MULTIVARIATE BEHAVIORAL RESEARCH, *40*(1), 25–51

Testing Linear Models for Ability Parameters in Item Response Models

Cees A. W. Glas and Irene Hendrawan
Faculty of Behavioral Science
University of Twente

Methods for testing hypotheses concerning the regression parameters in linear models for the latent person parameters in item response models are presented. Three tests are outlined: A likelihood ratio test, a Lagrange multiplier test and a Wald test. The tests are derived in a marginal maximum likelihood framework. They are explicitly formulated for the 3-parameter logistic model, but it is shown that the approach applies to a broad class of item response models. Since the distributions of the test statistics are derived asymptotically, simulation studies were performed to assess the Type I error rates of the tests for small realistic sample sizes. Overall, the Type I error rates for the null hypothesis that a regression coefficient equals zero, were close to the nominal significance level. A number of power studies were conducted. It is argued that on theoretical grounds the power of the Lagrange multiplier test might be less than the power of the other two tests, but this expectation was not corroborated.

The robustness of the tests to violation of the item response model was investigated with simulation studies of the power and Type I error rate. The results showed that the performance of the tests was acceptable in the cases where local independence and the constancy of the discrimination parameters over treatment groups were violated to the same extent for all treatment groups. The simulation studies also showed that the tests were biased if local independence was violated for one of the treatment groups.

Item response theory (IRT) models provide a useful and theoretically well-founded framework for educational and psychological measurement. They support the construction of measurement instruments, linking and equating measurements, and the evaluation of test bias and differential item functioning. Fur-

Correspondence concerning this article should be addressed to Cees A.W. Glas, Department of Educational Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. E-mail: Glas@edte.utwente.nl ther, IRT provides the theoretical underpinning for item banking, optimal test construction and various flexible test administration designs, such as multiple matrix sampling, flexi-level testing and computerized adaptive testing (see van der Linden & Hambleton, 1997). A common element in these applications is that individual respondents or groups of respondents are administered partially different tests, the consequences of which motivate the application of IRT outlined below.

In IRT models, the influence of the items and persons on the observed scores are modeled by separate sets of parameters, say person ability parameters θ_n, $n = 1, \ldots, N$ and item parameters, for instance, item difficulty parameters b_i, $i = 1, \ldots, K$. One of the consequences of this separation of parameters is that if the item parameters are calibrated on a common scale, the person parameters θ_n can be estimated even though the persons are not administered the same subset in items. In this case, comparing raw sum score means makes no sense because the numbers of items administered and the characteristics of the items differ across groups. However, the measures θ_n are on a common scale and the averages of these measures can be compared meaningfully regardless of the items administered.

An analogous argument holds for regression and analysis of variance models for raw sum scores: If individual respondents or groups are administered different items, raw sum scores lose their meaning. Therefore, as an alternative, linear regression models for the parameter θ_n can be used to model the differences between groups, or, more generally, to model ability parameters using covariates. For example, suppose one is interested in the influence of background variables on intelligence. An intelligence test is administered in a two-factor design defined by a variable y_{n1} which is equal to one if person n is male and equal zero if person n is female, and a variable y_{n2} which is equal to one if person n lives in an urban area and equal zero if person n lives in a rural area. To evaluate the main effects of gender and place of residence, and a possible interaction effect, we consider the model

$$\theta_n = \delta_0 + y_{n1}\delta_1 + y_{n2}\delta_2 + y_{n1}y_{n2}\delta_{12} + \epsilon_n, \tag{1}$$

where ϵ_n has a normal distribution with a mean equal to zero and a variance σ^2. Note that δ_0 is an overall mean, δ_1 is the main effect of gender, δ_2 is the main effect of place, and δ_{12} is the interaction effect between gender and place of residence.

Linear models for the latent parameters θ were considered by Zwinderman (1991, 1997), and Fox and Glas (2001, 2002). The latter two authors present applications in the field of school effectiveness research. Other applications in educational research where differences between groups are relevant are international comparative educational surveys such as those of the IEA (TIMSS and PIRLS) and the OECD (IALS and PISA); for statistical issues regarding these surveys refer to Adams and Wu (2002). Embretson (1996) discussed applications in the framework of psychological research.

