person-fit analysis revolves around fitting an item response theory (IRT) model to respondents' vectors of item scores on a test and drawing statistical inferences about fit or misfit of these vectors. Four person-fit measures were studied in order-restricted latent class models (OR-LCMs). To decide whether the OR-LCM fits an item score vector, a Bayesian framework was adopted and posterior predictive checks were used. First, simulated Type I error rates and detection rates were investigated for the four person-fit measures under varying test and item characteristics. Second, the suitability of the OR-LCM methodology in a nonparametric IRT context was investigated. The result was Type I error rates close to the nominal Type I error rates and detection rates close to the detection rates found in OR-LCMs. This means that the OR-LCM methodology is a suitable alternative for assessing person fit in nonparametric IRT models. *Index terms: Bayesian approach to person fit, nonparametric item response theory, order-restricted latent class analysis, person-fit analysis, person-fit statistics, posterior predictive checks*

Introduction

Person-fit analysis revolves around fitting an item response theory (IRT) model to respondents' vectors of item scores on a test and drawing statistical inferences about fit or misfit of these vectors (Drasgow, Levine, & Williams, 1985; Klauer, 1995; Levine & Rubin, 1979; Meijer & Sijtsma, 2001; Reise, 2000). Causes of misfit may be guessing for the correct answers, cheating by copying from another testee's answer sheet, test anxiety resulting in many errors on the first items of the test, lack of concentration toward the end of the test, and nonmastery of particular subabilities (see Haladyna, 1994, pp. 163-167; Meijer, 1994a; Meijer & Sijtsma, 1995, 2001). A misfitting item score vector may provide evidence of a biased and an unduly inaccurate test score estimate (e.g., Birenbaum & Nassar, 1994; Meijer, 1997, 1998; Meijer & Nering, 1997). To obtain a more valid estimate of test performance, respondents having misfitting item score vectors may be reassessed by means of another test. In the context of education, person misfit may lead to the decision of remedial teaching of certain abilities and skills so that a more valid test performance results. At the test administration level, results from person-fit analysis may help to improve test conditions. For example, the test instruction may be improved, and more practice items may be presented so as to prevent confusion resulting in odd answers when filling out the test form. At the data analysis level, misfitting item score vectors may be considered to be outliers (Bradlow & Weiss, 2001;

Applied Psychological Measurement, Vol. 27 No. 6, November 2003, 459–478
DOI: 10.1177/0146621603259270

Meijer, 2002). A data analysis may compare the results obtained from the complete data, including the outliers and the data without the outliers.

IRT models are parametric when the regression of the item score on the latent trait is defined by a parametric function, such as the logistic or the normal ogive (Boomsma, van Duijn, & Snijders, 2001; Van der Linden & Hambleton, 1997), and nonparametric when the regression is subjected to order restrictions only (Junker, 1993; Mokken & Lewis, 1982; Ramsay, 1991; Sijtsma & Molenaar, 2002; Stout, 1990). Parametric IRT models are special cases of nonparametric IRT (NIRT) models (Sijtsma & Hemker, 2000). Parametric models have the advantage that the sampling distributions of person-fit statistics often are known (e.g., Klauer, 1995; Molenaar & Hoijtink, 1990; Snijders, 2001). A disadvantage is that these models may be too critical, resulting in many misfitting item score vectors that may be fit by less restrictive IRT models. Being more flexible, NIRT models may be adequate candidates. Their disadvantage is that the sampling distributions of person-fit statistics are unknown (Meijer & Sijtsma, 2001), derived under unrealistic assumptions (Emons, Meijer, & Sijtsma, 2002), or conservative (Emons, 2003b; Sijtsma & Meijer, 2001). Order-restricted latent class models (OR-LCMs) (Croon, 1991; Heinen, 1996; Hoijtink & Molenaar, 1997) share the flexibility with NIRT models and the possibility in establishing sampling distributions of person-fit statistics with parametric IRT models.

In this study, OR-LCMs were fit to respondents' item score vectors, and Bayesian posterior predictive checks (PPCs) were adopted (Berkhof, van Mechelen, & Hoijtink, 2001; Gelman, Carlin, Stern, & Rubin, 1995; Glas & Meijer, in press) to test for misfit. The application of latent class models (LCMs) (Heinen, 1996) to person fit was pursued earlier by Van den Wittenboer, Hox, and De Leeuw (2000) in another context. Following Hoijtink and Molenaar (1997), Vermunt (2001), and Van Onna (2002), the OR-LCM was used to approximate NIRT models. OR-LCMs are mathematically almost identical to NIRT models, but unlike NIRT models, they assume a discrete latent trait. Discreteness of the latent trait fits in with the actual practice in IRT of estimating only a limited number of values of the continuous latent trait. One reason is that continuity cannot be maintained in principle due to finite sample size and sometimes, in addition, model structure (as in the Rasch [1960] model, in which the number of correct answers is a sufficient statistic for the latent trait). Another reason is that for practical test applications, only a limited number of estimated latent trait values are needed. Straightforward examples are mastery testing that uses only the classification of masters and nonmasters and grading that uses five ordered classes (A, B, C, D, and F). Because of the resemblance of OR-LCMs and NIRT models, and because OR-LCMs allow for the calculation of PPCs, the applicability of OR-LCMs to person fit in an NIRT context was investigated.

Sijtsma and Molenaar (2002, pp. 149-150) list many applications of NIRT models to data collected in psychological, sociological, marketing, and social medical research. So far, OR-LCMs have been studied as discrete approximations of NIRT models at the theoretical level. Their practical usefulness has only been shown in data from a general child intelligence test (Emons, 2003a). Here, the OR-LCM approach was successful in identifying some interesting cases of person misfit. The present study is a first contribution to the demonstration of the usefulness of OR-LCMs in person-fit analysis.

First, NIRT models and OR-LCMs are discussed. Second, four well-known person-fit statistics are redefined in the context of the OR-LCM: the normed number of Guttman errors (Meijer, 1994b), Van der Flier's (1980) $U3$ statistic, the log-likelihood (Levine & Rubin, 1979) of an item score vector, and Tatsuoka's (1984) ζ_1 statistic. Third, Bayesian estimation and the calculation of PPCs are discussed. Fourth, a simulation study was used to investigate the degree to which the LCM-adapted person-fit measures detected misfit of item score vectors using PPCs. In particular, for each of the four person-fit measures, simulated and nominal Type I error rates were compared and detection rates determined. One of the design factors was the distribution of the latent trait.

Data were simulated under (a) a discrete latent trait distribution of five latent classes, typical of OR-LCM analysis, and (b) a continuous normal distribution typical of NIRT (and IRT in general). This distinction enables the comparison of OR-LCM analysis under typical LCM assumptions and typical NIRT assumptions about the latent trait.

2. Item Response Models and Latent Class Models

Let a test consist of J items, and let $X_j (j = 1, \ldots, J)$ be the dichotomous item score random variable with $X_j = 1$ for a correct or coded response and 0 otherwise. Furthermore, let $(\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_J}))$ be the random vector of the item score variables with realizations $\mathbf{x} = (\mathbf{x_1}, \ldots, \mathbf{x_J})$, and let $(X_+ = \sum_1^J X_j)$ denote the unweighted sum score. Finally, let θ denote the latent trait.

NIRT models. The first assumption of the MH model is unidimensionality (UD). UD means that the latent trait θ is a scalar. Let the conditional probability of $X_j = 1$ be denoted by $(P_j(\theta))$. This conditional probability is known as the item response function (IRF). The assumption of local independence (LI) means that the item scores are statistically independent conditional on θ; that is,

$$P(\mathbf{X} = \mathbf{x}|\theta) = \prod_{j=1}^{J} P_j(\theta)^{x_j}[1 - P_j(\theta)]^{1-x_j}. \quad (1)$$

Furthermore, the monotonicity (M) assumption states that $P_j(\theta)$ is nondecreasing in θ; that is, for two arbitrary fixed values θ_a and θ_b,

$$P_j(\theta_a) \leq P_j(\theta_b), \text{ whenever } \theta_a < \theta_b. \quad (2)$$

Assumptions UD, LI, and M together define the MH model (Mokken, 1971, p. 117; Sijtsma & Molenaar, 2002, pp. 22-23).

