LSAC RESEARCH REPORT SERIES

Likelihood-based Statistics for Validating Continuous Response Models

Cees A. W. Glas
Oksana Korobko

University of Twente, Enschede, The Netherlands

Law School Admission Council
Research Report 05-03
October 2005

A Publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation whose members are more than 200 law schools in the United States and Canada. It was founded in 1947 to coordinate, facilitate, and enhance the law school admission process. The organization also provides programs and services related to legal education. All law schools approved by the American Bar Association (ABA) are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also included in the voting membership of the Council.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

The statistical theory of estimating and testing item response theory (IRT) models for items (questions) with discrete (correct or incorrect) responses has been thoroughly developed (recall that IRT is a mathematical model that is typically used to analyze test data). In contrast, the theory for IRT models for items with continuous responses has hardly received any attention. This omission is mainly due to the fact that, so far, the continuous response format has hardly been used by the testing industry. An exception may be the rating scale item format, where a respondent marks a position on a line to express his or her opinion about a topic. Recently, continuous responses have attracted interest as complementary information to accompany discrete item responses. One may think of the response time needed to answer an item in a computerized adaptive testing situation or of computer ratings of tasks performed in a simulated environment as continuous responses.

In the present report, an existing model for the analysis of continuous responses was extended to include a procedure for estimating the parameters in the model. Tests for evaluating the fit of the model were successfully evaluated. These tests can be used to detect problematic items and violations of assumptions of the model. The tests were also shown to have excellent control of their false positive error rate, as well as excellent ability to detect true effects.

Abstract

The theory for the estimation and testing of item response theory (IRT) models for items with discrete responses is by now very thoroughly developed. In contrast, the estimation and testing theory for IRT models for items with continuous responses has hardly received any attention. This is mainly due to the fact that the continuous response format is seldom used. An exception may be the so-called analogous-scale item format where a respondent marks the position on a line to express his or her opinion about a topic. Recently, continuous responses have attracted interest as covariates accompanying discrete responses. One may think of the response time needed to answer an item in a computerized adaptive testing situation. In the present report, the theory of estimating and testing a model for continuous responses, the model proposed by Mellenbergh in 1994, is developed in a marginal maximum likelihood framework. It is shown that the fit to the model can be evaluated using Lagrange multiplier tests. Simulation studies show that these tests have excellent properties in terms of control of Type I error rate and power.

Introduction

Item response theory (IRT) models are stochastic models for two-way data, for instance, the responses of students to items. An essential feature of these models is parameter separation, that is, the influences of the items and students on the responses are modeled by distinct sets of parameters. IRT provides the theoretical underpinning for computer adaptive testing, the use of incomplete assessment designs, equating and linking of assessments, evaluation of differences between groups, and applications to multilevel analyses as used in school effectiveness research. Most applications of IRT models pertain to categorical data (Samejima, 1969; Masters, 1982; Bock, 1972). However, situations may also arise where the responses of students to items are continuous. One might think of a computer adaptive test where the response time is recorded with every response to the actual assessment item. In the present report, the IRT model for continuous data proposed by Mellenbergh (1994) will be elaborated. The model will first be generalized to allow for a multidimensional proficiency structure, and then a marginal maximum likelihood (MML) estimation procedure and a method for testing model fit will be proposed. Finally, a number of simulation studies will be conducted to assess the Type I error rate and the power of the proposed tests.

Consider a two-dimensional data matrix $\boldsymbol{X}$ with entries x_{nk}, for $n = 1, ..., N$, and $k = 1, ..., K$. The matrix contains the response of students to items. It is assumed that the response of the student n on the item k is normally distributed, that is

$$P(x_{nk} \mid \boldsymbol{\theta}_n, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \cdot \exp\left(-\frac{(x_{nk} - \tau_{nk})^2}{2\sigma_k^2} \right). \tag{1}$$

The expectation of the item response is a linear function of the explanatory variables

$$\begin{aligned} \tau_{nk} &= \sum_{h=1}^{H} \alpha_{kh}\theta_{nh} - \beta_k \\ &= \boldsymbol{\alpha}_k^t \boldsymbol{\theta}_n - \beta_k \end{aligned} \tag{2}$$

where $\boldsymbol{a}_k$ is a vector of parameters $\alpha_{k1}, ..., \alpha_{kh}, ..., \alpha_{kH}$ which are usually called factor loadings and β_k is a location parameter. Further, $\boldsymbol{\theta}_n = (\theta_{n1}, ..., \theta_{nh}, ..., \theta_{nH})$ is the H-dimensional proficiency parameter of student n. In this report, we assume that the density of $\boldsymbol{\theta}_n$ is described by the normal distribution with average value $\boldsymbol{\mu}_\theta$ and the covariance matrix $\boldsymbol{\Sigma}_\theta$, which is given by $g(\boldsymbol{\theta}_n; \boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)$. The model is partially identified by the restriction $\boldsymbol{\mu}_\theta = 0$. Additional restrictions must be imposed to completely identify the model. This will be returned to below. Further, since the main application envisioned here is modeling response times, we assume that $\sigma_k^2 = 1$, for all k. That is, we assume that all the observed responses have the same scale.

In the case of discrete responses, the data are the response patterns of the students, and these counts are seldom, if ever, transformed. In the case of continuous responses, transformations can be applied to the responses. For instance, if the model given by (1) is used to analyze response times, the observations x_{nk} should be the logarithms of the response times.

