PSYCHOMETRIKA—VOL. 72, NO. 2, 159–180
JUNE 2007
DOI: 10.1007/s11336-003-1081-5

A PERSON FIT TEST FOR IRT MODELS FOR POLYTOMOUS ITEMS

C.A.W. GLAS AND ANNA VILLA T. DAGOHOY

UNIVERSITY OF TWENTE, THE NETHERLANDS

A person fit test based on the Lagrange multiplier test is presented for three item response theory models for polytomous items: the generalized partial credit model, the sequential model, and the graded response model. The test can also be used in the framework of multidimensional ability parameters. It is shown that the Lagrange multiplier statistic can take both the effects of estimation of the item parameters and the estimation of the person parameters into account. The Lagrange multiplier statistic has an asymptotic χ^2-distribution. The Type I error rate and power are investigated using simulation studies. Results show that test statistics that ignore the effects of estimation of the persons' ability parameters have decreased Type I error rates and power. Incorporating a correction to account for the effects of the estimation of the persons' ability parameters results in acceptable Type I error rates and power characteristics; incorporating a correction for the estimation of the item parameters has very little additional effect. It is investigated to what extent the three models give comparable results, both in the simulation studies and in an example using data from the NEO Personality Inventory-Revised.

Key words: item response theory, person fit, model fit, multidimensional item response theory, polytomous items, power, Type I error.

1. Introduction

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 used holds. Fit of items can be investigated across persons and fit of persons can be investigated across items. In psychological and educational measurement, instruments are developed that are used in a population of persons and item fit is used to evaluate to what extent an IRT model fits an instrument in a particular population (see, for instance, Andersen, 1973; Yen, 1981, 1984; Molenaar, 1983; Glas, 1988, 1999; Glas & Suárez-Falcón, 2003; Orlando & Thissen, 2000).

But although the IRT model may generally fit the data, specific persons may still produce patterns that are highly unlikely given the model. For instance, some persons may give random responses because they are unmotivated to take the test. Using person fit statistics, the fit of a score pattern can be determined under the null-hypothesis that the IRT model holds. Meijer and Sijtsma (1995, 2001) give an overview of person fit statistics proposed for various IRT models. Most person fit statistics were developed for IRT models for dichotomous items (Levine & Rubin, 1979; Wright & Stone, 1979; Tatsuoka, 1984; Smith, 1985, 1986; Klauer & Rettig, 1990; Drasgow, Levine, & McLaughlin, 1991; Sijtsma & Meijer, 2001). Person fit tests for polytomous items are far less numerous (such tests were developed by Drasgow, Levine, & Williams, 1985; Wright & Masters, 1982; van Krimpen-Stoop & Meijer, 2002).

One of the problems of person fit statistics is that the derivation of the distribution of the statistics has to account for the fact that item and person parameters are estimated. These estimates usually decrease the asymptotic variance of most statistics proposed in the literature. Therefore, their asymptotic distribution is usually unknown (see, for instance, Nering, 1995; Reise, 1995). There are several solutions to this problem. The first one is to avoid the estimation of the person

Requests for reprints should be sent to Cees A.W. Glas, Department of Research Methodology, Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500AE, Enschede, The Netherlands. E-mail: c.a.w.glas@gw.utwente.nl

parameter by applying the nonparametric IRT. Sijtsma and Meijer (2001) show that a count of Guttman errors is a good person fit index. In the present paper, however, we focus on applications that call for explicit parametric IRT models, such as computerized adaptive testing. A second solution pertains to the Rasch model (Rasch, 1960) and other parametric IRT models that belong to the class of exponential family models, such as the partial credit model (PCM) (Masters, 1982). These models have a sufficient statistic for the person parameter and conditioning on the observed value of the sufficient statistic can correct for the effect of estimation (Molenaar & Hoijtink, 1990; von Davier & Molenaar, 2003).

Further, for the Rasch model, Klauer (1995) presents a number of uniformly most-powerful tests that also do not require estimation of the ability parameter. In the present paper, however, the focus is on models outside the exponential family. Finally, Snijders (2001) proposed a method for standardization of a specific class of person fit statistics for dichotomous items, such that their asymptotic distribution can be properly derived. However, generalizations to polytomously scored items are not available. The relation between the test statistics presented below and the test statistics considered by Snijders (2001) will be returned to in the Discussion section of this paper.

This paper is organized as follows. First, three models for polytomously scored items will be introduced: the generalized partial credit model (GPCM) (Muraki, 1992), the sequential model (SM) (Tutz, 1990), and the graded response model (GRM) (Samejima, 1969, 1973). Also multidimensional generalizations of these models will be considered. Second, a person fit statistic for testing the constancy of the person ability parameter will be introduced. The test is based on the Lagrange multiplier (LM) test. In Appendix A it will be shown how the LM test can be used to test other model violations. Third, the Type I error rate and power of the test will be assessed using simulation studies. In most studies of person fit, the influence of the estimates of the item parameters is not considered. A study of the effect of uncertainty of item parameter-estimation on ability estimates by Tsutakawa and Johnson (1990) showed only minor effects. This will probably also hold for person fit statistics, but this point has not been systematically investigated. Therefore, three types of tests will be addressed:

(1) tests that do not take estimation effects into account;
(2) tests that take the effects of ability estimation into account; and
(3) tests that take both the effects of estimation of the item and person parameters into account.

Next, the robustness of the testing procedure will be assessed. This part of the study is related to reports that, although the rationales underlying the GPCM, SM, and GRM are very different, the models are hard to distinguish because their response functions are very close (Verhelst, Glas, & de Vries, 1997). It will be investigated whether the exchangeability of the three models in practical situations also extends to person fit tests, or, put another way, whether the three models can be distinguished using person fit tests. Finally, the performance of the person fit tests will be evaluated using data sets from the NEO Personality Inventory-Revised test.

2. IRT Models for Polytomous Items

Consider a test with polytomously scored items labeled $i = 1, \ldots, K$. Every item has response categories labeled $j = 0, \ldots, m_i$. Item responses will be coded by stochastic variables X_{ij} ($i = 1, \ldots, K$; $j = 0, \ldots, m_i$; in the sequel the index i of m is dropped for convenience) with realizations x_{ij}. $x_{ij} = 1$ if a response was given in category j, and zero otherwise. It will be assumed that the response categories are ordered, and that there exists a latent ability variable θ such that a response in a higher category reflects a higher ability level than a response in a lower

category. The probability of scoring in a response category j on item i is given by a response function $P_{ij}(\theta) = P(X_{ij} = 1 \mid \theta)$. In many measurement situations, such as in measurement of abilities, it is reasonable to assume that the response function of the category $j = 0$ decreases as a function of ability, the response function for $j = m$ increases as a function of ability and the response functions of the intermediate categories are single peaked. Mellenbergh (1995) showed that IRT models with such response functions can be divided into three classes. Though the rationales underlying the models in these classes are very different, their response functions appear to be very close (Verhelst et al., 1997), so the models might be hard to distinguished on the basis of empirical data. One of the topics addressed in this paper is whether this also holds when using person fit tests. We will now introduce three models from the three classes distinguished by Mellenbergh (1995).

2.1. The Graded Response Model

Using the abbreviation for the logistic function given by

$$\Psi(x) = \frac{\exp(x)}{1 + \exp(x)}, \tag{1}$$

the probability of a response in category j of item i, $P(X_{ij} = 1 \mid \theta)$, is given by

$$P_{ij}(\theta) = \begin{cases} 1 - \Psi(\alpha_i\theta - \beta_{i1}) & \text{if } j = 0, \\ \Psi(\alpha_i\theta - \beta_{ij}) - \Psi(\alpha_i\theta - \beta_{i(j+1)}) & \text{if } 0 < j < m, \\ \Psi(\alpha_i\theta - \beta_{im}) & \text{if } j = m, \end{cases} \tag{2}$$

(Samejima, 1969). To ensure that the probabilities $P_{ij}(\theta)$ are positive, the restriction $\beta_{i(j+1)} > \beta_{ij}$ for $0 < j < m$ is imposed.