Apart from the advantage of the possibility of using incomplete designs, there are other arguments for favoring IRT-based methods over regression analysis and analysis of variance of classical number-correct scores. Firstly, if the two- or three-parameter IRT models hold, rather than the one-parameter logistic IRT model (all models defined below), then using number-correct scores leads to loss of precision (Fox & Glas, 2001, 2002). Secondly, and more importantly, Embretson (1996) showed that zero interaction effects in IRT models can be estimated as non-zero effects in number-correct score models, vice versa. The appropriateness of the test difficulty level for the ability distribution determined both the direction and magnitude of the bias of the estimates. The absolute value of the interaction contrast decreased as test difficulty differed from the optimal level, and the bias of the interaction contrast increased inward as the test difficulty level was increasingly inappropriate.

The objective of this article is to investigate means of testing hypotheses concerning the regression parameters in linear models for the latent parameters θ. Usually, the null-hypothesis entails that one or more regression parameters are equal to zero. Three tests will be compared: The likelihood ratio (LR) test, the Lagrange multiplier (LM) or efficient score test and the Wald test. The Type I error rate and the power of the tests will be compared using simulation studies. These simulation studies pertain to a situation where item parameters are estimated concurrently with the regression parameters. A second set of simulation studies pertains to the situation where the item parameters are considered known and only the regression parameters and the variance σ^2 are estimated. This situation occurs in situations were tests are assembled from an item bank, in which case the item parameters are estimated separately in a so-called calibration study. Finally, the robustness of the tests to violation of the IRT model will be investigated.

ITEM RESPONSE THEORY

In this article, the focus will be on dichotomously scored items. However, in the discussion section it will be shown that the theory presented also applies to polytomously scores items. Let a response of a person *n* to an item labeled *i* be coded by a stochastic variable X_{ni}, that assumes a value $x_{ni} = 1$if the response was correct and $x_{ni} = 0$ otherwise. In IRT, the probability of giving a correct response to an item is a function of the latent person parameter θ and a number of item parameters. The probability of a correct response given the item and person parameters, denoted by $P_i(\theta_n)$, is called the item response function (IRF). Much used models are the 1-, 2-, or 3-parameter logistic models (1PLM, 2PLM, and 3PLM). The 3PLM (Lord & Novick, 1968, Chaps. 17–20) is defined as

$$P_i(\theta_n) = c_i + (1 - c_i)\frac{\exp[a_i(\theta_n - b_i)]}{1 + \exp[a_i(\theta_n - b_i)]}. \quad (2)$$

This item response function is determined by an item discrimination parameter a_i, an item difficulty parameter b_i, and the guessing parameter c_i. The item difficulty parameter b_i is the point on the latent θ scale where the probability of a correct response is $c_i + (1 - c_i)/2$. If $c_i = 0$, that is, if there is no guessing, this chance is equal to .5. The greater the value of b_i, the greater the ability that is required to give a correct response to the item, that is, the item is more difficult. When the latent ability scale θ is identified in such a way that its mean is 0 and its standard deviation is 1, the values of b_i vary typically from about –2 (very easy) to +2 (very difficult). The a_i parameter is proportional to the slope of the IRF at the point b_i on the ability scale. In practice, a_i ranges from 0 (flat IRF) to 2 (very steep IRF). Items with steeper slopes are more useful for separating persons near an ability level $\theta = b_i$. The guessing parameter c_i (ranging from 0 to 1) is the probability of a correct score for low-ability persons (that is, $\theta \rightarrow -\infty$).

The 2PLM is obtained by setting $c_i = 0$ for all items; the 1PLM or Rasch (1960) model is obtained by additionally setting $a_i = 1$ for all items. In the 2- and 3PLM the IRFs may cross, whereas in the Rasch model the IRFs do not cross.

MML ESTIMATION

Marginal maximum likelihood (MML) estimation is probably the most used technique for item calibration. For the 1PLM, 2PLM and 3PLM, the theory was developed by such authors as Bock and Aitkin (1981), Thissen (1982), Rigdon and Tsutakawa (1983), and Mislevy (1984, 1985, 1986), and computations can be made using the software packages Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) and ConQuest (Wu, Adams, & Wilson, 1997). In this section, a general MML framework for estimation of linear models for the ability parameters will be outlined. In the next section, this framework will be used to define the three test statistics.