Estimation

Preliminaries

Let $\boldsymbol{\xi}$ be a vector of model parameters, that is, $\boldsymbol{\xi}$ consists of the vectors $\boldsymbol{a}$, $\boldsymbol{\beta}$, $\boldsymbol{\mu}_\theta$, and $vec(\boldsymbol{\Sigma}_\theta)$, where $vec(\boldsymbol{\Sigma}_\theta)$ is a vector with all elements of $\boldsymbol{\Sigma}_\theta$. The marginal log-likelihood function can then be written as

$$\log L(\boldsymbol{\xi}, \mathbf{X}) = \sum_n \log Pr(\mathbf{x}_n; \boldsymbol{\xi}) \tag{3}$$

where $\mathbf{x}_n$ is the response pattern of the student n. The MML estimation equations require the vector of derivatives of the log-likelihood function. These first order derivatives can be derived using Fisher's identity (Louis, 1982; Glas, 1992). In the framework of IRT, Fisher's identity is given by

$$\mathbf{h}(\boldsymbol{\xi}) = \frac{\partial}{\partial \boldsymbol{\xi}} \log L(\boldsymbol{\xi}, \mathbf{X}) = \sum_n E(b_n(\boldsymbol{\xi}) \mid \mathbf{x}_n, \boldsymbol{\xi}), \tag{4}$$

where the expectation is with respect to the posterior expectation $P(\boldsymbol{\theta}_n \mid \mathbf{x}_n, \boldsymbol{\xi})$. Further,

$$b_n(\boldsymbol{\xi}) = \frac{\partial}{\partial \boldsymbol{\xi}} \log Pr(\mathbf{x}_n, \boldsymbol{\theta}_n; \boldsymbol{\xi}) = \frac{\partial}{\partial \boldsymbol{\xi}} [\log Pr(\mathbf{x}_n \mid \boldsymbol{\theta}_n, \alpha, \beta) + \log g(\boldsymbol{\theta}_n; \boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)]. \tag{5}$$

Notice that the derivative is a sum of the logarithm of the probability of the response pattern and the logarithm of the density of the student ability parameter. The power of Fisher's identity is that the derivatives are very easy to derive, while the derivation of $\mathbf{h}(\boldsymbol{\xi})$ is a cumbersome enterprise. Moreover, direct derivation of the matrix of second-order derivatives needed for the computation of the standard errors of the estimates is even more demanding. However, using Fisher's identity repeatedly, Louis (1982) shows that the Fisher information matrix

$$H(\boldsymbol{\xi}, \boldsymbol{\xi}) = -\frac{\partial^2 L(\boldsymbol{\xi}, \mathbf{X})}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}'} \tag{6}$$

is the sum over students n of terms (see Appendix A)

$$-E(B_n(\boldsymbol{\xi}, \boldsymbol{\xi}) \mid \mathbf{x}_n, \boldsymbol{\xi}) - E(b_n(\boldsymbol{\xi}) b_n(\boldsymbol{\xi})' \mid \mathbf{x}_n, \boldsymbol{\xi}) + E(b_n(\boldsymbol{\xi}) \mid \mathbf{x}_n, \boldsymbol{\xi})\, E(b_n(\boldsymbol{\xi}) \mid \mathbf{x}_n, \boldsymbol{\xi})', \tag{7}$$

where

$$B_n(\boldsymbol{\xi}, \boldsymbol{\xi}) = \frac{\partial^2 \log Pr(\mathbf{x}_n, \boldsymbol{\theta}_n; \boldsymbol{\xi})}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}'}.$$

Glas (1998, 1999) and Glas and Suarez-Falcon (2003) show that in the case of the two- and three- parameter logistic model and the nominal response model, the second derivatives can be approximated by

$$H(\boldsymbol{\xi}, \boldsymbol{\xi}) \approx \sum_n E(b_n(\boldsymbol{\xi}) \mid \mathbf{x}_n, \boldsymbol{\xi})\, E(b_n(\boldsymbol{\xi}) \mid \mathbf{x}_n, \boldsymbol{\xi})'. \tag{8}$$

Below, the precision of this approximation will be evaluated empirically.

The exact expressions for the information matrix derived using (7) and (8) are given in Appendix A.

Application to the IRT Model for Continuous Responses

In this article we consider no missing data. The marginal likelihood function is

$$L(\boldsymbol{\xi} \mid x) = \prod_n \int \ldots \int \prod_k p(x_{nk} \mid \boldsymbol{\theta}_n, \boldsymbol{\alpha}_k, \beta_k)\, g(\boldsymbol{\theta}_n \mid \boldsymbol{\Sigma}_\theta) d\boldsymbol{\theta}_n, \tag{9}$$

where $\boldsymbol{\xi}$ is the ensemble of item and population parameters $\boldsymbol{\xi} = (\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\mu}_\theta, vec(\boldsymbol{\Sigma}_\theta))'$. Here, $g(\boldsymbol{\xi}_n, \boldsymbol{\theta}_n \mid \boldsymbol{\Sigma})$ is the density of $\boldsymbol{\xi}_n$ and $\boldsymbol{\theta}_n$, which assumed to follow a multivariate normal distribution with mean vector $\mathbf{0}$ and variance-covariance Σ.

The logarithm of the marginal likelihood function is

$$\log L(\boldsymbol{\xi} \mid x) = \sum_n \log \int \ldots \int \prod_k p(x_{nk})\, g(\boldsymbol{\theta}_n \mid \boldsymbol{\Sigma}_\theta) d\boldsymbol{\theta}_n. \tag{10}$$

The maximization procedure provides the set of equations which can be used to estimate the values of parameters. To be more specific, the solution of the equation $\partial \log L(\boldsymbol{\xi} \mid x)/\partial \boldsymbol{\xi} = 0$ is the maximum likelihood estimate of $\boldsymbol{\xi}$.