2.2. The Sequential Model

In the SM (Tutz, 1990) the probability of a response in category j of item i is given by

$$P_{ij}(\theta) = \begin{cases} 1 - \Psi(\alpha_i\theta - \beta_{i1}) & \text{if } j = 0, \\ \prod_{h=1}^{j} \Psi(\alpha_i\theta - \beta_{ih})[1 - (\Psi(\alpha_i\theta - \beta_{i(j+1)}))] & \text{if } 0 < j < m, \\ \prod_{h=1}^{m} \Psi(\alpha_i\theta - \beta_{ih}) & \text{if } j = m. \end{cases} \tag{3}$$

Verhelst et al. (1997) note that in the SM every polytomous item can be viewed as a sequence of virtual dichotomous items. These dichotomous items are considered to be presented as long as a correct response is given, and the presentation stops when an incorrect response is given. An important consequence of this conceptualization of the response process is that estimation and testing procedures for the two-parameter logistic (2PL) model with incomplete data can be directly applied to the SM.

2.3. The Generalized Partial Credit Model

In the GPCM (Muraki, 1992) the probability of a response in category j of item i is given by

$$P_{ij}(\theta) = \frac{\exp(j\alpha_i\theta - \beta_{ij})}{1 + \sum_{h=1}^{m} \exp(h\alpha_i\theta - \beta_{ih})}. \tag{4}$$

The PCM (Masters, 1982) is the special case where $\alpha_i = 1$ for all items i. The item parameters are usually reparametrized as $\beta_{ij} = \sum_{h=1}^{j} \eta_{ih}$. In that case, η_{ij} can be interpreted as so-called boundary parameters: η_{ij} is the position on the latent θ-scale where $P_{i(j-1)}(\theta) = P_{ij}(\theta)$.

2.4. Multidimensional Generalizations

In many situations the assumption that an individual's response behavior can be explained by a unidimensional person parameter θ does not hold. In that case the assumption of a unidimensional person parameter can be replaced by the assumption of a multidimensional person parameter $\theta_1, \ldots, \theta_q, \ldots, \theta_Q$. The multidimensional versions of the models given by (2), (3), and (4) are defined by replacing $\alpha_i\theta$ by $\sum_{q=1}^{Q} \alpha_{iq}\theta_q$. Further, it is usually assumed that the parameters $\theta_1, \ldots, \theta_q, \ldots, \theta_Q$ have a joint Q-variate normal distribution (McDonald, 1997; Reckase, 1997).

3. The LM Test

Recently, LM tests for IRT models have been proposed by Glas (1998, 1999), Glas and Suárez-Falcón (2003), and Jansen and Glas (2005). The LM test (Aitchison & Silvey, 1958) is equivalent with the efficient score test (Rao, 1947) and the modification index that is commonly used in structural equation modeling (Sörbom, 1989). The purpose of the LM test is to compare two models, a model under the null-hypothesis and a more general model that is derived from the model under the null-hypothesis by adding parameters. Only the model under the null-hypothesis needs to be estimated.

The LM test is developed as follows. Consider some general parametrized model, and a special case of the general model, called the restricted model. The restricted model is derived from the general model by imposing constraints on the parameter space. In many instances, this is accomplished by fixing one or more parameters of the general model to constants. The LM test is based on the evaluation of the first-order partial derivatives of the log-likelihood function of the general model, evaluated using the maximum likelihood estimates of the restricted model. Because the parameters of the restricted model are estimated by maximum likelihood, their first-order derivatives are equal to zero at the solution of the likelihood equations. The magnitudes of the first-order partial derivatives corresponding to the other parameters determine the value of the statistic: the closer they are to zero, the better the model fit. Note that the likelihood function of the model under the null-hypothesis can be viewed as a special case of the general model where the parameters are constrained. Such constraints can be introduced to the likelihood function via the well-known LM method, which motivates the naming of the test.

More formally, the principle of the LM test can be described as follows. Consider a general model with parameters η. In the applications presented below, the restricted model is derived from the general model by fixing one or more parameters to zero. So, if the vector of the parameters of the general model, say η, is partitioned $\eta = (\eta_1, \eta_2)$, the null-hypothesis entails $\eta_2 = \mathbf{c}$, where $\mathbf{c}$ is a vector of constants. In the present application, $\mathbf{c}$ will be zero. Let $\mathbf{h}(\eta)$ be the first-order partial derivatives of the log-likelihood of the general model, that is, $\mathbf{h}(\eta) = \partial \log L(\eta)/\partial\eta$. This vector of partial derivatives gauges the change of the log-likelihood as a function of local changes in η. Let the vector of partial derivatives $\mathbf{h}(\eta)$ be partitioned as $(\mathbf{h}(\eta_1), \mathbf{h}(\eta_2))$. Then the test is based on the statistic

$$LM = \mathbf{h}(\eta_2)'\Sigma^{-1}\mathbf{h}(\eta_2), \tag{5}$$

where

$$\Sigma = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \tag{6}$$

and

$$\Sigma_{pq} = -\frac{\partial^2 \log L(\eta)}{\partial\eta_p\partial\eta_q'}, \tag{7}$$

for $p = 1, 2$ and $q = 1, 2$. The LM statistic is evaluated using the maximum likelihood estimates of the parameters of the restricted model. Therefore, at the maximum likelihood estimate of η_1, it holds that $\mathbf{h}(\eta_1) = 0$. In the applications presented below, the model under the null-hypothesis will be an IRT model. The LM statistic has an asymptotic χ^2-distribution with degrees of freedom equal to the number of parameters in η_2 (Aitchison & Silvey, 1958).

The variance of the parameter estimates plays an explicit role in the distribution of the LM statistics. Glas (1999) shows that the matrices Σ and Σ_{22} in (6) can be viewed as the asymptotic covariance matrices of $\mathbf{h}(\eta_2)$ with η_1 estimated and known, respectively. Further, Σ_{11}^{-1} is the asymptotic covariance matrix of the estimate of η_1, so the term $\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ accounts for the influence of the estimation of η_1 on the covariance matrix of $\mathbf{h}(\eta_2)$. Therefore, in the LM test, the variance of the estimates of the parameters is explicitly taken into account.

4. An LM Test for Constancy of Theta

To illustrate the application of the LM test as a test of person fit and to illustrate its relation to some existing tests, an LM test for the constancy of θ over partial response patterns for the GPCM will be presented. For the Rasch model, Smith (1985, 1986) introduced the UB test, which is a Pearson-type test for evaluating the constancy of the ability parameter across subtests. For the UB test, the complete response pattern is split up into a number of parts, say the parts $g = 0, \ldots, G$, and it is evaluated whether the same ability parameter θ can account for all partial response patterns. In this section, this approach will be generalized to polytomously scored items.

Let A_g be the set of the indices of the items in part g. Consider a model that is an alternative to the GPCM given by (4). In the alternative model it is assumed that the response pattern cannot be described by one ability parameter, that is, for $g > 0$, define

$$P(X_{ij} = 1 \mid \theta, z_{ig} = 1) = \frac{\exp(j\alpha_i(\theta + \delta_g) - \beta_{ij})}{1 + \sum_{h=1}^{m} \exp(h\alpha_i(\theta + \delta_g) - \beta_{ih})}, \tag{8}$$

where z_{ig} $(g = 1, \ldots, G)$ is an indicator assuming a value one if $i \in A_g$ and zero otherwise. For items where $z_{ig} = 0$ for $g = 1, \ldots, G$, the GPCM holds, so this partial response pattern on these items is used as a reference. For the remainder of the response pattern, it is hypothesized that additional ability parameters δ_g $(g = 1, \ldots, G)$ are necessary to describe the response behavior.