In addition to UD, LI, and M, it may be assumed that the IRFs do not intersect. This means that the item ordering is the same for every individual, with the possible exception of ties for some θs (Sijtsma & Junker, 1996). Formally, for two items j and i and a fixed value θ_0, if $P_j(\theta_0) > P_i(\theta_0)$, then

$$P_j(\theta) \geq P_i(\theta), \text{ for all } \theta. \quad (3)$$

This is the assumption of invariant item ordering (IIO) (Sijtsma & Junker, 1996), which is identical to Rosenbaum's (1987) concept of item i being uniformly more difficult than item j. An IIO is relevant when person-fit assessment is based on the assumption that the items have the same difficulty ordering for each θ (see, e.g., Sijtsma & Meijer, 2001). The model defined by the assumptions of UD, LI, M, and nonintersecting IRFs is Mokken's double monotonicity (DM) model (Mokken, 1971, p. 118; Sijtsma & Molenaar, 2002, pp. 23-25).

OR-LCMs. In OR-LCMs, Q latent classes are assumed, each with weight ω_q $(q = 1, \ldots, Q)$. Each class corresponds to a point on the latent continuum θ. This means that the latent classes can be ordered such that the first latent class represents the lowest latent trait level and the Qth latent class the highest latent trait level. For OR-LCMs, the assumptions of LI and M are adapted as follows. Let the conditional response probability of $X_j = 1$ within class q be denoted by π_{jq}, with $j = 1, \ldots, J$ and $q = 1, \ldots, Q$. Within each class q, the item scores are independent; that is, LI is adapted to

$$P(\mathbf{X} = \mathbf{x}|q) = \prod_{j=1}^{J} \pi_{jq}^{x_j} (1 - \pi_{jq})^{1-x_j}. \quad (4)$$

Assumption M states that the class-specific probabilities π_{jq} are nondecreasing in the latent class number; that is,

$$\pi_{j1} \leq \pi_{j2} \leq \cdots \leq \pi_{jQ}\ ,\ j = 1, \ldots, J. \tag{5}$$

The LCM defined by UD, LI, and M is a discrete version of the MH model. OR-LCMs that assume an IIO can be defined by restrictions on the item parameters, such that for items j and i and a latent class q_0, if it is known that $\pi_{jq_0} > \pi_{iq_0}$, then

$$\pi_{jq} \geq \pi_{iq}, \text{ for all } q. \tag{6}$$

OR-LCMs and NIRT models postulate flexible models that can be used as the starting point for an IRT analysis and to get insight into the peculiarities of the data (Junker & Sijtsma, 2001b).

3. Person-Fit Measures

Person-fit measures compare a person's observed item score vector with the expected item score vector (Drasgow, Levine, & McLaughlin, 1987; Meijer & Sijtsma, 2001; Schmitt, Chan, Sacco, McFarland, & Jennings, 1999). There are group-based statistics and IRT-based statistics. Group-based statistics use the expected item score vector on the basis of observed data, whereas IRT-based statistics use the expected item score vector on the basis of an IRT model. Here, two group-based person-fit statistics, an IRT-based person-fit statistic, and a statistic that combines information from the IRT model and the observed data were redefined for OR-LCMs. In earlier research (Glas & Meijer, 2003; Meijer, 1994b), these statistics were found to be relatively powerful. Also, they are much used in practical person-fit analysis (Birenbaum, 1986; Levine & Rubin, 1979; Meijer & Sijtsma, 2001; Nering, 1995, 1997; Reise & Waller, 1993; Zickar & Drasgow, 1996).

3.1 Normed Number of Guttman Errors

The first person-fit measure compares an observed item score vector with the expected item score vector under the deterministic Guttman (1950) model. Large deviations of the observed item score vector from the expected item score vector indicate misfit. For continuous θ and item location parameter δ_j, the expected item scores under the Guttman model are defined by

$$\theta < \delta_j \leftrightarrow X_j = 0, \tag{7}$$

and

$$\theta \geq \delta_j \leftrightarrow X_j = 1, \tag{8}$$

for all j. For OR-LCMs, the analog of the Guttman model, which is now defined for a discrete latent trait, can be specified as follows. For each item j and each arbitrary latent class, say q_0, the latent class Guttman model can be defined by the following pair of equations:

$$\text{if } \pi_{jq_0} = 0, \text{ then } X_j = 0 \text{ for all } q \leq q_0, \tag{9}$$

and

$$\text{if } \pi_{jq_0} = 1, \text{ then } X_j = 1 \text{ for all } q \geq q_0. \tag{10}$$

Note that the Guttman model and the latent class Guttman model exclude pairs of item scores for which the easier item was answered incorrectly and the more difficult item correctly. Such item score pairs are called Guttman errors.

The item difficulty ordering needed to calculate the number of Guttman errors in the data is based on the proportions of respondents who answered the item correctly, and this ordering is used for each respondent. For OR-LCMs defined by UD, LI, and M, the item difficulty ordering may vary over latent classes. Therefore, the response probabilities π_{jq} ($j = 1, \ldots, J$), were used, for the item difficulty ordering in each class separately. This means that for the OR-LCM, the number of Guttman errors was calculated given the item difficulty ordering in the class to which the respondent belongs.

Let in each class q ($q = 1, \ldots, Q$) the items be ordered by decreasing π_{jq}, and let $r_j(q)$ denote the rank number of item j in class q. For example, $r_5(1) = 3$ means that item 5 has rank number 3 in class 1. For a fixed respondent belonging to class q and observed item score pattern $\mathbf{x}$ containing x_+ correct answers, the number of Guttman errors equals

$$G(q) = \sum_{j=1}^{J} r_j(q)x_j - \sum_{j=1}^{x_+} j. \tag{11}$$

As the maximum value of $G(q)$ varies with x_+, measure $G(q)$ was normed by the maximum number of Guttman errors that is possible given J and x_+ to be able to compare $G(q)$ for patterns with different number-correct scores (also see Meijer, 1994b). Norming resulted in

$$G^*(q) = \frac{\sum_{j=1}^{J} r_j(q)x_j - \sum_{j=1}^{x_+} j}{x_+(J - x_+)}. \tag{12}$$

The minimum value of $G^*(q)$ equals 0 and is obtained if all correct answers were given to the x_+ easiest items. Such a pattern is called a Guttman vector because it is predicted by the latent class Guttman model (equations (9) and (10)). The maximum value of $G^*(q)$ equals 1 and is obtained if all correct answers were given to the x_+ most difficult items. Such a pattern is called a reversed Guttman pattern.

Three remarks are in order. First, the dependence of measure $G^*(q)$ on a latent parameter q fits into the Bayesian framework (to be discussed shortly), where model fit can be investigated using measures that are functions of both parameters and data (Gelman et al., 1995, p. 169). Second, measure $G^*(q)$ is not defined when $X_+ = 0$ or $X_+ = J$. In practice, this is not a problem when such extreme scores are rare. Third, Guttman errors are permitted to some degree under probabilistic OR-LCMs and NIRT models. However, a high degree of dissimilarity between the observed item score vector and the Guttman pattern is an indication of misfit.

3.2 Van der Flier's *U3*

The second person-fit measure that was adapted to OR-LCMs was Van der Flier's (1980) $U3$ statistic. Let s_j be the sample fraction of respondents that answered item j correctly, and assume a numbering of items corresponding with decreasing s_j; that is, item 1 is the easiest item and so on. Furthermore, in item score vector $\mathbf{x} = (x_1, \ldots, x_J)$, score x_1 is the score on the easiest item 1 and so on. Then, $U3$ is defined as

$$U3 = \frac{\sum_{j}^{x_+} \text{logit}(s_j) - \sum_{j}^{J} x_j \text{logit}(s_j)}{\sum_{j=1}^{x_+} \text{logit}(s_j) - \sum_{j=J-x_++1}^{J} \text{logit}(s_j)}. \tag{13}$$

Measure $U3$ is closely related to $G(q)$ (Van der Flier, 1980). Van der Flier (1980) derived expressions for $E(U3)$ and $Var(U3)$ given x_+ and proposed a standardized version of $U3$, called $ZU3$, for

which the null distribution was derived to be asymptotically standard normal. This derivation used the assumption of statistical independence of item scores. Emons et al. (2002) showed that in realistic test situations, the empirical sampling distributions differed from the standard normal distribution.