Applying Fisher's identity, from maximization of the marginal likelihood with respect to the covariance matrix, we arrive at

$$\boldsymbol{\Sigma}_\theta = \frac{1}{N} \sum_n E(\boldsymbol{\theta}_n \boldsymbol{\theta}_n^t \mid x_n, \boldsymbol{\xi}) \tag{11}$$

where

$$E(\boldsymbol{\theta}_n \boldsymbol{\theta}_n^t \mid x_n, \boldsymbol{\xi}) = \int \ldots \int \boldsymbol{\theta}_n \boldsymbol{\theta}_n^t f[\boldsymbol{\theta}_n \mid \mathbf{x}_n, \boldsymbol{\Sigma}_\theta] d\boldsymbol{\theta}_n$$

and the posterior density has a form

$$f[\boldsymbol{\theta}_n \mid \mathbf{x}_n, \boldsymbol{\Sigma}_\theta] = \frac{\prod_k p(x_{nk}) g(\boldsymbol{\theta}_n \mid \Sigma_\theta).}{\int \ldots \int \prod_k p(x_{nk}) g(\boldsymbol{\theta}_n \mid \boldsymbol{\Sigma}_\theta) d\boldsymbol{\theta}_n}. \tag{12}$$

In the same way, maximization with respect to β_k results in

$$\sum_n x_{nk} = \sum_n E(\tau_{nk} \mid \mathbf{x}_n, \boldsymbol{\xi}). \tag{13}$$

Finally, for α_{kh}, we obtain

$$\sum_n x_{nk} E(\theta_{nh} \mid \mathbf{x}_n, \boldsymbol{\xi}) = \sum_n E(\tau_{nk} \theta_{nh} \mid \mathbf{x}_n, \boldsymbol{\xi}). \tag{14}$$

All these expressions can be solved simultaneously. In practice, this is done by Newton-Raphson, expectation-maximization (EM) algorithm, or a combination of both (Bock & Aitkin, 1981). Below, we further comment on the use of the EM algorithm.

Identification of the Model

To identify the model the restriction $\boldsymbol{\mu}_\theta = 0$ was imposed. The model can be identified further in two ways. The first approach requires setting the covariance matrix to the identity matrix and introducing the constrains $\alpha_{jq} = 0$; $j = 1, \ldots, q-1$ and $q = j+1, \ldots, Q$. The latent ability dimensions are independent of each other. The first item loads on the first dimension only. The second item loads on the first two dimensions only, and so on, until item Q loads on the first $Q-1$ dimensions. All other items load on all dimensions.

The second approach to identify the model is setting the mean equal to the zero and considering the covariance matrix as a parameter of proficiency distribution that must be estimated. Further, the model is identified by imposing the restrictions, $\alpha_{jq} = 1$, if $j = q$, and $\alpha_{jq} = 0$, if $j \neq q$, for $j = 1, \ldots, Q$ and $q = 1, \ldots, Q$. So, here, the first item defines the first dimension, the second item defines the second dimension, and the third

item defines the third dimension. The covariance matrix Σ_θ describes the relation between the defined latent dimensions.

The transformation between the two parametrizations can be done as follows. Let $\mathbf{A}^\circ$ and $\mathbf{A}$ be the matrices of discrimination parameters for the first and the second approaches, respectively. According to Béguin and Glas (2001), $\boldsymbol{\theta}_i$ can be transformed to $\boldsymbol{\theta}_i^\circ$ by $\boldsymbol{\theta}_i^\circ = \mathbf{L}^{-1}\boldsymbol{\theta}_i$, where $\mathbf{L}$ is the Cholesky decomposition of $\boldsymbol{\Sigma}_\theta$. Since $\mathbf{L}$ is the lower triangular and $\mathbf{A}\boldsymbol{\theta}_i = \mathbf{A}\mathbf{L}\boldsymbol{\theta}_i^\circ = \mathbf{A}^\circ\boldsymbol{\theta}_i^\circ$, the restrictions $\alpha_{jq} = 1$, if $j = q$ and $\alpha_{jq} = 0$, if $j \neq q$ for $j = 1, ..., Q$ and $q = 1, ..., Q$, are transformed into $\alpha_{jq}^\circ = 1$ for $j = 1, ..., Q - 1$ and $q = j + 1, ..., Q$. Let us define the lower triangular matrix $\mathbf{F}$ as the first Q rows of $\mathbf{A}^\circ$ and using $\boldsymbol{\theta}_i = \mathbf{F}\boldsymbol{\theta}_i^\circ$, we obtain $\boldsymbol{\Sigma}_\theta = \mathbf{F}\mathbf{F}^T$ and $\mathbf{A} = \mathbf{A}^\circ\mathbf{F}^{-1}$, which in turn produces restrictions $\alpha_{jq} = 1$, if $j = q$ and $\alpha_{jq} = 0$, if $j \neq q$ for $j = 1, ..., Q$ and $q = 1, ..., Q$.

Computation

For solving the estimation equations, the EM algorithm (Dempster, Laird, & Rubin, 1977) can be used. This general iterative algorithm for maximum likelihood (ML) estimation in incomplete data problems handles missing data by first replacing missing values by a distribution of missing values, second, estimating new parameters given this distribution, and, third, reestimating the distribution of the missing values assuming the new parameter estimates are correct. This process is iterated until convergence is achieved. The multiple integrals that appear above can be evaluated using Gauss-Hermite quadrature. A critical point related to using Gauss-Hermite quadrature is the dimensionality of the latent space, that is, the number of latent variables that can be analyzed simultaneously. Wood et al. (2002) indicates that the maximum number of factors is 10 with adaptive quadrature, 5 with nonadaptive quadrature, and 15 with Monte Carlo integration. In the present study, it is possible to use adaptive quadrature points; however, for more scales and time points, this procedure may become infeasible. In the discussion section of this paper, two alternative estimation procedures will be given.