In this section, an LM test accounting for the effects of estimation of the person parameter θ will be derived. An LM test that also accounts for the effects of estimation of the item parameters will be treated in a following section. To define the statistic, an expression for the derivatives with respect to the ability parameters is needed. Note that the first-order derivatives with respect to θ of (4) and (8), under the null-hypothesis that $\delta_g = 0$ for all g, are the same. In Appendix A it is shown that the first-order derivative of the log-likelihood is given by

$$\frac{\partial \log L}{\partial \theta} = \sum_{i=1}^{k} \sum_{j=0}^{m} \left[x_{ij} \left(j\alpha_i - \sum_{h=1}^{m} h\alpha_i P_{ih}(\theta) \right) \right]$$

$$= \sum_{i=1}^{k} [y_i - E_\theta(Y_i)],$$

where $y_i = \sum_{j=0}^{m} x_{ij} j\alpha_i$, that is, it is the weighted score on item i, and $E_\theta(Y_i)$ is its expectation. In Appendix A it is also shown that

$$\frac{\partial \log L}{\partial \delta_g} = \sum_i z_{ig} [y_i - E_\theta(Y_i)], \tag{9}$$

and that the second-order derivatives are

$$\frac{\partial^2 \log L}{\partial \theta^2} = -\sum_{i=1}^{k}\sum_{j=0}^{m} j\alpha_i P_{ij}(\theta)\,[j\alpha_i - E_\theta(Y_i)]\,,$$

$$\frac{\partial^2 \log L}{\partial \delta_g^2} = -\sum_{i=1}^{k} z_{ig}\sum_{j=0}^{m} j\alpha_i P_{ij}(\theta)\,[j\alpha_i - E_\theta(Y_i)]\,,$$

$$\frac{\partial^2 \log L}{\partial \theta \partial \delta_g} = -\sum_{i=1}^{k} z_{ig}\sum_{j=0}^{m} j\alpha_i P_{ij}(\theta)\,[j\alpha_i - E_\theta(Y_i)]\,,$$

$$\frac{\partial^2 \log L}{\partial \delta_g \partial \delta_{g'}} = 0,$$

where $g \neq g'$. Inserting these expressions into (5) and (6) gives an expression for the LM statistic for testing the constancy of the ability parameter over the partial response patterns.

Now consider the case $G = 1$ where the test is split into two subtests: say the first part of the test and the second part of the test. Alternative partitionings of the test into two nonempty subtests are also possible. In this case, the null-hypothesis becomes $\delta = 0$. The matrices Σ_{11}, Σ_{22}, and Σ_{12} in (6) become scalars and the LM statistic specializes to

$$LM = \frac{h_2^2}{\sigma_{22} - \sigma_{12}^2\sigma_{11}^{-1}}, \tag{10}$$

where h_2 is given by (9). From the theory outlined in the previous section, it follows that this statistic has an asymptotic χ^2-distribution with one degree of freedom. Note that h_2 turns out to be a difference between observed and expected values. In Appendix A, where a general formulation of the test statistic is given, it can be verified that this also holds for the SM, so in these two cases h_2 can be viewed as a residual.

Note that $\sigma_{12}^2/\sigma_{11}$ takes into account the loss of variation due to the estimation of θ. In the simulation studies reported below, we shall also consider a version of the statistic where the term $\sigma_{12}^2/\sigma_{11}$ is deleted. Disregarding the effects of estimation of θ results in a statistic,

$$UB = \sum_{g=1}^{G} \frac{\left[\sum_{i\in A_g}[y_i - E_\theta(Y_i)]\right]^2}{\sum_{i\in A_g}^{k}\sum_{j=0}^{m} j\alpha_i P_{ij}(\theta)\,[j\alpha_i - E_\theta(Y_i)]}. \tag{11}$$

For the case of dichotomously scored items, the formulation of this statistic is similar to the formulation of the UB statistic by Smith (1985, 1986). The statistics given by (10) and (11) will be compared in the simulation studies reported below.

4.1. Incorporating Item Parameter Estimates

In marginal maximum likelihood (MML) (Bock & Aitkin, 1981; Bock, Gibbons, & Muraki, 1988), it is assumed that the ability parameters are independent and normally distributed. The approach derives its name from maximizing a log-likelihood that is marginalized with respect to θ, rather than maximizing the joint log-likelihood of all abilities parameters θ and all item parameters. The essential feature of MML estimation is that the number of parameters is constant, i.e., it does not grow with the number of observations. Therefore, for the present application, we consider a model where we marginalize over all ability parameters (because their number grows

with the number of respondents) except the ability parameter of the person we are interested in. Further, we now consider all available data. The log-likelihood is split into two parts, one pertaining to the marginal likelihood of N respondents (denoted by L_m) and one pertaining to the respondent that is the focus of attention (say, observation $N+1$; this likelihood is denoted by L_p). So we have

$$\begin{aligned}\log L &= \log L_m + \log L_p \\ &= \sum_{n=1}^{N} \log \int \prod_{i=1}^{k} \prod_{j=0}^{m} P_{ij}(\theta)^{x_{nij}} G(\theta)\, d\theta + \sum_{i=1}^{k} \sum_{j=0}^{m} x_{ij} \log P_{ij}(\theta), \end{aligned} \tag{12}$$

where x_{nij} are the responses of N persons, and $G(\theta)$ is a, usually normal, ability distribution. The log-likelihood in (12) is concurrently maximized with respect to the item parameters, the parameters of $G(\theta)$, and the ability parameter of the focal person. Mislevy (1986) shows that maximization of L_m can be further enhanced by introducing fixed and empirical priors.

In the section on the LM test, the parameter vector was partitioned $\eta = (\eta_1, \eta_2)$, where, under the restricted model, η_1 are the free parameters and η_2 are the fixed parameters. In the present case, the item parameters, the parameters of $G(\theta)$, and the ability parameter of the focal person are stacked in η_1 and the parameters representing model violations (δ_g, $g = 1, \ldots, G$) are stacked in η_2. The parameters in η_1 are partitioned into the θ of the focal person and all the other parameter which are denoted by ξ. To perform the test we proceed in two steps: first we estimate ξ, and then we compute the LM statistic defined by (5). The first step boils down to solving the simultaneous system $\partial[\log L_m + \log L_p]/\partial\xi = 0$ and $\partial \log L_p/\partial\theta = 0$. (It should be noted that it is assumed that the parameter estimates converge in the open parameter space, that is, there are no boundary values so that the gradient can be assumed to be arbitrarily close to zero at the final estimate.) First- and second-order derivatives of $\log L_m$ with respect to item and population parameters ξ can be found in Glas (1999), the derivatives of $\log L_p$ with respect to θ were given above, and the derivatives of $\log L_p$ with respect to the item parameters can be found in papers on joint maximum likelihood estimation for IRT, say, Wright and Linacre (1992). In practice, the estimates of the item and population parameters ξ will not change much when one person is singled out as a target; in practice, only a few iteration steps are needed.

With the parameter estimates available, the LM statistic (5) can be computed with $\mathbf{h}(\eta_2) = \partial \log L_p/\partial\eta_2$, where η_2 is δ_g ($g = 1, \ldots, G$) and a matrix of weights

$$\sum = \frac{\partial^2 \log L_p}{\partial \eta_2^2} - \left[\frac{\partial^2 \log L_p}{\partial \eta_2 \partial \xi^t} \quad \frac{\partial^2 \log L_p}{\partial \eta_2 \partial \theta}\right] \begin{bmatrix} \dfrac{\partial^2[\log L_m + \log L_p]}{\partial \xi\, \partial \xi^t} & \dfrac{\partial^2 \log L_p}{\partial \xi\, \partial \theta} \\ \dfrac{\partial^2 \log L_p}{\partial \theta\, \partial \xi^t} & \dfrac{\partial^2 \log L_p}{\partial \theta^2} \end{bmatrix}^{-1} \begin{bmatrix} \dfrac{\partial^2 \log L_p}{\partial \xi\, \partial \eta_2} \\ \dfrac{\partial^2 \log L_p}{\partial \theta\, \partial \eta_2} \end{bmatrix}.$$

Also, for this expression, derivatives of $\log L_m$ with respect to ξ can be found in Glas (1999), derivatives of $\log L_p$ with respect to ξ can be found in papers on joint maximum likelihood estimation, and derivatives of $\log L_p$ with respect to η_2 and θ can be found above.