Since the most interesting applications will probably by in the framework of incomplete test administration designs, we introduce an item administration variable $\mathbf{d}_n$, which is a vector with entries d_{ni}, $i = 1, \ldots, K$, assuming a value one if item i was administered to person n, and zero otherwise. Further, let $\mathbf{x}_n$ be the response pattern of person n ($n = 1, \ldots, N$), that is, $\mathbf{x}_n$ has elements x_{ni}, for $i = 1, \ldots, K$. If $d_{ni} = 1$, the entries x_{ni} are as defined above, if $d_{ni} = 0$, x_{ni} is an arbitrary constant. Given ability parameter θ_n and the item parameters of the 3PLM a_i, b_i, and c_i, the probability of response pattern $\mathbf{x}_n$ is given by

$$p(\mathbf{x}_n \mid \mathbf{d}_n \theta_n) = \prod_i P_i(\theta_n)^{d_{ni} x_{ni}} \left[1 - P_i(\theta_n)\right]^{d_{ni}(1 - x_{ni})}. \quad (3)$$

In the MML approach, it is assumed that the ability parameters θ_n are independently drawn from one or more distributions. It will be assumed that the distributions of θ_n is normal with density $g(.; \mu, \sigma)$. However, the principle of defining a linear model for θ_n does not depend on the assumption of normality, and the tests proposed here could also be applied to alternative distributional assumptions. In this respect, the program Bilog-MG (Zimowski et al., 1996) has several options. It will be assumed that the expectation μ, depends on observed covariates $\mathbf{y}_n$ and regression coefficients $\boldsymbol{\delta}$. MML estimation derives its name from maximizing the log-likelihood that is marginalized with respect to θ, rather than maximizing the joint log-likelihood of all abilities parameters θ and item parameters. Let $\boldsymbol{\eta}$ be a vector of all item parameters a_i, b_i and c_i, regression coefficients $\boldsymbol{\delta}$ and σ. Then, the marginal log-likelihood function of $\boldsymbol{\eta}$ is given by

$$\log L(\boldsymbol{\eta}) = \sum_n \log\left\{\int p(\mathbf{x}_n \mid \mathbf{d}_n, \theta)\, g\left[\theta; \mu(\boldsymbol{\delta}, \mathbf{y}_n), \sigma\right] d\theta\right\}. \tag{4}$$

The reason for maximizing the marginal rather than the joint likelihood is that maximizing the latter does not lead to consistent estimates. This is related to the fact that the number of ability parameters grows proportional with the number of observations and in general this lead to inconsistency (Neyman & Scott, 1948). Simulation studies by Wright and Panschapakesan (1969) and Fischer and Scheiblechner (1970) show that these inconsistencies can indeed occur in IRT models. Kiefer and Wolfowitz (1956) have shown that MML estimates of an IRT model, are consistent under fairly reasonable regularity conditions, which motivates the general use of MML in IRT models.

Glas (1999) shows that the marginal likelihood equations in IRT models can be easily derived using Fisher's identity (see B. Efron, p. 29, in Dempster, Laird, & Rubin, 1977; Louis, 1982). Applying this to the present model, the first order derivatives of the log-likelihood function with respect $\boldsymbol{\eta}$ can be written as

$$\mathbf{h}(\boldsymbol{\eta}) = \frac{\partial \log L(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \sum_n E\left[\boldsymbol{\omega}_n(\boldsymbol{\eta}) \mid \mathbf{x}_n, \mathbf{y}_n, \boldsymbol{\eta}\right], \tag{5}$$

with

$$\boldsymbol{\omega}_n(\boldsymbol{\eta}) = \frac{\partial \log p(\mathbf{x}_n \mid \mathbf{d}_n, \theta)}{\partial \boldsymbol{\eta}} + \frac{\partial \log g\left[\theta; \mu(\boldsymbol{\delta}, \mathbf{y}_n), \sigma\right]}{\partial \boldsymbol{\eta}}, \tag{6}$$

where the expectation in Equation 5 is with respect to the posterior distribution $p(\theta|\mathbf{x}_\mathrm{n}, \mathbf{y}_\mathrm{n}, \boldsymbol{\eta})$.

Application of this framework to derive the likelihood equations for the item parameters of the 3PLM and the mean μ and standard deviation σ of the ability distribution $g(\theta;\ \mu,\ \sigma)$ can be found in Glas (1998, 2001). Here, only the derivation of the likelihood equations for the regression coefficients will be outlined. To apply Equations 5 and 6, first the first order derivatives of the log of the density of the ability parameters with respect to μ will be derived. In the present case, $g(\theta;\ \mu,\ \sigma)$ is the well-known expression for the normal distribution with mean μ and standard deviation σ, so it is easily verified that this derivative is given by

$$\frac{\partial \log g(\theta,\mu,\sigma)}{\partial \mu}=\frac{(\theta-\mu)}{\sigma^2}. \tag{7}$$