Testing the Model

Preliminaries

The Lagrange Multiplier (LM) test by Aitchison and Silvey (1958) is grounded on the following rationale: Consider some general parameterized model and a special case of the general model, the so-called 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. The unrestricted elements of the vector of first-order derivatives are equal to zero because their values originate from solving the likelihood equations. The magnitudes of the elements of the vector of first-order partial derivatives corresponding to restricted parameters determine the value of the statistic: The closer they are to zero, the better the model fits.

More formally, let us consider a null hypothesis about a model with parameters $\boldsymbol{\eta}_0$. This model is derived from the general model with parameters $\boldsymbol{\eta}$ by fixing one or more parameters to known constants. We can make a partition of $\boldsymbol{\eta}_0$ as $\boldsymbol{\eta}_0 = (\boldsymbol{\eta}_{01}', \boldsymbol{\eta}_{02}')'$, and postulate constants described by vector $\boldsymbol{\eta}_{02}$ via $\mathbf{c} = \boldsymbol{\eta}_{02}$. The partial derivatives of the log-likelihood function of first and second order are $\mathbf{h}(\boldsymbol{\eta}) = \partial \log L(\boldsymbol{\eta})/\partial\boldsymbol{\eta}$ and $\mathbf{H}(\boldsymbol{\eta}, \boldsymbol{\eta}) = -\partial^2 \log L(\boldsymbol{\eta})/\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}'$ accordingly. Then, the LM statistic is given by

$$LM = \mathbf{h}(\boldsymbol{\eta}_0)'\mathbf{H}(\boldsymbol{\eta}_0, \boldsymbol{\eta}_0)^{-1}\mathbf{h}(\boldsymbol{\eta}_0). \tag{15}$$

For the case of a partitioned η, at the point of the LM estimates $\boldsymbol{\eta}_{01}$, the free parameters have partial derivatives equal to zero, $\mathbf{h}(\boldsymbol{\eta}_{01}) = 0$. The last equation can be computed through

$$LM(\mathbf{c}) = \mathbf{h}(\mathbf{c})'\,\mathbf{W}^{-1}\mathbf{h}(\mathbf{c}), \tag{16}$$

where

$$\mathbf{W} = \mathbf{H}_{22}(\mathbf{c}, \mathbf{c}) - \mathbf{H}_{21}(\mathbf{c}, \boldsymbol{\eta}_{01})\mathbf{H}_{11}(\boldsymbol{\eta}_{01}, \boldsymbol{\eta}_{01})^{-1}\mathbf{H}_{12}(\boldsymbol{\eta}_{01}\,\mathbf{c}), \tag{17}$$

and the partitioning of $\mathbf{W}(\eta_0, \eta_0)$ is according to the partition $\boldsymbol{\eta}_0 = (\boldsymbol{\eta}'_{01}, \mathbf{c}')'$. The LM statistic has an asymptotic χ^2 - distribution with degrees of freedom equal to the number of parameters in $\boldsymbol{\eta}_2$ (Rao, 1947; Aitchison & Silvey, 1958).

In the next section, we will introduce three LM statistics to test for differential item functioning, the shape of the item response curve, and local independence.

Differential Item Functioning

Differential item functioning (DIF) is a difference in item response behavior between equally proficient members of two or more groups. As an example, consider the difference in response behavior between boys and girls. It could be that performance of boys on science and mathematical items is better than performance for girls. On the other hand, the performance of girls on language items could be better than the performance of boys. By itself, however, this does not indicate DIF. DIF arises when, for a certain item, the level of performance of equally proficient boys and girls is different, probably because the item refers to irrelevant knowledge that is more ubiquitous in one population than in the other.

There are several techniques for detection of DIF and most of them are based on the evaluation of differences in response probabilities between groups conditional on a measure of proficiency.

Let $\boldsymbol{\eta}_{01}$ be a vector of parameters describing the explanatory parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and the covariance matrix $vec(\boldsymbol{\Sigma}_\theta)$ of the ability of students on the different subjects. Thus $\boldsymbol{\eta}_{01} = (\alpha, \beta, vec(\boldsymbol{\Sigma}_\theta))$ and we have to deal with null model. The alternative model is introduced by $\boldsymbol{\eta}_{02}$ which is $\boldsymbol{\eta}_{02} = (\delta_k)$.

The expectation of the item response τ_{nk} is a linear function of parameters as it was described by (2). The alternative model, then, can be written as

$$\tau_{nk} = \boldsymbol{\alpha}_k^T \boldsymbol{\theta}_n - \beta_k + \delta_k Y_n. \tag{18}$$

If $\delta = 0$, we arrive at τ_{nk} for the null model. This approach can be used to describe the populations of males and females; we then take Y_n to be

$$Y_n = \begin{cases} 1 & \text{if student } n \text{ is male} \\ 0 & \text{if student } n \text{ is female.} \end{cases} \tag{19}$$

It is easy to see that the difficulty parameters for males and females will be different. For boys we have $\beta_k + \delta_k$ (alternative model), whereas for girls we obtain β_k (null model).

Having an expression for τ_{nk}, we can estimate the first derivatives of the log-likelihood function $\mathbf{h}(\eta_{02}) = \partial \log L(\eta) / \partial \delta_k$ as

$$\mathbf{h}(\eta_{02}) = \sum_n x_{nh} Y_n - \sum_n Y_n E(\tau_{nk} \mid x). \tag{20}$$

Substitution of this expression into (17) provides the expression for the Lagrange multiplier to be evaluated:

$$LM = \frac{\left(\sum_n x_{nk} Y_n - \sum_n Y_n E(\tau_{nk} \mid x) \right)^2}{\mathbf{W}}, \tag{21}$$

where $\mathbf{W}$ now is a scalar. $\mathbf{W}$ can be interpreted as the variance of $\mathbf{h}(\eta_{02})$ given the parameter estimates.