5. Simulation Studies

Three sets of simulation studies will be reported. In the first set, a comparison is made between the Type I error rate and the power of tests that do and do not take the ability estimates into account, and tests that take both the estimates of the item and person parameters into account. In the second set of simulations, the robustness of the tests with respect to the choice of the specific

model for polytomous responses (the GRM, the SM, or the GPCM) is studied. In the third set of simulations, the multidimensional versions of the test statistics are studied. These versions will also be used in the real data example.

In all simulation studies reported below, all statistics were computed using a partition of the items into two subtests of equal size, that is, the first and second parts of the test. All tests were computed using a 5% significance level, and 100 replications were made for each branch of the study.

5.1. Simulation Study I: Influence of Estimation on Type I Error Rate and Power

5.1.1. Type I Error. The aim of this study is to assess whether the theoretical advantage of taking the effects of estimation into account pays off in practice. To keep the presentation simple, the 2PL model for dichotomous items (which is the common special case of the GRM, the SM, and the GPCM for $m = 1$) will be used for this first set of simulations. As can be seen in the report on the second set of simulations, the results generalize to polytomous items.

Sample sizes of $N = 100$, $N = 1000$, and $N = 4000$ were crossed with test lengths of $K = 20$, $K = 40$, and $K = 60$. For the test length $K = 20$, the item parameters were equal to $\beta_i = -2.00 + 0.20(i - 1)$, $i = 1, \ldots, 20$. For the test lengths $K = 40$ and $K = 60$ these values were repeated two and three times, respectively.

The UB statistic was computed in three conditions:

(1) using the true item and person parameters;
(2) using true item parameters and estimated ability parameters; and
(3) using both estimated item and ability parameters.

The LM statistic that takes the effects of the ability parameter estimates into account will be labeled LM_1, the statistic that also takes the effect of estimation of the item parameters into account will be labeled LM_2. LM_2 was computed in two conditions:

(1) using true item parameters and estimated ability parameters; and
(2) using both estimated item and ability parameters.

Finally, LM_2 was computed using estimates of the item and person parameters obtained as outlined in the previous section.

The results are shown in Table 1. In the third column of the table, it can be seen that when the true values for all parameters were used, the Type I error rates of the UB test were very close to the nominal significance level. This is as expected, because in this case the test does not involve estimation. In the next two columns, it can be seen that the Type I error rate decreased substantially when parameter estimates were used. The Type I error rates of the LM_1 test were close to 5% in both cases. It must be noted that the Type I error rate of LM_1 was slightly inflated for the case where $N = 100$ and the true item parameters were used. The effect vanished when estimates of the item parameters were used. Finally, the Type I error rate of the LM_2 test was not substantially closer to 5% than the Type I error rates of the LM_1 test. So explicitly taking the effects of the estimation of the item parameters into account did not result in a marked improvement.

5.1.2. Power of tests. Next, the power of the LM_1 and LM_2 tests was studied. The MML estimation procedure was run using the data of all simulees, both the aberrant and nonaberrant ones. In all simulations, 10% of the simulees were aberrant. The presence of the aberrant simulees did, of course, produce some bias in the parameter estimates, but this setup was considered realistic because in many situations it is not a priori known which respondents are aberrant, and which

TABLE 1.
Type I error rate for tests for constancy of theta.

		UB			LM_1		LM_2
	Theta						
	Beta	True	Estimated	Estimated	Estimated	Estimated	Estimated
K	N	True	True	Estimated	True	Estimated	Estimated
20	100	.042	.008	.007	.049	.041	.044
	1000	.042	.007	.007	.043	.042	.042
	4000	.042	.007	.007	.041	.041	.041
40	100	.049	.008	.005	.068	.050	.055
	1000	.047	.005	.005	.051	.045	.045
	4000	.047	.005	.005	.046	.045	.045
60	100	.047	.014	.004	.087	.050	.054
	1000	.048	.006	.005	.055	.051	.051
	4000	.048	.005	.005	.051	.051	.051

are not. Item parameters were equal to the item parameters in the previous study and the ability parameters θ were again drawn from a standard normal distribution. In all simulations, test length was equal to $K = 40$ and $K = 60$. The samples sizes were $N = 400$ and $N = 1000$. These sample sizes determine the precision of the estimates of the item parameters.

Two model violations were studied: different abilities for different subsets of items, and guessing on part of the items. An example of the first violation occurs in computerized adaptive testing when part of the items has been exposed and become known to the examinees. If the test administrator has a specific hypothesis about which items this may concern, this hypothesis can support a specific formulation of the LM test in terms of the composition of the subtests $g = 1, \ldots, G$. An example of the second violation, random guessing, occurs in educational testing when the examinees have several opportunities to take a test and use their first opportunity to test the ice without much studying.

For the study regarding different ability parameters, the model violation was imposed by assuming that the ability parameter in the last part of the test changed by a shift of 1.0. This shift

TABLE 2.
Power and Type I error of LM_1 and LM_2 to differences in ability and guessing.

			Difference in ability				Guessing			
			LM_1		LM_2		LM_1		LM_2	
K	N	Items infected	Type I error	power	Type I error	power	Type I error	power	Type I error	power
40	400	10	.06	.07	.06	.07	.05	.55	.05	.56
		20	.06	.15	.06	.15	.06	.85	.06	.84
	1000	10	.05	.07	.05	.07	.05	.55	.05	.55
		20	.05	.23	.05	.22	.06	.86	.06	.84
60	400	15	.06	.07	.06	.08	.05	.72	.05	.72
		30	.07	.20	.07	.21	.06	.83	.06	.84
	1000	15	.05	.09	.05	.09	.05	.72	.05	.72
		30	.06	.26	.06	.26	.06	.83	.06	.83

pertained either in the last half or the last quarter of the test. The results are displayed in Table 2 in the columns under the heading "Difference in ability." The column labeled "Items infected" gives the number of items where the model violation was imposed. The columns labeled "Type I error" give the proportion of the 90% nonaberrant simulees where the test was significant at 5%. The columns labeled "Power" give the proportion of significant tests for the 10% aberrant simulees. Note that the power of LM_1 and LM_2 for the detection of the imposed difference in ability is limited. The explanation is that the amount of information with respect to a person's ability and response behavior in 20 or 40 dichotomously scored items is quite low. There are main effects on power of the test length and of the number of items where the model violation was imposed. The explanation of the two effects is that the model violation can be better detected if the amount of aberrant data is larger. Further, a longer test length may provide more information for the estimation of the ability parameter. Comparing the cases $N = 400$ and $N = 1000$, it can be seen that the increased precision of the parameter estimates had a mild positive effect on the power when half of the items were infected. The power of LM_1 is comparable to the power of LM_2, so taking the effects of the estimation of item parameters into account in the definition of the test did not have a marked effect.

As a second example of a change in ability the simulees guess the response on part of the test. Guessing occurred for a quarter or a half of the test items. For the guessed items, the probability of a correct response was equal to 0.2. This violation is more severe than the previous one. In the previous simulation the item parameters were not changed, so for aberrant simulees the probabilities of correct responses were uniformly shifted for the affected part of the test. In the present simulation, guessing implies that the original items parameters lose their meaning, that is, all items are equally difficult. The results are shown in the last four columns of Table 2. Note that the power is now much larger than in the previous study. This is as expected, because the model violation is more serious.