The regression parameters are modeled as linear functions of the group means, that is,

$$\boldsymbol{\mu} = \mathbf{Y}\boldsymbol{\delta},$$

where $\boldsymbol{\mu}$ is a vector of expected abilities, $\mathbf{Y}$ is some known matrix of observed covariates, and $\boldsymbol{\delta}$ is the vector of regression parameters. To obtain MML estimates of the regression parameters, we need to calculate the first order derivatives of the log-likelihood with respect to $\boldsymbol{\delta}$. This is done using

$$\frac{\partial \log L}{\partial \boldsymbol{\delta}}=\left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\delta}}\right]^t\left[\frac{\partial \log L}{\partial \boldsymbol{\mu}}\right]=\mathbf{Y}^t\left[\frac{\partial \log L}{\partial \boldsymbol{\mu}}\right],$$

where the vector $\partial \log L/\partial \boldsymbol{\mu}$ has entries

$$\frac{\partial \log L}{\partial \mu}=E\left[\frac{(\theta-\mu)}{\sigma^2}\,\middle|\,\mathbf{x}_n,\mathbf{y}_n,\boldsymbol{\eta}\right]. \tag{8}$$

Also the standard errors are easily derived in this framework: Mislevy (1986) points out that the information matrix can be approximated as

$$\mathbf{H}(\boldsymbol{\eta},\boldsymbol{\eta})\approx\sum_n E\left[\boldsymbol{\omega}_n(\boldsymbol{\eta})\mid\mathbf{x}_n,\mathbf{y}_n,\boldsymbol{\eta}\right]E\left[\boldsymbol{\omega}_n(\boldsymbol{\eta})\mid\mathbf{x}_n,\mathbf{y}_n,\boldsymbol{\eta}\right]^t, \tag{9}$$

and the standard errors are the diagonal elements of the inverse of this matrix.

THE WALD TEST

The Wald test is used for testing a special model against a more general alternative (Wald, 1943; also see Buse, 1982). The Wald test is evaluated using the maximum likelihood estimates of parameters of the general model. In the framework of the previous section, a hypothesis concerning some function of the item parameters, $\boldsymbol{\delta}$ and σ, say

$$\mathbf{f}(\boldsymbol{\eta}) = 0,$$

can be tested using the statistic

$$W_l = \mathbf{f}(\boldsymbol{\eta})^t \left\{ \left[\frac{\partial \mathbf{f}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \right]^t \mathbf{H}(\boldsymbol{\eta}, \boldsymbol{\eta})^{-1} \left[\frac{\partial \mathbf{f}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \right] \right\}^{-1} \mathbf{f}(\boldsymbol{\eta}).$$

Note that this is a quadratic form in $\mathbf{f}(\boldsymbol{\eta})$ where the matrix of weights is the inverse of the covariance matrix of $\mathbf{f}(\boldsymbol{\eta})$. The statistic has an asymptotic χ^2-distribution with degrees of freedom equal to the number of elements in $\mathbf{f}(\boldsymbol{\eta})$.

This definition of the Wald test also envelopes hypotheses about item parameters and σ, but below we are only interested in hypotheses about δ. For instance, the special case where the null hypothesis is $\delta_s = 0$ (where $1 \le s \le Q$) can be evaluated using the statistic

$$W_l = \left[\frac{\hat{\delta}_s}{\text{standard error of } \hat{\delta}_s} \right]^2, \tag{10}$$

Where $\hat{\delta}_s$ is the MML estimate of the parameter of interest. The test will be labeled W_l (the subscript l refers to the fact that the test focuses on means of latent variables). The standard error in the denominator is the diagonal of the inverted information matrix given by Equation 9.

LIKELIHOOD RATIO TEST

The likelihood function is a sufficient summary of sample information, and the likelihood principle entails that inferences are based on likelihood functions. When there are competing models, as possible explanations of a set of data, the likelihood approach is to compare their likelihoods. When H_0 and H_1 are a null and alternative hypothesis, we compare their likelihoods by means of the likelihood ratio statistic

$$LR = -2\log\frac{L(H_0)}{L(H_1)}, \tag{11}$$

where $L(H_0)$ and $L(H_1)$ are the likelihood under the null hypothesis and the alternative hypothesis, respectively (Lehmann, 1986). The LR statistic has an asymptotic χ^2-distribution with degrees of freedom equal to the difference in the number of parameters in H_1 and H_0. As intuition suggests, we prefer the explanation offered by H_1 when LR is small, that is, when H_1 is much more "likely" than H_0 (Lindgren, 1993).