To assess person fit in OR-LCMs, $U3$ was implemented as follows. The response probabilities π_{jq} $(j = 1, \ldots, J)$ were used for establishing the item difficulty ordering within latent classes q. Let $I[r_j(q) \leq x_+]$ be the indicator function with value 1 if the rank of item j in class q, $r_j(q)$, is lower than or equal to x_+ and 0 otherwise. Consider that a respondent who belongs to class q has a response vector $\mathbf{x}$ containing x_+ correct answers. Substituting s_j by π_{jq} in equation (13) and using the class-specific rank ordering of the items in class q defined by $r_j(q)$, person-fit measure $U3(q)$, which is now a function of q, is defined as

$$U3(q) = \frac{U3(q)_{\min} - \sum_{j=1}^{J} x_j \operatorname{logit}(\pi_{jq})}{U3(q)_{\min} - U3(q)_{\max}} \tag{14}$$

with

$$U3(q)_{\min} = \sum_{j=1}^{J} I[r_j(q) \leq x_+] \operatorname{logit}(\pi_{jq})$$

and

$$U3(q)_{\max} = \sum_{j=1}^{J} I[r_j(q) \geq J - x_+ + 1] \operatorname{logit}(\pi_{jq}).$$

$U3(q)$ is not defined for $x_+ = 0$ and $x_+ = J$. It has values in the interval [0,1].

3.3 Log-Likelihood

The third person-fit measure used was the log-likelihood of $\mathbf{x}$, denoted $\operatorname{Log}L(q)$ (see Levine & Rubin, 1979). Given the observed response vector $\mathbf{x}$ and class-specific item probabilities π_{jq}, for a respondent who is a member of class q, person-fit measure $\operatorname{Log}L(q)$ equals

$$\operatorname{Log}L(q) = \sum_{j=1}^{J} \{x_j \log \pi_{jq} + (1 - x_j) \log[1 - \pi_{jq}]\}. \tag{15}$$

Measure $\operatorname{Log}L(q)$ is similar to the l statistic (Levine & Rubin, 1979). Snijders (2001) derived an asymptotic sampling distribution for a standardized version of l (Drasgow et al., 1985). This standardized version serves the purposes of reducing the dependence of the outcome of the person-fit analysis on the θ level and of having a statistic that has a known sampling distribution. In this study, distributions of statistics were simulated for separate θs, and as a result, standardization was not necessary; see also Glas and Meijer (2003) and Molenaar and Hoijtink (1990).

3.4 Person-Fit Measure ζ

The fourth measure studied was

$$\zeta(q) = \sum_{j=1}^{J} [\pi_{jq} - x_j](s_j - \bar{s}), \tag{16}$$

Table 1

Example of $G^*(q)$, $U3(q)$, Log$L(q)$, and $\zeta(q)$, Applied to Six Item Score Vectors, With $J = 6$, and Known Item Response Probabilities (π_{qj})

j	1	2	3	4	5	6				
π_{jq}	.78	.70	.47	.35	.25	.11				
	Item Score Pattern						$G^*(q)$	$U3(q)$	Log$L(q)$	$\zeta(q)$
1	1	1	1	0	0	0	0.000	0.000	–2.195	–.2144
2	1	1	0	0	1	0	0.133	0.169	–3.174	–.0244
3	0	0	0	1	1	1	1.000	1.000	–7.996	0.0956
4	1	1	1	1	0	0	0.000	0.000	–2.814	–.1744
5	1	1	1	1	1	0	0.000	0.000	–3.913	–.0244
6	1	0	1	0	0	0	0.091	0.182	–3.043	–.0244

Note. In this example, only one latent class is considered.

with $\bar{s}$ being the mean of the proportions s_j across all j. This measure is based on the ζ_1 statistic (Tatsuoka, 1984). Measure $\zeta(q)$ uses both group information from the observed data and information about the discrepancy between the expected item scores under the OR-LCM and the observed item scores. Measure $\zeta(q)$ increases if an item with $s_j > \bar{s}$ is answered incorrectly or if an item with $s_j < \bar{s}$ is answered correctly. Thus, large positive values of $\zeta(q)$ indicate misfit. Glas and Meijer (2003) used Tatsuoka's (1984) ζ_1 under the three-parameter normal ogive model and found Type I error rates that were close to the nominal significance level and also high detection rates. For small numbers of items and small sample sizes, the Type I error rates tended to be somewhat higher than the nominal significance level.

Comparison of the four statistics. The four person-fit measures record misfit in different ways. Measures $G^*(q)$ and $U3(q)$ are based on deviations from the expected ranking of correct and incorrect scores under the OR-LCM. When the correct answers are given to the easiest items, there is no indication of misfit. Measures Log$L(q)$ and $\zeta(q)$, however, compare observed scores with the expectation under the OR-LCM. As a result, person-fit measures may produce different orderings of item score vectors according to increasing magnitude of misfit. To illustrate this, Table 1 gives the values of $G^*(q)$, $U3(q)$, Log$L(q)$, and $\zeta(q)$ for six item score vectors with known (but arbitrarily chosen) values of π_{jq} ($j = 1, \ldots, 6$). For example, $G^*(q)$ indicates perfect fit for the fifth vector, but Log$L(q)$ indicates that this is the least likely vector but one (i.e., the third vector).

4. Bayesian Goodness-of-Fit Assessment

Model fit can be investigated in a Bayesian context using PPCs (Berkhof et al., 2001; Gelman, Meng, & Stern, 1996; Gelman et al., 1995). A PPC compares the value of a goodness-of-fit statistic based on observed data with the values it would have were the study replicated with the same statistical model and the same values of the model parameters that produced the observed data. If the model fits, replicated data generated under the model should be similar to the observed data (Gelman et al., 1996, p. 165). Compared with classical approaches to goodness-of-fit assessment, the Bayesian approach is useful in particular when complex models are used for which it is impossible to derive the asymptotic distribution of goodness-of-fit statistics or when the asymptotic distribution is insufficiently accurate for realistic sample sizes (Berkhof et al., 2001). A property of PPCs is that

the fit measure that quantifies the discrepancy between the model and the data can be a function of both the unknown parameters and the data. This property facilitates a flexible way to assess the fit of model characteristics (see Berkhof et al., 2001) and takes the uncertainty about the model parameters explicitly into account. Next, the computation of the PPCs is discussed for $G^*(q)$, $U3(q)$, $\text{Log}L(q)$, and $\zeta(q)$, and Bayesian model estimation is outlined, which precedes the assessment of model fit. Then, the calculation of PPCs and its application to person-fit assessment in OR-LCMs is explained.

Bayesian Model Estimation

Bayesian model estimation methods fit a full probability model to a set of data and summarize the results by a posterior probability distribution on the parameters of the model (Gelman et al., 1995, p. 1). For posterior distributions that cannot be expressed in closed form, Markov chain Monte Carlo (MCMC) simulation methods are used to obtain a sample from the posterior distribution of the parameters of the statistical model of interest. This MCMC sample describes the posterior of the model parameters from which point estimates or interval estimates, such as the posterior mode and the standard deviation, can be obtained.

Hoijtink and Molenaar (1997; see also Van Onna, 2002) implemented the Gibbs sampler (Gelfand & Smith, 1990) to obtain samples from the posterior distribution for estimating and assessing model fit of OR-LCMs. The Gibbs sampler starts with the imputation of initial values for the model parameters, π_{jq}^0 and ω_q^0. Then, an iterative three-step procedure is used in which each parameter or set of parameters is sampled from their posterior distribution conditional on the data and the current values of all other parameters.

Suppose the Gibbs sampler is cycling through the lth iteration. In the first step, for each person v, class membership is sampled from its conditional posterior distribution given the current values of the parameters, π_{jq}^{l-1} and ω_q^{l-1}, and given the observed item score vector for person v (which is $\mathbf{x}_v$). In the second step, the class-specific probabilities are sampled given the appropriate inequality constraints on the π_{jq}s and conditional on class membership of each person obtained in Step 1. In the third step, the class weights ω_q ($q = 1, \ldots, Q$) are sampled conditional on class membership of each person from Step 1 and the current values of the class-specific probabilities from Step 2. This iterative sampling scheme is repeated until a stable form of the densities is obtained. Hoijtink and Molenaar's (1997) methodology was adopted in the present study.