5.2. Simulation Study 2: Robustness of Test Statistics

Above, it was already mentioned that the GRM, the SM, and the GPCM have response functions that are very close (Verhelst et al., 1997). Therefore, the models are often hard to distinguish. The second set of simulation studies addresses whether this also holds for the person fit test presented here. Because the previous simulation studies show that the effect of accounting for the estimation of the item parameters has little impact, in the remaining simulation studies only the LM test that takes the effect of the estimation of the ability parameter into account will be reported.

5.2.1. Type I Error Rate. First, the Type I error rate of the UB and LM tests for the GRM, the SM, and the GPCM was studied. The setup of the study was as follows. For all three models, ability parameters were drawn from a standard normal distribution. The parameters α_i were drawn from a log-normal distribution with a mean equal to zero and a standard deviation of 0.25. The item location parameters were fixed as follows. For the GPCM, the values of the category-bounds parameters for the items $i = 1, \ldots, 5$ were chosen as

$$\eta_{ij} = -2 + (i-1)/2 + (j-1)/2 \quad \text{for } j \leq 2$$

$$\text{and} \quad \eta_{ij} = -1.5 + (i-1)/2 + (j-1)/2 \quad \text{for } j > 2.$$

Note that the parameters of item 3 is located in such a way that the category bounds are located symmetric with respect to the standard normal ability distribution. The first two items are shifted to the left on the latent scale, the last two items are shifted to the right. The item parameters for the SM and the GRM were chosen in such a way that the item category response functions were

TABLE 3.
Type 1 error rates of UB and LM.

Generating model	Estimation model	K	UB	LM
GPCM	GPCM	20	.005	.051
		40	.006	.054
	SM	20	.007	.054
		40	.009	.055
	GRM	20	.006	.048
		40	.007	.050
SM	GPCM	20	.007	.056
		40	.008	.057
	SM	20	.004	.052
		40	.008	.057
	GRM	20	.006	.045
		40	.009	.050
GRM	GPCM	20	.004	.042
		40	.008	.048
	SM	20	.003	.030
		40	.005	.039
	GRM	20	.004	.052
		40	.009	.050

close to the response functions under the GPCM. To achieve this, data were generated under the GPCM, and using these data, the item parameters of the SM and the GRM were estimated using MML. These estimated values were then used as generating values for the simulation of data following the SM and the GRM. The sample size was always equal to 1000.

Table 3 gives the Type I error rate of the UB and LM tests. The column labeled "Generating model" gives the model used for generating the data. The model used for estimation and testing is given in the column labeled "Estimation model." The column labeled "K" gives the number of items, and the columns labeled UB and LM give the proportions of rejections at a 5% significance level. Analogous to the previous study, the results show that the empirical Type I error rate of the UB tests were much smaller than the nominal ones. The Type I error rate of the LM test attained its nominal value. Note that estimation with the "wrong model" still gives an acceptable Type I error rate.

5.2.2. Power of tests. The setup of the simulation study to the power of the tests in terms of test lengths and choice of parameters was analogous to the setup of the simulation study of the Type I error rate. However, a model violation was created by shifting the ability parameter for the second half of the test. The shifts were equal to either -0.5 or -1.0. As argued above, a realistic situation was created by obtaining MML estimates of the item parameters using the data of all 1000 simulees, both the aberrant and nonaberrant ones. In all simulations, 10% of the simulees were aberrant. Item parameters were equal to the item parameters in the previous study and the ability parameters θ were again drawn from a standard normal distribution. In all simulations, the samples size was always equal to 1000 and the test length was equal to $K = 20$ and $K = 40$. Therefore, the item parameter values defined above were repeated four and eight times.

The results for a shift of the ability parameter are shown in Table 4. The column labeled "Type I error" pertains to the proportion of incorrectly flagged respondents in the sample of 900 nonaberrant simulees, the column labeled "Power" refers to the proportion of correctly flagged

TABLE 4.
Detection of differences in ability.

K	Generating model	Estimation model	δ	Type I error	power
20	GPCM	GPCM	−0.5	.051	.131
			−1.0	.054	.358
		SM	−0.5	.054	.136
			−1.0	.056	.373
		GRM	−0.5	.049	.123
			−1.0	.051	.339
	SM	GPCM	−0.5	.045	.145
			−1.0	.047	.384
		SM	−0.5	.029	.148
			−1.0	.031	.420
		GRM	−0.5	.036	.140
			−1.0	.037	.376
	GRM	GPCM	−0.5	.042	.151
			−1.0	.045	.393
		SM	−0.5	.030	.151
			−1.0	.031	.426
		GRM	−0.5	.033	.146
			−1.0	.035	.386
40	GPCM	GPCM	−0.5	.052	.235
			−1.0	.057	.641
		SM	−0.5	.056	.240
			−1.0	.060	.645
		GRM	−0.5	.051	.225
			−1.0	.055	.616
	SM	GPCM	−0.5	.055	.289
			−1.0	.056	.647
		SM	−0.5	.034	.312
			−1.0	.036	.677
		GRM	−0.5	.043	.283
			−1.0	.045	.639
	GRM	GPCM	−0.5	.050	.294
			−1.0	.057	.662
		SM	−0.5	.035	.315
			−1.0	.039	.678
		GRM	−0.5	.040	.294
			−1.0	.043	.638

simulees in the sample of the 100 aberrant simulees. The Type I error rate of the LM test was relatively close to the nominal significance level. Also, the power of the test for the GRM was lower than the power of the test for the GPCM with the power of the test for the SM performing better than the other two models. In general, the LM tests have quite reasonable power to detect model violations of constancy of the ability parameter.

TABLE 5.
Percentage of agreement between the GRM, the SM, and the GPCM.

K	Generating model	Estimation model	δ	Normal	Aberrant
20	GPCM	SM	−0.5	98.5	81.4
			−1.0	98.3	84.7
		GRM	−0.5	99.1	80.8
			−1.0	99.0	83.2
	SM	SM	−0.5	99.4	62.5
			−1.0	99.1	74.6
		GRM	−0.5	99.3	71.4
			−1.0	99.3	79.1
	GRM	SM	−0.5	99.3	66.7
			−1.0	99.1	76.2
		GRM	−0.5	99.4	72.6
			−1.0	99.3	78.5
40	GPCM	SM	−0.5	98.3	82.2
			−1.0	98.2	88.1
		GRM	−0.5	98.9	81.9
			−1.0	98.9	87.0
	SM	SM	−0.5	99.0	65.1
			−1.0	99.1	78.2
		GRM	−0.5	99.1	72.8
			−1.0	99.1	82.2
	GRM	SM	−0.5	98.9	69.3
			−1.0	98.9	78.3
		GRM	−0.5	99.0	74.1
			−1.0	99.2	80.1

5.2.3. Agreement Between the Models. To investigate to what extent the three models give comparable results the degree of agreement to detect normal and aberrant responses between the three models in the previous simulation was determined.

Table 5 gives the results of the degree of agreement. The degree of agreement for normal simulees was higher than for aberrant simulees. It was not greatly improved as the number of items increased from 20 to 40. Note that the highest degree of agreement regarding the detection of aberrant simulees occurred with the GPCM as the generating model, and it was lowest when the generating model was the SM.

6. Simulation Study 3: The Test in a Multidimensional Setting

If a scale consists of more than one subscale, a person fit statistic pertaining to one of the subscales can be computed in two ways: using the estimate of the relevant ability obtained on the focused subscale alone, or using an estimate of all ability parameters pertaining to all subscales. The purpose of this simulation study is to assess the effect of using auxiliary information from other subscales on the power as a function of the correlation between the subscales. The expectation is that using auxiliary information pays off most when the correlation is high.