Shape of the Item Response Function

We defined a number of boundaries for the score obtained on the other items. Let the item of interest be labeled k and the other items be labeled $j = 1, 2, ..., k-1, k+1, ..., K$. Let us also introduce the function $r(\mathbf{x}^{(k)})$, where $r(\mathbf{x}^{(k)})$ is the number-correct score on this partial response pattern, and $\mathbf{x}^{(k)}$ is the response pattern without item k, that is

$$r(\mathbf{x}^{(k)}) = \sum_{j \neq k} x_j. \tag{22}$$

$r(\mathbf{x}^{(k)})$ is often called a rest score. The range of possible scores $r(\mathbf{x}^{(k)})$ is partitioned into S_k intervals. Furthermore, we define

$$\mathbf{w}(s, \mathbf{x}^{(k)}) = \begin{cases} 1 & \text{if } r_{s-1} \le r\left(\mathbf{x}_n^{(k)}\right) < r_s, \\ 0 & \text{otherwise,} \end{cases} \tag{23}$$

for $s = 1, ..., S_k$ with $r_0 = -\infty$ and $r_{S_k} = \infty$. So, $w(s, \mathbf{x}^{(k)})$ is an indicator function assuming a value equal to 1 if the number correct score of response pattern $\mathbf{x}^{(k)}$ is in score range s. The expectation of the item response under the alternative model has the form

$$\tau_{nk} = \boldsymbol{\alpha}^T \boldsymbol{\theta} - \beta_k + \sum_{s=1}^{S-1} \mathbf{w}(s, \mathbf{x}^{(k)})\, \delta_s. \tag{24}$$

Note that $\mathbf{w}(s, \mathbf{x}^{(k)})$ is equal to 1 for only one of the S score segments, so the summation defined in (24) only selects one of the parameters δ_s. The parameter δ_s gauges the shift in item parameter β_k for score group s. Finally, note that there is no parameter δ_S, that is, the highest score level is used as a base line. If δ_S were also present, the model defined by (23) would no longer be identified. The expression for the first derivative with respect to δ_s is

$$\mathbf{h}(\eta_{02}) = \sum_n \mathbf{w}(s, \mathbf{x}^{(k)}) x_{nk} - \sum_n \mathbf{w}(s, \mathbf{x}^{(k)}) E(\tau_{nk} \mid x). \tag{25}$$

Note that the first-order derivative is the difference between the observed scores and expected scores of persons in subgroup s. The simplest form of the test emerges if only two score levels are considered, that is, if $S_k = 2$. In that case, one could set the cutoff score r_1 somewhere in the middle of the score range, say $r_1 = 0$, and test whether students with a high rest score $r(\mathbf{x}^{(k)})$ perform better or worse as expected on the target item k.

Local Independence

The assumption of local stochastic independence requires the association between the items to vanish given the parameters. If, for instance, we want to test whether an item response depends on the previous item, we define the indicator function

$$\mathbf{w}(s, x_{i(k-1)}) = \begin{cases} 1 & \text{if } r_{s-1} \le x_{i(k-1)} < r_s, \\ 0 & \text{otherwise,} \end{cases} \tag{26}$$

for $s = 1, ..., S_k$ with $r_0 = -\infty$ and $r_{S_k} = \infty$. As before, the simplest form of the test emerges if only two score levels are considered, and tests whether students with a high score on the previous item perform better or worse than expected on the target item.

The expression for expectation of the item response τ_{nk} has the form

$$\tau_{nk} = \boldsymbol{\alpha}^T \boldsymbol{\theta} - \beta_k + \sum_{s=1}^{S-1} \mathbf{w}(s, x_{i(k-1)}) \delta_s, \tag{27}$$

and the last contribution describes the effect of item $k - 1$ on item k. The parameter δ_k reflects the alternative model and we can get the expression for first derivative:

$$\mathbf{h}(\eta_{02}) = \sum_n x_{nk}\, \mathbf{w}(s, x_{i(k-1)}) - \sum_n \mathbf{w}(s, x_{i(k-1)})\, E(\tau_{nk} \mid x). \tag{28}$$

Note that, analogous to the test for the shape of the response functions, in this case, the first-order derivative is equal to the difference between the observed scores and expected scores of persons in subgroup s again. Also in this case, the simplest form of the test emerges if only two score levels are considered, that is, if $S_k = 2$. In that case, one could set the cutoff score r_1 somewhere in the middle of the score range of item $k - 1$ and test whether students with a high score on item $k - 1$ perform better or worse than expected on the target item k.

Simulation Study on the Type I Error Rate and Power of the Tests

The Type I error rate or significance level of a test is the probability of rejecting the null hypothesis of perfect model fit when the null model is true. In the present study, a significance level of 10% was used. On the other hand, power is the probability of rejecting the null hypothesis when a model violation occurs. One could call this the detection rate or hit rate. For all three tests described above, both the Type I error rate and the power were studied using simulation studies. In these studies, data were generated according to the model under the null hypothesis or the model under the alternative hypothesis, that is, under the null model with an added model violation. In all studies, the sample size was varied as 500, 1,000 and 4,000, and the number of items was varied as 10, 20, and 40. A unidimensional version of the model was used where the student parameters θ_n were drawn from a standard normal distribution. The item location parameters β were equally spaced between –1.0 and 1.0. Finally, the item discrimination parameters α were all equal to 0.5. In Appendix B, it is shown that in this way, the reliability of the scores, that is, the ratio of the within- and between-person variance, was equal to 0.60 for a test length of 10 items, 0.80 for a test length of 20 items, and 0.90 for a test length of 40 items.