Posterior Predictive Checks

Suppose a discrepancy measure, $T(\mathbf{x}, \boldsymbol{\xi})$, is a scalar summary of the data $\mathbf{x}$ and one or more of the model parameters collected in the vector $\boldsymbol{\xi}$. A PPC for $T(\mathbf{x}, \boldsymbol{\xi})$ compares the value of $T(\mathbf{x}, \boldsymbol{\xi})$ for observed data ($\mathbf{x}^{obs}$) with the values $T(\mathbf{x}, \boldsymbol{\xi})$ has for replicated data ($\mathbf{x}^{rep}$) under the hypothesized model. Replicated data under this model are data described by the posterior predictive distribution of $\mathbf{x}^{rep}$,

$$P(\mathbf{x}^{rep}|\mathbf{x}^{obs}) = \int P(\mathbf{x}^{rep}|\,\boldsymbol{\xi})\,P(\boldsymbol{\xi}|\mathbf{x}^{obs})d\,\boldsymbol{\xi}. \tag{17}$$

Lack of fit of the posterior predictive distribution to the data can be quantified by evaluating the tail-area probability or p value for $T(\mathbf{x}, \boldsymbol{\xi})$. This quantity defines the probability that the replicated data are more extreme than the observed data, as measured by the discrepancy measure (Gelman et al., 1995, p. 169):

$$\text{Bayes } p \text{ value } = P[T(\mathbf{x}^{rep},\ \boldsymbol{\xi}) \geq T(\mathbf{x}^{obs},\ \boldsymbol{\xi})|\mathbf{x}]. \tag{18}$$

The probability in equation (18) is taken over the joint posterior predictive distribution of $\mathbf{x}^{rep}$ and $\boldsymbol{\xi}$ (Gelman et al., 1995, p. 169). In practice, the p value (equation (18)) is computed using posterior simulations of $\boldsymbol{\xi}$ and $\mathbf{x}^{rep}$ (Gelman et al., 1996, pp. 169-170) with a two-step procedure. First, simulate L draws from the posterior distribution of $\boldsymbol{\xi}$, $P(\boldsymbol{\xi}|\mathbf{X})$. Second, for each sampled vector $\boldsymbol{\xi}$, draw one $\mathbf{x}^{rep}$ from the predictive distribution, $P(\mathbf{x}^{rep}|\,\boldsymbol{\xi})$. The result is a sample from the joint posterior distribution ($P(\mathbf{x}^{rep},\,\boldsymbol{\xi})$). The Bayesian p value is the proportion of replications for which $T(\mathbf{x}^{rep},\,\boldsymbol{\xi})$ was more extreme than $T(\mathbf{x}^{obs},\,\boldsymbol{\xi})$.

4.3 Application of PPCs for the Assessment of Person Fit

From the Gibbs sampler, a sample from the posterior distribution of item parameters and class weights is obtained. This sample was also used to simulate the posterior predictive distribution of the person-fit measures $G^*(q)$, $U3(q)$, $\mathrm{Log}L(q)$, and $\zeta(q)$. Because these person-fit measures are a function of both item parameters and class membership q (see equations (12) and (14)-(16)), posterior simulations of q for each individual to calculate the PPC were needed.

The following iterative sampling scheme was used. Let θ_v^* be the discrete variable for class membership of respondent v: $\theta_v^* \in \{1, \ldots, Q\}$. In addition, $\boldsymbol{\pi}_q = (\pi_{1q}, \ldots, \pi_{Jq})$ is the vector of item parameters in class q, and $\boldsymbol{\Pi} = (\boldsymbol{\pi}_1, \ldots,\, \boldsymbol{\pi}_Q)'$ is the $J \times Q$ matrix of item parameters, and $\omega = (\omega_1, \ldots, \omega_q)$ is the vector of class weights. Suppose L draws are taken from the joint posterior of $\boldsymbol{\Pi}$ and $\boldsymbol{\omega}$ with $l = 1, \ldots, L$. Then, for *each simulee*, the following steps were carried out for each sampled matrix ($\boldsymbol{\Pi}^l$) and vector $\boldsymbol{\omega}^l$:

1. Sample class membership, θ_v^{*l}, from the posterior distribution of θ_v^*, given $\boldsymbol{\Pi}^l$, $\boldsymbol{\omega}^l$, and $\mathbf{x}_v^{obs}$. For the next two steps, $q = \theta_v^{*l}$.
2. Sample $\mathbf{x}^{rep}$ from the predictive distribution of $\mathbf{x}$, given $\boldsymbol{\pi}_q^l$ and $\mathbf{x}_v^{obs}$. When $\mathbf{x}^{rep}$ contained only 1s or 0s, this vector was ignored because $G^*(q)$ and $U3(q)$ are not defined for such patterns, and a new item score vector $\mathbf{x}^{rep}$ was sampled. For comparison purposes, these extreme item score vectors were also ignored for evaluating Bayesian p values for $\mathrm{Log}L(q)$ and $\zeta(q)$, even though these measures are defined for such extreme item score vectors. Note that for a simulee, the predictive distribution of $\mathbf{x}^{rep}$ (see equation (17), with $\boldsymbol{\xi}$ replaced by q and $\boldsymbol{\pi}$) does not depend on the class weights.
3. Calculate the values of the person-fit measures for $\mathbf{x}^{rep}$ and $\mathbf{x}^{obs}$, given q and $\boldsymbol{\pi}_q^l$: This yields $T(\mathbf{x}^{rep}, q,\, \boldsymbol{\pi}_q^l)$ and $T(\mathbf{x}^{obs}, q,\, \boldsymbol{\pi}_q^l)$.

Step 3 led to the evaluation of the Bayesian p values. Finally, it may be noted that $\zeta(q)$ was computed using fixed s_j and $\bar{s}$, which were calculated from the observed data matrix.

5. Simulation Study

5.1 Purpose of the Study

The simulation study had two purposes. First, the ability to detect misfitting item score vectors under OR-LCMs by means of the Bayesian approach using PPCs was compared for the four person-fit measures: $G^*(q)$, $U3(q)$, $\mathrm{Log}L(q)$, and $\zeta(q)$. In particular, data under OR-LCMs were simulated, and the Type I error rates and detection rates were studied as a function of test length, kind of model violation representing different types of aberrant response behavior, and number of items that showed misfit. Second, the feasibility of using OR-LCMs to assess person-fit under NIRT models was investigated. In particular, the Type I error rates and detection rates were studied

for the LCM-adapted person-fit methods when applied to data simulated under a flexible IRT model.

5.2 Method

Data simulation. The IRFs were defined using the four-parameter logistic model (4-PLM) (Hambleton & Swaminathan, 1985, p. 48),

$$P_j(\theta) = \gamma_j + (\lambda_j - \gamma_j)\frac{\exp\left[\alpha_j(\theta - \delta_j)\right]}{1 + \exp\left[\alpha_j(\theta - \delta_j)\right]}, \tag{19}$$

where γ_j is the lower asymptote for $\theta \to -\infty$, λ_j is the upper asymptote for $\theta \to \infty$, α_j is the slope parameter, and δ_j is the location parameter. Because this IRF can be bent in many ways, the 4-PLM was considered to have enough flexibility to approximate NIRT models that place only order restrictions on the IRF (e.g., Sijtsma & Meijer, 2001). Data were simulated for $J = 20$ and $J = 40$. These test lengths represented short and long tests, respectively. For both test lengths, two setups for the parameters of the 4-PLM were used, resulting in one configuration of intersecting IRFs (few restrictions) and one configuration of nonintersecting IRFs (many restrictions). The item parameter values were chosen to be consistent with values that have been observed in practice or that have been used in related simulation studies (see, e.g., Emons et al., 2002; Glas & Meijer, 2003; Hambleton & Swaminathan, 1985, p. 240). In particular, the location parameters (Table 2) of the 4-PLM were chosen to be equally spaced along the latent trait scale (typical of many psychological tests) and well located with respect to the latent trait distribution (i.e., for the particular population, the test contains both easy and difficult items and several items of moderate difficulty in between). IRF slopes (Table 2) were sometimes fixed at an average value for all items and adapted to the dispersion of the latent trait distribution, and sometimes they were drawn from a distribution so as to represent more realistically some variation in item quality. Lower and upper asymptotes (Table 2) were chosen to vary a little but always close to 0 and 1, respectively. This was done to allow the possibility that, for some low-ability respondent, the probability of having a 1 score may be a little higher than 0 (some know just enough to solve the item), but for a high-ability respondent, some items may be easy to solve but not trivial.