We considered three subscales, $t = 1, \ldots, 3$, associated with three ability parameters $\theta_1, \theta_2, \theta_3$. It was assumed that the three ability parameters had a three-variate normal distribution, so the model was a special case of the general model given by (14) in Appendix A. The variances of the three ability parameters were all equal to one. Every subscale had 16 items, so there were 48 items in total. The null-hypothesis was that the last eight items of the first subscale relate to the same ability parameter θ_1 as the first eight items. For all items, the number of categories equaled five, that is, $m = 4$, and the item parameters were the same as in the previous study. Every data set consisted of $N = 1000$ simulees. The design of the simulation studies was crossed with three facets:

(1) the values for the correlations between the ability dimensions: $\rho_{\theta\theta} = 0.4$ and $\rho_{\theta\theta} = 0.8$;
(2) the effect size of the model violation: $\delta = -0.5$ and $\delta = -1.0$; and
(3) using an estimate of θ_1 alone or using an estimate of all person parameters $\theta_1, \theta_2, \theta_3$ simultaneously, that is, an estimate that maximizes the likelihood (14) given in Appendix A.

The results are given in Table 6. The column marked "Generating model" gives the model used for generating the data and the column marked "Estimation model" gives the model used for the estimation and testing of the model. The tests were computed as indicated in Appendix A.

Note that for both the estimation procedures, the Type I error rates were again close to the nominal significance level if the correct estimation model was used. Using the wrong estimation model generally produced inflated Type I error rates, especially if the GPCM was used as an estimation model. An exception is the case where data generated using the SM were analyzed using the GRM. In that case, the Type I error rate was too low. There are no obvious explanations for these effects. Overall, the combination of the GPCM as a generating and estimation model produced the best combination of power and Type I error rate characteristics.

As expected, there were clear main effects of the effect size on the power of the test. The effect of the size of the correlation was negligible. Concurrently estimating the person parameters $\theta_1, \theta_2, \theta_3$ had a systematic positive effect on the power, but this effect was very small.

7. An Empirical Example

Data from the NEO Personality Inventory data were used to get an impression of the degree of agreement between the three IRT models in a real data set. The NEO Personality Inventory is a personality test designed to provide a general description of normal personality that is relevant to clinical, counseling, and educational situations. It is based on the Five-Factor model of personality (Costa & McCrae, 1992). The NEO Personality Inventory consists of five broad domains and for each of these domains, six facet scores or subfactors have been developed to provide specific levels of information. Each of the six facets is measured by eight items. All items are rated on a five-point scale. Three validity items are also included.

The empirical example presented here pertains to the neuroticism domain. To obtain subscales of reasonable length, pairs of two facets within the domain were grouped together on the basis of their correlation. That is, facets with the highest mutual correlation were grouped together. So three subscales of 16 items each were analyzed. Note that this setup is analogous to the setup used in the second simulation. Further, also in the present analysis, the null-hypothesis tested was that the last eight items of the first subscale related to the same ability parameter as the first eight items. Note that these two groups of eight items related to two different facets. The test was performed in two versions: one version using the parameter estimates of the first subscale only, and the other using the item parameter estimates of all three subscales. The MML estimates of the item parameters for the unidimensional model were computed using Multilog (Thissen, Chen, & Bock, 2003) and the MML estimates of the item and latent covariance

TABLE 6.
Power of the LM as a function of the correlation between the subscales.

Generating model	Estimation model	δ	$\rho_{\theta\theta}$	Estimation of θ_1		Estimation of $\theta_1, \theta_2, \theta_3$	
				Type I error	Power	Type I error	Power
GPCM	GPCM	−0.5	.40	.047	.138	.047	.142
			.80	.050	.138	.049	.150
		−1.0	.40	.047	.415	.047	.425
			.80	.050	.414	.049	.450
	SM	−0.5	.40	.068	.070	.071	.071
			.80	.071	.068	.052	.078
		−1.0	.40	.068	.209	.071	.213
			.80	.071	.212	.052	.259
	GRM	−0.5	.40	.040	.047	.046	.037
			.80	.043	.043	.047	.035
		−1.0	.40	.040	.112	.046	.087
			.80	.043	.112	.047	.105
SM	GPCM	−0.5	.40	.130	.228	.132	.232
			.80	.138	.224	.132	.224
		−1.0	.40	.130	.397	.132	.402
			.80	.138	.409	.129	.421
	SM	−0.5	.40	.051	.098	.052	.100
			.80	.055	.097	.052	.103
		−1.0	.40	.051	.230	.052	.235
			.80	.055	.230	.052	.260
	GRM	−0.5	.40	.025	.060	.024	.049
			.80	.025	.058	.026	.045
		−1.0	.40	.025	.151	.024	.123
			.80	.025	.148	.026	.127
GRM	GPCM	−0.5	.40	.105	.164	.116	.189
			.80	.108	.165	.104	.186
		−1.0	.40	.105	.291	.116	.326
			.80	.108	.277	.104	.325
	SM	−0.5	.40	.072	.075	.066	.083
			.80	.075	.077	.064	.101
		−1.0	.40	.072	.140	.066	.169
			.80	.075	.139	.064	.203
	GRM	−0.5	.40	.043	.053	.043	.054
			.80	.046	.054	.043	.057
		−1.0	.40	.043	.112	.043	.118
			.80	.046	.114	.043	.135

TABLE 7.
Observed and latent correlations.

Observed			Latent-GPCM		
1.000	0.694	0.596	1.000	0.848	0.758
0.694	1.000	0.585	0.848	1.000	0.778
0.596	0.585	1.000	0.758	0.778	1.000
Latent-GRM			Latent-SM		
1.000	0.864	0.767	1.000	0.861	0.772
0.864	1.000	0.797	0.861	1.000	0.802
0.767	0.797	1.000	0.772	0.802	1.000

parameters of the multidimensional model were computed using dedicated software developed by the authors.

Table 7 gives the manifest correlations between the total scores on the three subscales (upper right-hand matrix under label "Observed") and the MML estimated latent correlations for the three IRT models (the matrices under the labels "Latent-GPCM," "Latent-GRM" and "Latent-SM," respectively). Note that, as expected, the manifest correlations are attenuated, that is, they are lower than the latent correlations. Note that the pattern of the correlations is similar for each of the three models.

Given the MML estimates of the item parameters and the covariance matrices, the θ-parameters were estimated by maximum likelihood, and the fit statistics were computed. Table 8 gives a cross-tabulation of the persons identified as aberrant and nonaberrant under the three

TABLE 8.
Agreement between the SM, the GRM, and the GPCM and between the unidimensional and multidimensional models.

Unidimensional item-parameter estimates												
		SM				GRM				SM		
		+	−			+	−			+	−	
GPCM	+	.128	.061	GPCM	+	.074	.038	GRM	+	.108	.132	
	−	.083	.728		−	.096	.792		−	.081	.679	
Kappa = 0.55				Kappa = 0.45				Kappa = 0.37				
Multidimensional item-parameter estimates												
		SM				GRM				SM		
		+	−			+	−			+	−	
GPCM	+	.109	.061	GPCM	+	.061	.049	GRM	+	.098	.103	
	−	.089	.741		−	.077	.813		−	.088	.711	
Kappa = 0.50				Kappa = 0.42				Kappa = 0.39				
Agreement between the unidimensional and multidimensional models												
		GPCM				GRM				SM		
		UNI				UNI				UNI		
		+	−			+	−			+	−	
MULTI	+	.112	.019	MULTI	+	.093	.065	MULTI	+	.074	.099	
	−	.048	.821		−	.051	.791		−	.055	.772	
Kappa = 0.73				Kappa = 0.55				Kappa = 0.40				

models for the unidimensional and multidimensional cases, respectively. The "plus" and "minus" signs in all tables refer to persons flagged as aberrant and normal, respectively. Coefficient Kappa was used as a measure of agreement, the values are given at the bottom of each section of the table. The values of Kappa indicate that the degree of agreement between the models is moderate. The largest agreement is between the GPCM and the GRM and between the GPCM and the SM, respectively. This holds for both parameter estimation procedures. The degree of agreement between the unidimensional and multidimensional versions of the models is presented in the last panel of Table 8. It can be seen that the agreement between the unidimensional and multidimensional models was highest for the GPCM.