Type I Error Rate

The study with respect to the Type I error rate was conducted using both the exact expressions for the second-order derivatives given in (7) and Appendix A, and the approximation given by (8). The number of replications in the simulation study was 100 for each combination of the sample size and test length. For the test on DIF, the numbers of simulees in each group were equal. For the tests for the item response function and local independence, two score groups were formed (so $S_k = 2$ for all k) and the cutoff score was always equal to zero. The Type I error rate was computed as the number of significance tests significant at the 10% level aggregated over all items. The results are presented in Table 1.

TABLE 1
Type I error rate of three test statistics computed using exact and approximated matrices of second-order derivatives

		DIF Test		IRF Test		LID Test	
N	*K*	Exact	Approx.	Exact	Approx.	Exact	Approx.
500	10	.10	.11	.08	.07	.06	.07
	20	.10	.08	.09	.08	.08	.13
	40	.10	.13	.10	.16	.06	.14
1,000	10	.10	.09	.14	.08	.07	.08
	20	.13	.10	.09	.09	.11	.11
	40	.10	.12	.11	.13	.10	.15
4,000	10	.11	.10	.11	.09	.08	.11
	20	.10	.10	.11	.11	.07	.09
	40	.10	.13	.11	.13	.10	.13

It can be seen that the control of Type I error rate was generally good. There were no main effects of sample size and test length. Further, there were no striking differences between the two versions of the statistic.

Differential Item Functioning

In the simulation study on the power of the tests to detect differential item functioning (DIF), three values were chosen for the effect size: $\delta = 0.1$, $\delta = 0.2$, and $\delta = 0.5$. Following the terminology of Cohen (1988), these effect sizes can be labeled as minimal, small, and large. Within every one of the 100 replications, the model violation was imposed on one randomly chosen item. The results are given in Table 2.

TABLE 2
Detection of differential item functioning

			DIF Test		IRF Test		LID Test	
N	K	δ	Hits	False Alarms	Hits	False Alarms	Hits	False Alarms
500	10	.1	.69	.12	.09	.07	.10	.07
		.2	1.00	.15	.09	.07	.07	.07
		.5	1.00	.29	.13	.07	.08	.09
	20	.1	.68	.12	.10	.10	.09	.11
		.2	1.00	.13	.10	.09	.09	.11
		.5	1.00	.17	.11	.10	.13	.10
	40	.1	.74	.14	.19	.15	.15	.13
		.2	1.00	.14	.14	.15	.13	.14
		.5	1.00	.15	.12	.15	.13	.14
1,000	10	.1	.90	.12	.09	.08	.09	.08
		.2	1.00	.19	.10	.08	.05	.07
		.5	1.00	.45	.24	.08	.09	.08
	20	.1	.94	.11	.11	.09	.13	.12
		.2	1.00	.13	.14	.08	.12	.12
		.5	1.00	.20	.14	.09	.10	.12
	40	.1	.96	.12	.10	.14	.11	.15
		.2	1.00	.13	.12	.13	.17	.14
		.5	1.00	.14	.14	.13	.11	.14
4,000	10	.1	1.00	.15	.11	.09	.12	.12
		.2	1.00	.45	.24	.10	.07	.12
		.5	1.00	.91	.62	.11	.07	.14
	20	.1	1.00	.11	.14	.11	.23	.21
		.2	1.00	.21	.10	.09	.22	.20
		.5	1.00	.45	.31	.10	.14	.23
	40	.1	1.00	.11	.14	.11	.24	.25
		.2	1.00	.12	.09	.11	.23	.25
		.5	1.00	.20	.13	.11	.15	.25

The columns labeled "Hits" give the proportion of replications for which the test on the differentially functioning item was significant at the 10% level. The columns labeled "False Alarms" give the proportion of significant results for the items conforming to the model, aggregated over replications and all model conforming items.

Note that the test on DIF displayed the largest proportion of hits; in most instances, this proportion was equal to 1.00. Note further that the proportion of hits for the test targeted to DIF has main effects of test length and sample size. Finally, the control of Type I error rate, that is, the proportion of false alarms, remained generally close to the nominal significance level. The main exceptions occurred for the large effect size in combination with a short test. The explanation is that in these cases the imposed model violation was such that every combination led to a global model violation affecting all items. The two other statistics had both the proportion of hits and false alarms at the nominal significance level. From a diagnostic perspective, it is desirable that tests have power against specific model violations, so this is a positive result.

Item Response Functions

The results of the simulation studies with respect to the power of the three tests to detect violation of the item response function (IRF) are shown in Table 3. The power is reported in the columns labeled "Hits." It can be seen that in the present case the test on DIF had no power. The test on the fit of the IRF had the highest power. But the test targeted at local independence (local item dependence (LID) test column) also had power to detect violation, although its power was of course less than the power of the specific test. In both cases, there were clear main effects of the effect size δ, sample size, and test length. Further, it can be seen that the Type I error rate was well under control.