Furthermore, data were simulated for a discrete θ (OR-LCM) and a continuous θ (IRT). The distribution of θ was chosen to be a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 2.75$. The parameters of the θ distribution were chosen such that, in combination with the IRFs, the sum score X_+ based on simulated data was reliable. For the condition with discrete θ, five fixed θ points were chosen (see lowest panel of Table 2), which resulted in an OR-LCM with five classes. The fixed θ values were chosen such that for a normally distributed θ with $\mu = 0$ and $\sigma^2 = 2.75$, they corresponded to percentile scores of 10, 30, 50, 70, and 90. The response probabilities are defined by $\pi_{jq} = P_j(\theta_q)$ (equation (19)), for $j = 1, \ldots, J$ and $q = 1, \ldots, Q$. At each θ_q level, 200 item score vectors were simulated. This resulted in 1,000 score vectors in total that were generated by a five–latent class model with uniformly distributed class membership. For continuous θ, the θs were drawn from a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 2.75$. Here, also 1,000 item score vectors were simulated. Because the fit of an item score vector was evaluated conditionally on θ, the person-fit results are independent of the particular shape of the θ distribution. It may be noted that the shape of the θ distribution may influence the accuracy of parameter estimates and thus indirectly the person-fit results. However, the effects of the shape of the θ distribution on the accuracy of parameter estimates were beyond the scope of the present study. To keep the calculations straightforward, the simulations were based on a uniform θ distribution (OR-LCM) and a normal θ distribution (IRT). Furthermore, to verify whether X_+ based on the data was reliable, Cronbach's

Table 2

Item and Ability Parameter Values Used in the Simulation Study

Item Parameter Values
Nonintersecting item response functions (IRFs), $J = 20$
$\alpha = 1$, for all j $\delta = -2.0, -1.8, \ldots, 2.0$, with $\delta \neq 0.0$ $\gamma = .25, .24, \ldots, .06$ $\lambda = .90$ for $j = 1, \ldots, 10$; .85 for j = 11, . . . , 15; .80 for $j = 16, \ldots, 20$
Nonintersecting IRFs, J = 40
α = 1 for all j δ = –2.0, –1.9, . . . , 2.0, with $\delta \neq 0.0$ γ = .40, .39, . . . , .01 λ = .90 for j = 1, . . . , 20; .85 for j = 21, . . . , 30; .80 for j = 31, . . . , 40
Intersecting IRFs, J = 20
α drawn from uniform distribution, U[0.6,1.4] δ = –2.0, –1.8, . . . , 2.0, with $\delta \neq 0.0$ $\gamma = .25, .24, \ldots, .06$ λ = .90 for $j = 1, \ldots, 10$; .85 for j = 11, . . . , 15; .80 for j = 16, . . . , 20
Intersecting IRFs, J = 40
α drawn from uniform distribution, U[0.6,1.4] $\delta = -2.0, -1.9, \ldots, 2.0$, with $\delta \neq 0.0$ $\gamma = .40, .39, \ldots, .01$ $\lambda = .90$ for $j = 1, \ldots, 20$; .85 for j = 21, . . . , 30; .80 for $j = 31, \ldots, 40$
Ability parameters for discrete θ
θ_1 = –3.52; θ_2 = –1.43; θ_3 = 0.00; θ_4 = 1.43; θ_5 = 3.52

(1951) alpha was calculated for each combination of test length, discrete or continuous θ, and the two configurations of IRFs. All alphas were at least .84.

Simulating item score vectors under aberrant response behavior. Item score vectors were simulated for two types of aberrant response behavior, labeled as *Cheating* and *Inattention*. For Cheating, it was assumed that students make a strategic choice when they do not master the textbook well enough. So they are not copying from just anyone who happens to sit next to them, but they make a rational choice to sit next to a high-ability student. It is also assumed that this high-ability neighbor has the answers correct. To simulate this, item scores were generated under the 4-PLM, but for the items with the highest δ values, the probability of a 1 score was fixed to 1.00. It may be noted that this also simulates preknowledge of the more difficult items. The sample contained only cheaters because the focus was on the detection rate of this kind of aberrant response behavior. For Inattention, it was assumed that some students underestimated the level of some of the easier items and misread them quite seriously. This lowered their probability of a correct answer to .25. This was simulated by a response probability of .25 for the items with the lowest δ values. Such spuriously low probabilities may also simulate test anxiety or fumbling (Haladyna, 1994; Meijer, 1994a). In both cases, a deterministic parameter setup was chosen for simplicity.

The Achilles heel of person-fit research is that the sample size is J, which in most tests is a small number. This small sample size also produces unreliable total scores (often, number-correct scores) in classical test theory and inaccurate latent trait estimates in IRT. Based on this unreliability, it was expected that no statistical method is able to accurately find person misfit on the basis of 1 (or 2 or 3) binary item score only. Thus, for both types of aberrant response behavior, simulees showed misfit to $J_{\text{misfit}} = 5$, 8, or 10 items.

The result is a design with 2 (discrete θ and continuous θ) $\times$ 2 (intersecting and nonintersecting IRFs) $\times$ 2 (test length) $\times$ 2 (type of aberrant behavior) $\times$ 3 (level of misfit) = 48 cells. Furthermore, to investigate the stability of the results, 50 replications were simulated for several representative cells.

Dependent variables. The dependent variables were the mean and the standard error of the Type I error rates and the detection rates, evaluated at three nominal significance levels: .01, .05, and .10. The detection rates were investigated using a two-step procedure. First, a data set of 1,000 item score vectors was simulated under the null model of normal response behavior. This data set is the calibration sample, denoted by $\mathbf{X}_{cal}$. The OR-LCM was fit to $\mathbf{X}_{cal}$ using the Bayesian estimation approach, yielding samples from the posterior distributions of all item parameters. Second, data sets were simulated under the model of aberrant response behavior, meaning that *all* item score vectors showed misfit. Then, the person-fit measures were applied to each of these aberrant behavior data matrices, and the PPCs were calculated using the sample from the posterior distributions of the item parameters obtained in Step 1. The detection rate is the proportion of item score vectors classified as aberrant. It may be noted that in most applications, the items are calibrated before they are put into actual practice (see Meijer, 1996).

Estimating the OR-LCM. The program ORCA (Van Onna, 2002) was used to estimate the OR-LCM and to sample from the posterior distributions of the item and class weight parameters. For the item parameters, the beta distribution with hyperparameters equal to 1 was used as the prior distribution; for the latent class weights, the Dirichlet distribution with hyperparameters equal to 1 was used as the prior distribution. Furthermore, the proportions s_j were used as starting values for the item parameters. For all class weights, Q^{-1} was used as the starting value. The number of iterations for the Gibbs sampler was fixed to 13,250, and the first 2,000 iterations served as burn-in iterations that were ignored in the statistical analysis. For the remaining iterations, sampled values from each 15th iteration were saved, yielding 750 samples from the posterior. These draws were used for the statistical analysis.

5.3 Results

5.3.1 Results for the Type I Error Rates

Table 3 shows the Type I error rates for the 20-item and 40-item tests for simulated data based on discrete θ (left-hand panel) and continuous θ (right-hand panel). The upper panel shows the results for intersecting IRFs, and the lower panel shows the results for nonintersecting IRFs. For intersecting IRFs and discrete θ, in most conditions, the Type I error rates were smaller than the nominal significance level, meaning that the person-fit tests were conservative. The most conservative person-fit measure was $\text{Log}L(q)$. In addition, simulated Type I error rates that exceeded the nominal significance level were found mainly for $\zeta(q)$. Differences between the simulated Type I error rates and nominal significance levels were smaller than .002 for $\alpha = .01$, smaller than .017 for $\alpha = .05$, and smaller than .028 for $\alpha = .10$.