8. Discussion

An LM test statistic for assessing person fit was introduced where the effects of estimation of the item and person parameters are explicitly taken into account. Simulation studies showed that taking the effects of the estimation of the person parameter has a substantial effect on the precision of the Type I error rate and on the power. The simulation studies also showed that accounting for estimation of item parameters had little additional effect. The other goal was to compare the robustness of person fit of tests across three IRT models for polytomous items: the GPCM, the SM, and the GRM. Simulation studies for the unidimensional cases of these models showed that the Type I error rate was close to its nominal value, both if the correct and wrong model were used for estimation and computation of the statistics.

The test statistics were generalized to multidimensional versions of the GRM, the SM, and the GPCM. Simulation studies showed that the power of the tests was quite acceptable. The conclusion regarding the robustness of the person fit tests across three IRT models for polytomous items did not hold here: in the multidimensional case, using the "wrong model" often resulted in an inflation of Type I error. The conclusion is that searching for the best fitting model for the majority of the persons before searching for aberrant persons remains the best strategy. Of course, in many situations, searching for item fit across persons and searching for person fit across items may be an iterative process.

For the multidimensional versions of the GRM, the SM, and the GPCM, the effect of using auxiliary information obtained from concurrent estimation of the ability parameters of all subscales on the power of the fit tests was assessed. Results showed that the effect of this auxiliary information was small, and also the main effect of the size of the correlation between the subscales was very small. In an empirical example, data from the NEO Personality Inventory-Revised were used to get an impression of the degree agreement between the three IRT models in a real situation. Results showed that the degree of agreement between the three models was only moderate. Also, the degree of agreement between the unidimensional and multidimensional versions of the models was moderate.

The final remarks of this discussion pertain to the relation of the proposed test to other tests of person fit. An essential feature of the test is that a model violation is translated into an explicit alternative model by introducing extra parameters that represent the model violation. The test then amounts to the evaluation whether the additional parameters are equal to zero. This distinguishes the present approach from the use of more general tests such as the test based on the likelihood statistic l_z by Drasgow et al. (1985) and the test based on the Pearson-type W-statistic by Wright and Stone (1979). These tests have an unspecified general alternative, so they have a more global nature. The present approach allows for targeting specific model violations, and in Appendix A it is shown that the approach presented here can also be used to target a number of other model violations rather than the one studied here in detail.

It may be of interest to contrast the approach to adjusting for the estimation of θ with the approach by Snijders (2001). In the framework of dichotomously scored items, Snijders (2001) notes that statistics such as l_z and W have a form

$$W(\theta) = \sum_i (X_i - p_i(\theta))w_i(\theta).$$

For dichotomously scored items, it holds that $X_i^2 = X_i$, so also statistics of the form

$$\sum_i (X_i - p_i(\theta))^2 v_i(\theta)$$

belong to this class. For a class of estimators that includes maximum likelihood and Bayesian modal estimators, adjusting the weights $w_i(\theta)$ and dividing results in a statistic $W(\theta)$ that has an asymptotic standard normal distribution if the person parameter is estimated. Consequently, the squared statistic has an asymptotic χ^2-distribution with one degree of freedom. To align the approach of the present paper with the approach by Snijders (2001) two elements are needed: first, for dichotomous items, Snijders' approach must be generalized to test statistics with an asymptotic χ^2-distribution with more than one degree of freedom; and, second, the approach should be generalized to polytomous responses, which involves taking the dependencies of the response variables X_{ij} $(j = 1, \ldots, m)$ within an item i into account. These topics, however, are beyond the scope of the present paper and remain points for further study.

Appendix A

Detailed characterization of the test statistics.

A detailed characterization of the LM tests will be given for the multidimensional versions of the GPCM, the SM, and the GRM; the unidimensional versions follow directly as a special case. A general formulation for all statistics considered above is given by

$$LM = \mathbf{h}(\eta_2)^t \left(\Sigma_{22} - \Sigma_{12}^{'} \Sigma_{11}^{-1} \Sigma_{12}^{-1}\right)^{-1} \mathbf{h}(\eta_2), \tag{13}$$

where $\mathbf{h}(\eta_2)$, Σ_{11}, Σ_{12}, and Σ_{22} are first- and second-order derivatives as defined in (5), (6), and (7).

The log-likelihood of a response pattern $\mathbf{x}$ for the general alternative model for multidimensional data is given by

$$\log L(\theta, \delta) = \sum_{i=1}^{k} \sum_{j=0}^{m} x_{ij} \log P_{ij}(\eta_i) + \log g\left(\theta | \Sigma_\theta\right), \tag{14}$$

where $g\left(\theta | \Sigma_\theta\right)$ is the multivariate normal density with a mean set equal to zero to identify the latent scaler, and covariance matrix Σ_θ. Further, $P_{ij}(\eta_i)$ is the probability of a response on item i in category j, given by either the GPCM, the SM, or the GRM. In general, this probability depends on $\eta_i = \alpha_i^t \theta + \mathbf{z}_i^t \delta$, where $\alpha_i^t = (\alpha_{i1}, \ldots, \alpha_{iQ})$, $\theta^t = (\theta_1, ...\theta_Q)$, $z_i^t = (z_{i1}, \ldots, z_{iG})$ and $\delta^t = (\delta_1, \ldots, \delta_G)$. The parameters δ are the parameters added to the restricted model (i.e., the IRT model) to obtain a more general model, so the null-hypothesis tested is $\delta = 0$. In the application considered above, the covariate z_i is an indicator function specifying to which of the G subtests item i belongs. However, this definition also includes other applications. For instance, Klauer (1995) and Glas (1999) show that violation of local independence can be modeled (and, therefore, tested) by assuming that z_i is the response on one or more previous items. Further, the framework presented here can also be applied to evaluate person fit to the testlet

model, because Glas, Wainer, and Bradlow (2000) note that the testlet model (an IRT model that takes the dependencies of items clustered in testlets into account) can be seen as a full-information factor analysis model as given by (14) (Gibbons & Hedeker, 1992). A detailed study of the power and practical usefulness of such generalizations is however beyond the scope of the present paper.

To derive the specific expressions for (13) for the IRT models considered here, we define

$$d_{ij} = \frac{\partial \log P_{ij}(\eta_i)}{\partial \eta_i}$$

and

$$D_{ij} = \frac{\partial^2 \log P_{ij}(\eta_i)}{\partial \eta_i^2}.$$

Using the chain rule, it directly follows that

$$\frac{\partial \log P_{ij}(\eta_i)}{\partial \theta_q} = \frac{\partial \log P_{ij}(\eta_i)}{\partial \eta_i}\frac{\partial \eta_i}{\partial \theta_q} = \alpha_{iq} d_{ij},$$

and

$$\frac{\partial \log P_{ij}(\eta_i)}{\partial \delta_g} = \frac{\partial \log P_{ij}(\eta_i)}{\partial \eta_i}\frac{\partial \eta_i}{\partial \delta_g} = z_{ig} d_{ij}.$$

Analogously,

$$\frac{\partial^2 \log P_{ij}(\eta_i)}{\partial \theta_q \partial \theta_{q'}} = \alpha_{iq}\alpha_{iq'} D_{ij},$$

$$\frac{\partial^2 \log P_{ij}(\eta_i)}{\partial \theta_q \partial \delta_g} = \alpha_{iq} z_{ig} D_{ij},$$

$$\frac{\partial^2 \log P_{ij}(\eta_i)}{\partial \delta_g \partial \delta_{g'}} = z_{ig} z_{ig'} D_{ij},$$

and

$$\frac{\partial \log g\,(\theta|\Sigma_\theta)}{\partial \theta} = -\Sigma^{-1}\theta.$$

These expressions can be used to obtain the first-order derivatives in (13), which are given by

$$\mathbf{h}(\eta_2) = \frac{\partial \log L(\theta, \delta)}{\partial \delta} = \sum_i \sum_{j=0}^{m} x_{ij} d_{ij},$$

and the second-order derivatives in (13), which are given by

$$\Sigma_{11} = -\frac{\partial^2 \log L(\theta, \delta)}{\partial \theta \partial \theta'} = -\sum_i \sum_{j=0}^{m} \frac{x_{ij}\partial^2 \log P_{ij}(\eta_i)}{\partial \theta \partial \theta'} + \frac{\partial^2 \log g\,(\theta|\Sigma_\theta)}{\partial \theta \partial \theta'},$$

$$\Sigma_{22} = -\frac{\partial^2 \log L(\theta, \delta)}{\partial \delta \partial \delta'} = -\sum_i \sum_{j=0}^{m} \frac{x_{ij}\partial^2 \log P_{ij}(\eta_i)}{\partial \delta \partial \delta'},$$

$$\Sigma_{12} = -\frac{\partial^2 \log L(\theta, \delta)}{\partial \theta \partial \delta'} = \sum_i \sum_{j=0}^{m} \frac{x_{ij}\partial^2 \log P_{ij}(\eta_i)}{\partial \theta \partial \delta'}.$$

Finally, the specific expressions for d_{ij} and D_{ij} for the three IRT models need to be derived.