TABLE 3
Detection of violation of the item response function

			DIF Test		IRF Test		LID Test	
N	*K*	δ	Hits	False Alarms	Hits	False Alarms	Hits	False Alarms
500	10	.1	.09	.11	.24	.06	.09	.07
		.2	.12	.12	.71	.07	.12	.08
		.5	.15	.11	1.00	.08	.23	.08
	20	.1	.09	.12	.27	.10	.14	.11
		.2	.10	.12	.86	.10	.18	.10
		.5	.09	.12	1.00	.11	.29	.10
	40	.1	.17	.14	.49	.15	.19	.14
		.2	.15	.13	.96	.14	.20	.14
		.5	.18	.14	1.00	.13	.29	.14
1,000	10	.1	.11	.19	.26	.07	.14	.06
		.2	.10	.12	.94	.07	.24	.07
		.5	.13	.10	1.00	.08	.42	.07
	20	.1	.07	.11	.37	.09	.20	.12
		.2	.09	.10	.97	.10	.23	.11
		.5	.09	.10	1.00	.09	.37	.11
	40	.1	.11	.12	.60	.13	.18	.14
		.2	.09	.12	1.00	.13	.29	.15
		.5	.13	.12	1.00	.11	.43	.13
4,000	10	.1	.10	.10	.69	.10	.21	.11
		.2	.09	.10	1.00	.09	.53	.09
		.5	.08	.10	1.00	.10	.90	.08
	20	.1	.14	.10	.91	.11	.34	.20
		.2	.15	.10	1.00	.11	.59	.19
		.5	.10	.10	1.00	.13	.88	.20
	40	.1	.13	.10	.97	.11	.44	.24
		.2	.06	.10	1.00	.10	.60	.24
		.5	.10	.10	1.00	.11	.86	.24

Local Independence

The results for the detection of violations of local independence are shown in Table 4. It can be seen that this test has now attained the highest power. The test on the shape of the IRFs had considerable power, but the power of the test on DIF hardly exceeded the nominal significance level. For all three tests, the Type I errors were virtually similar to their nominal levels.

TABLE 4
Detection of violation of local independence

			DIF Test		IRF Test		LID Test	
N	K	δ	Hits	False Alarms	Hits	False Alarms	Hits	False Alarms
500	10	.1	.11	.10	.09	.08	.11	.07
		.2	.14	.11	.13	.08	.41	.08
		.5	.09	.11	.23	.08	.95	.07
	20	.1	.14	.11	.11	.10	.17	.10
		.2	.10	.11	.12	.11	.40	.11
		.5	.09	.12	.14	.10	.93	.10
	40	.1	.14	.13	.14	.15	.17	.13
		.2	.12	.14	.17	.15	.38	.13
		.5	.17	.14	.18	.15	.90	.14
1,000	10	.1	.09	.10	.11	.08	.12	.06
		.2	.12	.10	.12	.07	.69	.08
		.5	.10	.10	.40	.08	1.00	.08
	20	.1	.10	.11	.14	.09	.13	.12
		.2	.11	.11	.12	.09	.64	.11
		.5	.10	.11	.26	.09	.98	.12
	40	.1	.09	.11	.10	.13	.11	.15
		.2	.13	.11	.12	.13	.60	.15
		.5	.12	.12	.14	.13	1.00	.14
4,000	10	.1	.10	.10	.19	.10	.38	.11
		.2	.09	.11	.49	.10	1.00	.11
		.5	.12	.10	.91	.13	1.00	.10
	20	.1	.09	.10	.12	.09	.18	.21
		.2	.09	.10	.29	.10	.99	.22
		.5	.13	.10	.57	.10	1.00	.21
	40	.1	.10	.10	.12	.11	.20	.25
		.2	.11	.10	.19	.11	.95	.25
		.5	.11	.10	.27	.11	1.00	.25

Conclusion

An MML framework for estimation and testing of an extension of a model for continuous responses proposed by Mellenbergh (1994) was presented, and simulation studies were conducted to test the Type I error rate and power. The simulation studies showed that these tests had excellent properties.

An advantage of MML estimation is that the item parameters and the covariance matrix can be estimated simultaneously. A disadvantage is the limited number of time points or the limited number of latent variables that can be analyzed. Earlier, it was mentioned that the maximum number of factors is 10 with adaptive quadrature, 5 with nonadaptive quadrature, and 15 with Monte Carlo integration. There are two alternatives that do not have these limitations. The first is a Bayesian procedure using a Markov Chain Monte Carlo (MCMC) algorithm (see, for instance, Gelman, Carlin, Stern, & Rubin, 1995) which was suggested by Béguin and Glas (2001). In this procedure, apart from the identification restrictions, the structure of the matrix factor loadings α_{ih} is entirely free. The second approach specifically applies to the case of a simple structure with unidimensional subscales loading on specific unidimensional latent variables used above. For that case, Rubin and Thomas (2001) discuss a two-stage procedure with a first stage consisting of calibrating the unidimensional subscales using a unidimensional IRT model, such as the Generalized Partial Credit Model (GPCM), and the second stage consisting of estimating the covariance matrix between the latent variables using a combination of parameter expansion and the EM algorithm.

A final remark concerns the relative merits of the likelihood-based and Bayesian methods. As mentioned, the main drawback of MML estimation is the possible limitation on the dimensionality of the latent space. Bayesian estimation methods based on the MCMC algorithm, usually combined with data augmentation methods, do not have these limitations. On the other hand, two potential problems with the Bayesian framework must also be considered. First, there are indications (Hendrawan, 2004; Dagohoy, 2005) that the MCMC estimation procedure is not robust to model violations. The reason may be that model violations cause disturbances in the data augmentation scheme that may lead to very slow convergence of the MCMC algorithm. Second, the procedures for testing model fit in a Bayesian framework are not yet satisfactorily developed. At this moment, two approaches to testing model fit based on a philosophy comparable to the one used above are being studied. The first approach is to use likelihood-based statistics as posterior predictive checks (Hoijtink, 2001; Glas & Meijer, 2003). As a general approach, this may have problems because, as was pointed out by Maris (2005), the power characteristics of posterior predictive checks are far from optimal. An alternative approach labeled Bayesian modification indices has been recently proposed by Fox and Glas (2005) but this approach has not yet been tested broadly for a general class of models. So for the time being, the proven robustness of MML estimation and testing methods still justifies their widespread use.

References

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

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*(4), 541–561.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Dagohoy, A. V. T. (2005). *Person fit for tests with polytomous responses*. Unpublished doctoral thesis, University of Twente, the Netherlands.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

Fox, J. P., & Glas, C. A. W. (2005). Bayesian modification indices for IRT models. *Statistical Neerlandica*, *59*, 95–106.