The Type I error rates found for nonintersecting IRFs were close to the Type I error rates for intersecting IRFs. The largest difference found was .039 for $\zeta(q)$ at $\alpha = .10$. Thus, intersection of

Table 3

Type I Error Rates for Three Significance Levels, for Data Simulated Based on Discrete θ and Continuous θ, for Two Levels of Test Length and Two Configurations of Item Response Functions (IRFs)

	Discrete θ						Continuous θ					
	$J = 20$			$J = 40$			$J = 20$			$J = 40$		
	.01	.05	.10	.01	.05	.10	.01	.05	.10	.01	.05	.10
Intersecting IRFs												
$G^*(q)$	.005	.038	.090	.007	.044	.106	.004	.035	.083	.009	.062	.111
$U3(q)$	.006	.039	.093	.013	.043	.097	.004	.036	.087	.013	.062	.114
Log$L(q)$	.005	.032	.075	.003	.030	.082	.004	.028	.070	.006	.045	.083
$\zeta(q)$	.012	.067	.128	.011	.042	.082	.008	.041	.086	.007	.045	.093
Nonintersecting IRFs												
$G^*(q)$	.005	.044	.092	.009	.044	.083	.005	.045	.098	.009	.047	.102
$U3(q)$	.007	.045	.098	.012	.040	.089	.005	.041	.104	.009	.049	.104
Log$L(q)$	.002	.025	.071	.002	.028	.064	.003	.032	.081	.005	.035	.078
$\zeta(q)$	.011	.050	.089	.007	.032	.071	.006	.053	.111	.009	.042	.084

IRFs had a minor effect on the Type I error rates. Furthermore, the Type I error rates for continuous θ were close to those for discrete θ. For the 20-item test, differences between Type I error rates for continuous and discrete θ ranged from .001 to .005 for $\alpha = .01$, from .001 to .026 for $\alpha = .05$, and from .001 to .019 for $\alpha = .10$. It was concluded that the distribution of θ had only a small effect on the person-fit tests.

For the 40-item test, the Type I error rates were comparable to those found for the 20-item test. In general, the Type I error rates showed only small deviations from the nominal Type I error rates, and no main trends were found. For both discrete θ and continuous θ, differences were found that ranged from .001 to .007 for $\alpha = .01$, from .002 to .027 for $\alpha = .05$, and from .004 to .028 for $\alpha = .10$. Comparison of the Type I error rates for discrete θ with those for continuous θ again showed small differences ranging from .000 to .004 for $\alpha = .01$, from .003 to .019 for $\alpha = .05$, and from .001 to .019 for $\alpha = .10$. The person-fit measures $G^*(q)$ and $U3(q)$ were less conservative for continuous θ than for discrete θ. At the 5% significance level, the Type I error rates for $G^*(q)$ and $U3(q)$ were somewhat higher than the nominal significance level.

Stability of Type I error rates. For simulated data based on a discrete θ, the standard errors (*SE*s) of the simulated Type I error rates for $G^*(q)$, $U3(q)$, and Log$L(q)$ were below .003 for $\alpha = .01$, below .008 for $\alpha = .05$, and below .015 for $\alpha = .10$ (not tabulated). The *SE*s of the Type I error rates for $\zeta(q)$ were, in general, twice as high as the others. For continuous θ, the *SE*s were somewhat smaller than those for discrete θ. Moreover, the *SE*s for $\zeta(q)$ were similar to those for the other person-fit measures.

5.3.2 Results for the Detection Rates

Table 4 shows the detection rates of the four person-fit measures at three nominal significance levels, for data based on a discrete θ and intersecting IRFs, for two test lengths, three levels of misfit, and two types of aberrant response behavior. For Cheating (left-hand panel of Table 4), it can be seen that, except for Log$L(q)$, the detection rates were increased with larger numbers of misfitting

Table 4

Detection Rates for Three Significance Levels, for Data Simulated Based on Discrete θ and Intersecting Item Response Functions (IRFs), Generated for Cheating and Inattention

	Cheating						Inattention					
	$J = 20$			$J = 40$			$J = 20$			$J = 40$		
	.01	.05	.10	.01	.05	.10	.01	.05	.10	.01	.05	.10
						$J_{misfit} = 5$						
$G^*(q)$	.311	.619	.767	.325	.581	.718	.090	.336	.488	.077	.271	.439
$U3(q)$	.322	.640	.791	.335	.587	.740	.089	.333	.491	.064	.254	.427
Log$L(q)$	.387	.504	.571	.297	.525	.631	.209	.410	.514	.141	.353	.466
$\zeta(q)$	.414	.607	.735	.379	.578	.702	.204	.359	.463	.153	.295	.420
						$J_{misfit} = 8$						
$G^*(q)$	.438	.802	.907	.514	.764	.862	.197	.444	.574	.214	.457	.580
$U3(q)$	.469	.824	.918	.571	.829	.919	.206	.451	.575	.185	.449	.589
Log$L(q)$	.363	.460	.517	.480	.608	.667	.311	.449	.523	.328	.497	.593
$\zeta(q)$	.528	.747	.868	.554	.753	.862	.312	.459	.549	.295	.437	.531
						$J_{misfit} = 10$						
$G^*(q)$	.552	.906	.967	.639	.874	.934	.209	.437	.559	.296	.529	.626
$U3(q)$	.595	.912	.969	.719	.925	.974	.227	.447	.564	.259	.518	.636
Log$L(q)$	.303	.399	.470	.495	.600	.658	.297	.427	.500	.411	.545	.619
$\zeta(q)$	.581	.815	.927	.610	.839	.932	.322	.457	.544	.383	.510	.591

items. The detection rates for Log$L(q)$, however, were lower, and for $J_{misfit} = 8$ or 10, they were much smaller than for the other three person-fit measures. For example, for $J_{misfit} = 8$ and $\alpha = .05$, the detection rate for Log$L(q)$ was .460, whereas the detection rates for the other three person-fit measures were .75 or higher. The differences between the detection rates for $G^*(q)$, $U3(q)$, and $\zeta(q)$ ranged from .033 to .103 for $J_{misfit} = 5$, from .050 to .090 for $J_{misfit} = 8$, and from .042 to .097 for $J_{misfit} = 10$. Furthermore, it can be seen that for $\alpha = .05$ and $\alpha = .10$, measure $U3(q)$ yielded the highest detection rates at all levels of misfit. For $\alpha = .01$, however, $\zeta(q)$ was most effective for $J_{misfit} = 5$ and $J_{misfit} = 8$, and $U3(q)$ was most effective for $J_{misfit} = 5$.

In general, for Inattention (right-hand panel of Table 4), the detection rates were substantially lower than for Cheating. In particular, for $J_{misfit} = 5$, the person-fit measure Log$L(q)$ performed better than the other three measures. For a larger number of misfitting items, however, differences between the detection rates of Log$L(q)$ and the other measures were smaller; for $J_{misfit} = 10$, measure $G(q)$ and $U3(q)$ performed better than Log$L(q)$ and $\zeta(q)$; differences between detection rates for $G^*(q)$, $U3(q)$, and $\zeta(q)$ were .11 for $\alpha = .01$, .03 for $\alpha = .05$, and .020 for $\alpha = .10$. Thus, at significance levels of .05 and .10 and $J_{misfit} = 10$, these three person-fit measures performed equally well. For the 40-item test, the detection rates were comparable with those found for the 20-item test. It is interesting to note that for $J_{misfit} = 5$, the detection rates for the 20-item test were close to the detection rates for the 40-item test. Thus, it was the number of misfitting items, not the proportion of misfitting items, that affected the detection rates of the person-fit measures.