8.1. First- and Second-Order Derivatives for the Graded Response Model

We introduce a concise notation $P_{ij} = \Psi_{ij} - \Psi_{i(j+1)}$ with $\Psi_{ij} = \Psi\left(\eta_i - \beta_{ij}\right)$, $\Psi(\cdot)$ is the logistic function defined by (1), $\Psi_{i0} = 1$, and $\Psi_{i(m+1)} = 0$. Further, $\Psi_{ij}^{'}$ and $P_{ij}^{'}$ are first-order derivatives with respect to η_i. Then

$$d_{ij} = \frac{\partial \log P_{ij}(\eta_i)}{\partial \eta_i} = [1 - \Psi_{ij} - \Psi_{i(j+1)}]$$

for $j = 0, \ldots, m$. Note that for $j = 0$ we have $d_{ij} = -\Psi_{i(j+1)}$, and for $j = m$ we have $d_{tij} = -(1 - \Psi_{ij})$.

For the second-order derivatives of the log-likelihood we obtain

$$D_{ij} = \frac{\partial^2 \log P_{ij}(\eta_i)}{\partial \eta_i^2} = -[\Psi_{ij}(1 - \Psi_{ij}) + \Psi_{i(j+1)}(1 - \Psi_{i(j+1)})]$$

for $j = 0, \ldots, m$. Note that for $j = 0$ we have $d_{tij} = -\Psi_{i(j+1)}(1 - \Psi_{i(j+1)})$, and for $j = m$ we have $d_{tij} = -\Psi_{ij}(1 - \Psi_{ij})$.

8.2. First- and Second-Order Derivatives for the Sequential Model

We introduce a concise notation

$$P_{ij} = \left[\prod_{h=1}^{j} \Psi_{ih}\right][1 - \Psi_{i(j+1)}]$$

where the product from $j = 1$ to 0 is assumed to result in unity, and $1 - \Psi_{i(m+1)} = 1$. Then

$$d_{ij} = \left[\sum_{h=1}^{j}(1 - \Psi_{ih}) - \Psi_{i(h+1)}\right].$$

The second-order derivative is given by

$$D_{ij} = \frac{\partial^2 \log P_{ij}(\eta_i)}{\partial \eta_i^2} = -\sum_{h=1}^{j+1} \Psi_{ih}(1 - \Psi_{ih}).$$

8.3. First- and Second-Order Derivatives for the Generalized Partial Credit Model

We introduce a concise notation $P_{ij} = P_{ij}(\eta_i)$. Then

$$d_{ij} = \frac{\partial \log P_{ij}(\eta_i)}{\partial \eta_i} = \left(j - \sum_{h=1}^{m} h P_{ih}\right)$$

and

$$D_{ij} = \frac{\partial^2 \log P_{ij}(\eta_i)}{\partial {\eta_i}^2} = -\sum_{j=0}^{m} j P_{ij}\left[j - \sum_{h=1}^{m} h P_{ih}\right].$$

References

Aitchison, J., & Silvey, S.D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics, 29,* 813–828.

Andersen, E.B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123–140.
Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM-algorithm. *Psychometrika, 46,* 443–459.
Bock, R.D., Gibbons, R.D., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement, 12,* 261–280.
Costa, P.T., Jr., & McCrae, R.R. (1992). Normal personality assessment in clinical practice: The NEO personality inventory. *Psychological Assessment*, *4*, 5–13.
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.
Gibbons, R.D., & Hedeker, D.R. (1992). Full-information bi-factor analysis. *Psychometrika, 57,* 423–436.
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.
Glas, C.A.W., & Suárez Falcón, J.C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27,* 87–106.
Glas, C.A.W., Wainer, H., & Bradlow. (2000). MML and EAP estimates for the testlet response model. In W.J. van der Linden, & C.A.W. Glas (Eds.), *Computer adaptive testing: Theory and practice* (pp. 271–287). Boston: Kluwer-Nijhoff.
Jansen, M.G.H., & Glas, C.A.W. (2005). Checking the assumptions of Rasch's model for speed tests. *Psychometrika, 70,* 671–684.
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.
Klauer, K.C., & Rettig, K. (1990). An approximately standardized person test for assessing consistency with a latent trait model. *British Journal of Mathematical and Statistical Psychology, 43,* 193–206.
Levine, M.V., & Rubin, D.B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269–290.
Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.
McDonald, R.P. (1997). Normal-ogive multidimensional model. In W.J. van der Linden, & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257–269). New York: Springer-Verlag.
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.
Mellenbergh, G.J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19,* 91–100.
Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–195.
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.
Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.
Nering, M.L. (1995). The distribution of person fit statistics using true and estimated person parameters. *Applied Psychological Measurement, 19*, 121–129.
Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50–64.
Rao, C.R. (1947). Large sample tests of statistical hypothesis concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society, 44,* 50–57.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.
Reckase, M.D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W.J. van der Linden, & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer-Verlag.
Reise, S.P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement*, *19*, 213–229.
Samejima, F. (1969). *Estimation of latent ability using a pattern of graded scores*. Psychometrika Monograph Supplement, No. 17. Greensboro, NC: Psychometric Society.
Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38,* 203–219.
Sijtsma, K., & Meijer, R.R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66,* 191–207.
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. (2001). Asymptotic distribution of person-fit statistics with estimated person parameter. *Psychometrika, 66,* 331–342.
Sörbom, D. (1989). Model modification. *Psychometrika, 54,* 371–384.

Tatsuoka, K.K. (1984). Caution indices based on item response theory. *Psychometrika, 49*, 95–110.
Thissen, D., Chen, W.-H., & Bock, R.D. (2003). *Multilog*. Lincolnwood, IL: Scientific Software International.
Tsutakawa, R.K., & Johnson, J.C. (1990). The effect of uncertainty of item parameter-estimation on ability estimates. *Psychometrika, 55,* 371–390.
Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43,* 39–55.
van Krimpen-Stoop, E.M.L.A., & Meijer, R.R. (2002). Detection of person misfit in computerized adaptive tests with polytomous items. *Applied Psychological Measurement, 26,* 164–180.
Verhelst, N.D., Glas, C.A.W., & de Vries, H.H. (1997). A steps model to analyze partial credit. In W.J. van der Linden, & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 123–138). New York: Springer-Verlag.
von Davier, M., & Molenaar, I.W. (2003). A person-fit index for polytomous Rasch models, latent class models, and their mixture generalizations. *Psychometrika, 68,* 213–228.
Wright, B.D., & Linacre, J.M. (1992). *Bigsteps* (Computer software). Chicago: MESA Press.
Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis*. (Computer software). Chicago: Mesa Press.
Wright, B.D., & Stone, M.H. (1979). *Best test design*. Chicago: 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.

Manuscript received 25 April 2003
Final version received 24 April 2006
Published Online Date: 17 NOV 2006