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

Glas, C. A. W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson (Ed.), *Objective measurement: Theory into practice (*Vol. 1, pp. 236–258). New Jersey: Ablex Publishing Co.

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica*, *8*(1), 647–667.

Glas, C. A. W. (1999). Modification indices for the 2–PL and the normal response model. *Psychometrika*, *64*(3), 273–294.

Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, *27*(3), 217–233.

Glas, C. A. W., & Suarez-Falcon, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*(2), 87–106.

Hendrawan, I. (2004). *Statistical tests for item response models: Power and robustness*. Unpublished doctoral thesis, University of Twente, the Netherlands.

Hoijtink, H. (2001). Conditional independence and differential item functioning in the two-parameter logistic model. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on Item Response Theory* (pp. 109–130). New York: Springer.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society*, *Series B*, *44*, 226–233.

Maris, G. (2005). Posterior predictive p-values for classical null hypotheses. *Statistica Neerlandica*, *59*(1), 70–81.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, *29*(3), 223–236.

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.

Rubin, D. B., & Thomas, N. (2001). Using parameter expansion to improve the performance of the EM algorithm for multidimensional IRT population-survey models. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on Item Response Theory* (pp. 193–204). New York: Springer.

Samejima, F. (1969) Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, *17*.

Wood, R., Wilson, D. T., Gibbons, R. D., Schilling, S. G., Muraki, E., & Bock, R. D. (2002). *TESTFACT: Test scoring, item statistics, and item factor analysis*. Chicago: Scientific Software International, Inc.

Appendix A

The information matrix is the sum over students n of terms

$$-E(B_n(\xi, \xi) \mid \mathbf{x}_n, \xi) - E(b_n(\xi)b_n(\xi)' \mid \mathbf{x}_n, \xi) + E(b_n(\xi) \mid \mathbf{x}_n, \xi)\, E(b_n(\xi) \mid \mathbf{x}_n, \xi)', \tag{29}$$

where

$$b_n(\xi) = \frac{\partial}{\partial \xi} \log Pr(\mathbf{x}_n, \boldsymbol{\theta}_n; \xi) \tag{30}$$

and

$$B_n(\xi, \xi) = \frac{\partial^2 \log Pr(\mathbf{x}_n, \boldsymbol{\theta}_n; \xi)}{\partial \xi \partial \xi'}. \tag{31}$$

The last term in (29) can be directly inferred from the estimation equations given by (14) and (13).

The kernel of the log-likelihood per student and item is given by

$$\log L_{nk} = -\frac{1}{2}(x_{nk} - \tau_{nk})^2, \text{ with } \tau_{nk} = \sum_h \alpha_{kh}\theta_{nh} - \beta_k.$$

For the items, the following derivatives are easily checked:

$$\frac{\partial \log L_{nk}}{\partial \alpha_{kh}} = \theta_{nh}(x_{nk} - \tau_{nk})$$

$$\frac{\partial \log L_{nk}}{\partial \beta_k} = (x_{nk} - \tau_{nk})$$

$$\frac{\partial^2 \log L_{nk}}{\partial \alpha_{kh}^2} = -\theta_{nh}^2$$

$$\frac{\partial^2 \log L_{nk}}{\partial \beta_k^2} = -1$$

$$\frac{\partial^2 \log L_{nk}}{\partial \alpha_{kh} \partial \alpha_{kp}} = -\theta_{nh}\theta_{np}$$

$$\frac{\partial^2 \log L_{nk}}{\partial \alpha_{kh} \partial \beta_k} = \theta_{nh}$$

Inserting these identities into (29) gives the information matrix for the items.

Appendix B

The reliability of a test is the ratio of systematic variance and total variance. It can be inferred from the variance decomposition

$$Var(\theta) = E(Var(\theta \mid x)) + Var(E(\theta \mid x)),$$

where θ is the ability parameter and x stands for a response pattern. Reliability is defined as

$$\rho = \frac{Var(E(\theta|x))}{Var(\theta)}$$

$$= \frac{Var(\theta) - E(Var(\theta|x))}{Var(\theta)}.$$

In the simulations reported above, $Var(\theta) = 1.0$. The term $Var(\theta \mid x)$ can be defined using the concept of Fisher information. Fisher information is the negative of the second-order derivative of the log-likelihood. Information is additive in the item responses for locally dependent items. So we have

$$\frac{\partial \log L_{x_{nk}}(\theta_n)}{\partial \theta_n} = \frac{\partial}{\partial \theta_n} \frac{-(x_{nk} - (\alpha_k \theta_n - \beta_k))^2}{2\sigma_k^2}$$

$$= \frac{\alpha_k (x_{nk} - (\alpha_k \theta_n - \beta_k))}{\sigma_k^2}$$

and

$$\frac{\partial^2 \log L_{x_{nk}}(\theta_n)}{\partial \theta_n^2} = \frac{\partial}{\partial \theta_n} \frac{\alpha_k (x_{nk} - (\alpha_k \theta_n - \beta_k))}{\sigma_k^2}$$

$$= -\frac{\alpha_k^2}{\sigma_k^2}.$$

Note that the second-order derivative does not depend on the response pattern. Therefore, we can drop the subscript n and write

$$\mathrm{E}(Var(\theta|x)) = \frac{1}{\sum_k \frac{\alpha_k^2}{\sigma_k^2}}.$$

Above, $\sigma_k^2 = 1.0$ and $\alpha_k = 0.5$ for all items, so the reliability was equal to 0.60 for a test length of 10 items, 0.80 for a test length of 20 items, and 0.90 for a test length of 40 items.