Table 5

Detection Rates for Three Significance Levels, for Data Simulated Based on Discrete θ and Nonintersecting Item Response Functions (IRFs), Generated for Cheating and Inattention

	Cheating						Inattention					
	$J = 20$			$J = 40$			$J = 20$			$J = 40$		
	.01	.05	.10	.01	.05	.10	.01	.05	.10	.01	.05	.10
						$J_{\text{misfit}} = 5$						
$G^*(q)$	.338	.656	.791	.311	.502	.622	.090	.335	.484	.073	.285	.425
$U3(q)$	.348	.678	.825	.318	.494	.632	.084	.316	.470	.067	.237	.367
Log$L(q)$	.396	.511	.582	.463	.633	.698	.155	.369	.471	.194	.429	.567
$\zeta(q)$	.435	.665	.807	.309	.450	.541	.222	.365	.460	.129	.245	.363
						$J_{\text{misfit}} = 8$						
$G^*(q)$	.453	.858	.948	.439	.695	.825	.231	.444	.578	.212	.459	.603
$U3(q)$	.516	.897	.968	.449	.724	.847	.217	.432	.573	.193	.394	.528
Log$L(q)$	.359	.472	.549	.554	.631	.677	.309	.452	.519	.337	.554	.657
$\zeta(q)$	.576	.809	.941	.474	.662	.776	.311	.445	.555	.252	.414	.528
						$J_{\text{misfit}} = 10$						
$G^*(q)$	.624	.990	1.000	.522	.769	.884	.244	.460	.579	.332	.573	.682
$U3(q)$	.691	.996	1.000	.545	.807	.920	.239	.461	.582	.288	.505	.632
Log$L(q)$	.303	.418	.482	.534	.600	.646	.320	.445	.510	.441	.586	.645
$\zeta(q)$	.634	.875	.999	.551	.755	.869	.346	.471	.580	.334	.503	.619

Because the results in the other conditions were, to a large extent, similar to the results for discrete θ with intersecting IRFs (Table 4), the discussion of these results is brief. In Table 5, it can be seen that for the 20-item test based on nonintersecting IRFs and discrete θ, the detection rates were somewhat higher than for intersecting IRFs (see Table 4), but for the 40-item test, the reverse result was found (cf. Tables 4 and 5). An exception was found for Log$L(q)$, which performed better for tests with nonintersecting IRFs for both the $J = 20$ and the $J = 40$. In addition, compared with the other measures, for nonintersecting IRFs, $J = 40$ and $J_{\text{misfit}} = 5$ (upper panel of Table 5), statistic Log$L(q)$ had the highest detection rates. Thus, test length and nonintersection of IRFs both had an effect on the effectiveness of Log$L(q)$.

The detection rates for continuous θ, the 20-item test and the 40-item, and intersecting IRFs can be found in Table 6. Compared with the detection rates for discrete θ (see Table 4), the detection rates for the Cheating condition were smaller for $G^*(q)$, $U3(q)$, and Log$L(q)$, whereas $\zeta(q)$ showed higher detection rates. For Inattention, the results were similar. Person-fit measure $\zeta(q)$ performed better for continuous θ in almost all conditions. There were no large effects of the θ distribution on the detection rates of the person-fit measures.

For continuous θ, $J = 20$, and nonintersecting IRFs, the detection rates for discrete θ (Table 5) were higher than those for continuous θ (Table 7). Also, Table 7 shows that for Cheating, person-fit measure $\zeta(q)$ performed well for α = .01, and $J_{\text{misfit}} = 8$ and $J_{\text{misfit}} = 10$. For the 40-item test, for $G^*(q)$ and $U3(q)$, the detection rates for Cheating were higher for continuous θ (see Table 5) than

Table 6

Detection Rates for Three Significance Levels, for Data Simulated Based on Continuous θ and Intersecting Item Response Functions (IRFs), Generated for Cheating and Inattention

	Cheating						Inattention					
	$J = 20$			$J = 40$			$J = 20$			$J = 40$		
	.01	.05	.10	.01	.05	.10	.01	.05	.10	.01	.05	.10
						$J_{\text{misfit}} = 5$						
$G^*(q)$	.308	.609	.768	.295	.554	.682	.052	.303	.468	.100	.317	.463
$U3(q)$	.313	.634	.790	.317	.566	.699	.056	.316	.478	.091	.298	.448
Log$L(q)$	.368	.473	.541	.314	.539	.629	.207	.404	.499	.140	.349	.476
$\zeta(q)$	.426	.654	.780	.342	.550	.667	.227	.384	.496	.150	.308	.423
						$J_{\text{misfit}} = 8$						
$G^*(q)$	.443	.806	.914	.505	.739	.847	.215	.434	.562	.225	.476	.601
$U3(q)$	.478	.832	.926	.518	.775	.873	.221	.454	.566	.204	.462	.599
Log$L(q)$	.356	.462	.512	.491	.590	.648	.329	.452	.512	.326	.508	.594
$\zeta(q)$	.526	.754	.892	.521	.705	.818	.321	.472	.568	.291	.459	.550
						$J_{\text{misfit}} = 10$						
$G^*(q)$	.481	.868	.959	.584	.824	.911	.215	.422	.556	.296	.521	.634
$U3(q)$	.527	.890	.962	.618	.861	.935	.229	.433	.566	.274	.515	.637
Log$L(q)$	.279	.369	.438	.498	.581	.629	.300	.409	.487	.407	.546	.621
$\zeta(q)$	.542	.797	.950	.591	.798	.900	.323	.476	.561	.352	.489	.583

for discrete θ (Table 7) in most cases; for Log$L(q)$ and $\zeta(q)$, the detection rates for continuous θ were somewhat lower than for discrete θ. For Inattention, the effects of the θ distribution on the detection rates were similar to those for Cheating, except for Log$L(q)$.

Stability of the detection rates. For discrete θ, the *SE*s of the simulated detection rates for $G^*(q)$ and $U3(q)$ ranged from .02 to .06, except for $J_{\text{misfit}} = 10$ and $\alpha = .01$, for which the *SE*s were .10 or .11. The *SE*s of the detection rates for Log$L(q)$ ranged from .01 to .02, and for $\zeta(q)$, they ranged from .09 to .17. For continuous θ, the *SE*s of the detection rates for $G^*(q)$, $U3(q)$, and Log$L(q)$ ranged from .02 to .04 for all nominal significance levels and all levels of J_{misfit}. For $\zeta(q)$, the *SE*s ranged from .07 to .12, meaning that $\zeta(q)$ is more sensitive to sampling fluctuations than the other three person-fit measures.

6. Discussion

This study investigated person-fit assessment using OR-LCMs. Two topics were investigated. First, the Type I error rates and the detection rates of four person-fit measures were compared. Simulations showed that, with a few exceptions, all person-fit measures were somewhat conservative with respect to Type I error rates. No condition was found in which Type I error rates were substantially greater than the nominal significance level. Furthermore, the detection rate largely depended on the number of misfitting items and the type of aberrant response behavior. None of the person-fit measures stood out with respect to detection rates. Thus, a general preference for one person-fit measure cannot be put forward. As shown by the results, the best choice for a specific

Table 7

Detection Rates for Three Significance Levels, for Data Simulated Based on Continuous θ and Nonintersecting Item Response Functions (IRFs), Generated for Cheating and Inattention

	Cheating						Inattention					
	$J = 20$			$J = 40$			$J = 20$			$J = 40$		
	.01	.05	.10	.01	.05	.10	.01	.05	.10	.01	.05	.10
						$J_{\text{misfit}} = 5$						
$G^*(q)$	.238	.503	.675	.313	.545	.656	.087	.348	.515	.109	.323	.484
$U3(q)$	.238	.542	.703	.345	.563	.681	.075	.339	.502	.083	.294	.460
$\text{Log}L(q)$	.396	.509	.571	.346	.531	.612	.157	.350	.454	.146	.341	.491
$\zeta(q)$	.382	.578	.691	.342	.531	.630	.235	.402	.505	.168	.311	.425
						$J_{\text{misfit}} = 8$						
$G^*(q)$	.331	.661	.825	.500	.722	.853	.215	.474	.602	.221	.467	.589
$U3(q)$	.363	.675	.835	.531	.777	.878	.205	.460	.598	.189	.444	.582
$\text{Log}L(q)$	.338	.430	.493	.508	.622	.673	.302	.457	.519	.314	.481	.578
$\zeta(q)$	.509	.722	.848	.515	.685	.795	.364	.512	.610	.281	.432	.532
						$J_{\text{misfit}} = 10$						
$G^*(q)$	.336	.781	.941	.575	.835	.929	.253	.482	.591	.281	.534	.640
$U3(q)$	.388	.766	.940	.623	.890	.959	.241	.484	.591	.251	.527	.641
$\text{Log}L(q)$	.285	.378	.461	.522	.591	.628	.318	.445	.514	.409	.530	.593
$\zeta(q)$	.572	.775	.886	.573	.781	.905	.341	.490	.584	.367	.517	.603

person-fit measure should be based on, for example, the type of aberrant behavior to be studied and the hypothesized IRT model. For example, if cheating is expected, $G^*(q)$ or $U3(q)$ should be used but not $\text{Log}L(q)$ due to the reduced detection rates for larger numbers of misfitting items.

Second, the feasibility of the OR-LCM approach to investigate person fit in NIRT models was studied. Apart from minor differences, the Type I error rates and the detection rates for continuous θ were comparable to those obtained for discrete θ. From this, it may be concluded that OR-LCMs may be used to investigate person fit in NIRT models. Moreover, the results also showed that an OR-LCM with relatively few latent classes was sufficiently accurate to approximate the NIRT model for person-fit assessment.

This study simulates person-fit assessment in applications in which data have been collected before the test is used in actual practice. In various test applications, previously collected data may not be available due to, for example, privacy or security reasons. In these cases, the model parameters have to be estimated from the sample at hand. Future research may focus on the sensitivity of the Type I error rates and the detection rates of the OR-LCM person-fit measures when the OR-LCMs are estimated from data that contain misfitting item score vectors. Meijer (1996) showed that the power of person-fit measures was reduced when the item and test characteristics were obtained in a sample that contained misfitting item scores. Subsequent research may focus on the effect sizes of misfitting item score vectors in the sample used to estimate the model on the performance of the person-fit measures.

Other research may based on person misfit that is the result of multidimensionality of measurement for some respondents but not all (Klauer, 1995). Thus, an appropriate multidimensional latent

trait vector may be simulated and the performance of OR-LCM person-fit statistics studied. Also, performance of OR-LCM person-fit statistics may be investigated for aberrant response behavior that is typical of personality testing, as opposed to cheating and inattention, which are more typical of ability testing. In a personality measurement context, the use of polytomous item scores also would seem to be an obvious choice.

Another research topic is the use of a mixture modeling approach, whereby the OR-LCM that explains response behavior has additional classes that specify certain types of aberrant response behavior. Such an approach was advocated by Van den Wittenboer et al. (2000), who used Guttman-based LCMs plus one or more latent classes that represent certain types of aberrant response behavior, such as guessing. It may be noted that person-fit analysis using the OR-LCM and NIRT approaches is useful, particularly when a parametric IRT model does not fit the data. This is due to their often weaker assumptions and, as a consequence, their greater orientation toward data analysis than model fitting (e.g., Junker & Sijtsma, 2001a). The main problems with person-fit analysis, both nonparametric and parametric, are the limited sample size (J) and the inherent unreliability of individual item scores (Meijer, Sijtsma, & Molenaar, 1995). There is little that can be done about this, unless tests are made much longer and only high-quality items are admitted to tests. Very long tests are not realistic given limited testing time and a finite motivation of respondents for answering more and more items, and high-quality items are sparse. Research has been started that combines a person-fit methodology (a combination of global, graphical, and local person-fit methods; see Emons, 2003a) with the use of auxiliary information from background variables, such as educational results and school performance. The person-fit methodology allows one to look at the same data from different angles. By combining results with other information about the respondents, it is hoped that more powerful and stable conclusions about individuals' performance can be reached.

References

Berkhof, J., van Mechelen, I., & Hoijtink, H. (2001). Posterior predictive checks: Principles and discussion. *Computational Statistics*, *15*, 337-354.

Birenbaum, M. (1986). Effect of dissimulation motivation and anxiety on response pattern appropriateness measures. *Applied Psychological Measurement*, *10*, 167-174.

Birenbaum, M., & Nassar, F. (1994). On the relationship between test anxiety and test performance. *Measurement and Evaluation in Counseling and Development*, *27*, 293-301.

Boomsma, A., van Duijn, M. A. J., & Snijders, T. A. B. (Eds.). (2001). *Essays on item response theory*. New York: Springer.

Bradlow, E. T., & Weiss, R. E. (2001). Outlier measures and norming methods for computerized adaptive tests. *Journal of Educational and Behavioral Statistics*, *26*, 85-104.

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

Croon, M. A. (1991). Investigating Mokken scalability of dichotomous items by means of ordinal latent class models. *British Journal of Mathematical and Statistical Psychology*, *44*, 315-331.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, *11*, 59-79.

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*, 67-68.

Emons, W. H. M. (2003a). *Detection and diagnosis of misfitting item-score vectors.* Unpublished doctoral dissertation, Tilburg University, Tilburg, The Netherlands.

Emons, W. H. M. (2003b). Investigating the local fit of item-score vectors. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics* (pp. 289-296). Tokyo: Springer.

Emons, W. H. M., Meijer, R. R., & Sijtsma, K. (2002). Comparing simulated and theoretical sampling distributions of the *U*3 person-fit statistic. *Applied Psychological Measurement, 26*, 88-108.

Gelfand, A. E., & 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., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.

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

Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 60-90). Princeton, NJ: Princeton University Press.

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. Englewood Cliffs, NJ: Lawrence Erlbaum.

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

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Thousand Oaks, CA: Sage.

Hoijtink, H., & 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.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21*, 1359-1378.

Junker, B. W., & Sijtsma, K. (2001a). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement, 25*, 211-220.

Junker, B. W., & Sijtsma, K. (Eds.). (2001b). Nonparametric item reponse theory [Special issue]. *Applied Psychological Measurement, 25*(3).

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: Springer-Verlag.

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

Meijer, R. R. (1994a). *Nonparametric person fit analysis*. Unpublished doctoral dissertation, Vrije Universiteit, Amsterdam, The Netherlands.

Meijer, R. R. (1994b). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18*, 311-314.

Meijer, R. R. (1996). The influence of the presence of deviant item score patterns on the power of a person-fit statistic. *Applied Psychological Measurement, 20*, 141-154.

Meijer, R. R. (1997). Person fit and criterion-related validity: An extension of the Schmitt, Cortina, and Whitney study. *Applied Psychological Measurement, 21*, 99-113.

Meijer, R. R. (1998). Consistency of test behaviour and individual difference in precision of prediction. *Journal of Occupational and Organizational Psychology, 71*, 147-160.

Meijer, R. R. (2002). Outlier detection in high-stakes certification testing. *Journal of Educational Measurement, 39*, 219-233.

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 of recent 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.

Meijer, R. R., Sijtsma, K., & Molenaar, I. W. (1995). Reliability estimation for single dichotomous items based on Mokken's IRT model. *Applied Psychological Measurement, 19*, 323-335.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. New York: De Gruyter.

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*, 417-430.

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

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement*, *19*, 121-129.

Nering, M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement*, *21*, 115-127.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*, 611-630.

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

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research*, *35*, 543-568.

Reise, S. P., & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology*, *65*, 143-151.

Rosenbaum, P. R. (1987). Probability inequalities for latent scales. *British Journal of Mathematical and Statistical Psychology*, *40*, 157-168.

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement*, *23*, 41-53.

Sijtsma, K., & Hemker, B. T. (2000). A taxonomy of IRT models for ordering of persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics*, *25*, 391-415.

Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, *49*, 79-105.

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika*, *66*, 191-208.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*, 331-342.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*, 293-325.

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

Van den Wittenboer, G., Hox, J. J., & De Leeuw, E. (2000). Latent class analysis of respondent scalability. *Quality & Quantity*, *34*, 177-191.

Van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties* (*Comparability of individual test performance*). Lisse: Swets & Zeitlinger.

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Van Onna, M. J. H. (2002). Bayesian estimation and model selection in ordered latent class models for polytomous items. *Psychometrika*, *67*, 519-538.

Vermunt, J. K. (2001). The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. *Applied Psychological Measurement*, *25*, 283-294.

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement*, *20*, 71-88.

Author's Address

Address correspondence to Wilco H. M. Emons, Department of Methodology and Statistics, FSW, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands; e-mail: w.h.m.emons@uvt.nl.