LAGRANGE MULTIPLIER TEST

Applications of the LM test to the framework of IRT have been described by Glas (1998, 1999). The principle of the LM test (Aitchison & Silvey, 1958) and the equivalent efficient score test (Rao, 1947) can be summarized as follows. The arrangement of the LM test is the same as the arrangement of the likelihood-ratio test and the Wald test; all these three tests are used for testing a special model against a more general alternative (Buse, 1982). The LM test is based on evaluating a quadratic function of partial derivatives of the log-likelihood function of the general model evaluated at the maximum likelihood estimates of the special model. The LM test is evaluated using the maximum likelihood estimates of parameters of the special model. The vector of the first order derivatives of the special model are equal to zero because their value originate from solving the likelihood equations. The magnitude of the elements of the vector of the first order derivatives corresponding with special parameters determine the value of the statistic: the closer they are to zero, the better the model fits.

More formally, the principle can be described as follows. Consider a null hypothesis about a model with parameters ϕ_0. This model is special case of a general model with parameters ϕ. In this case, the special model is derived from the general model by setting one or more parameters to zero. Let the parameter vector ϕ_0 be partitioned as $\phi_0 = (\phi_{01}, \phi_{02})$, the null hypothesis is $\phi_{02} = 0$. Let $\mathbf{h}(\phi)$ be the partial derivatives of the log-likelihood of the general model, so $\mathbf{h}(\phi) = \partial \log L(\phi)/\partial\phi$. Let $\mathbf{H}(\phi, \phi)$ be defined as $-\partial^2 \log L(\phi)/\partial\phi\partial\phi^t$. [Note that $\mathbf{H}(.,.)$ is used as a generic symbol for a matrix of the opposite of second order derivatives of the log-likelihood function and the variables with respect to which derivatives are taken are the arguments of $\mathbf{H}(.,.)$. An analogous definition is used for $\mathbf{h}(.)$] The LM statistic is given

$$LM = \mathbf{h}(\phi_0)^t \mathbf{H}(\phi_0, \phi_0)^{-1} \mathbf{h}(\phi_0). \tag{12}$$

If Equation 12 is calculated using maximum likelihood estimate of ϕ_{01}, it has an asymptotic χ^2-distribution with degrees of freedom equal to the number of parameter in ϕ_{02} (Aitchison & Silvey, 1958).

Evaluated at the maximum likelihood estimate of ϕ_{01}, the partial derivatives $\mathbf{h}(\phi_{01})$ are equal to zero. Therefore, the Equation 12 can be computed as

$$LM(\phi_{02}) = \mathbf{h}(\phi_{02})^t \, \mathbf{W}^{-1}\mathbf{h}(\phi_{02}), \tag{13}$$

where

$$\mathbf{W} = \mathbf{H}(\phi_{02},\phi_{02}) - \mathbf{H}(\phi_{02},\phi_{01})\mathbf{H}(\phi_{01},\phi_{01})^{-1}\,\mathbf{H}(\phi_{01},\phi_{02}), \tag{14}$$

and where all expressions are evaluated using the maximum likelihood estimates of ϕ_{01} and the values of ϕ_{02} under the null hypothesis. The statistic has an asymptotic χ^2-distribution with degrees of freedom equal to the number of parameter in ϕ_{02}.

Note that $\mathbf{H}(\phi_{01}, \phi_{01})$ also plays a role in the Newton-Raphson procedure for solving the estimation equations and in computation of the observed information matrix or standard error. So its inverse will generally be available at the end of the estimation procedure. Further, if the validity of the model of the null-hypothesis is tested against various alternative models, the computational work is reduced because the inverse of $\mathbf{H}(\phi_{01}, \phi_{01})$ is already available and the order of $\mathbf{W}$ is equal to the number of parameters fixed.

The interpretation of the test is supported by observing that the value of Equation 13 depends on the magnitude of $\mathbf{h}(\phi_{02})$. If the absolute values of these derivatives are large, the fixed parameters are bound to change once they are set free. It means that the test is significant, that is, the special model is rejected. If the absolute values of these derivatives are small, the fixed parameters will probably show little change should they be set free. It means that the test is not significant, that is, the special model is adequate.

Besides a test of significance, this approach also provides information with respect to the direction in which the fixed parameters will change when set free. This is done by computing a new value of the fixed parameters, say ϕ^*_{02}, by performing one Newton-Raphson step, that is,

$$\phi^*_{02} = \mathbf{W}^{-1}\mathbf{h}(\phi_{02}). \tag{15}$$

Below, this new value ϕ^*_{02}, is often called a modification index. Testing whether ϕ^*_{02} significantly differs from zero can be done using Rao (1947) efficient score test. Rao shows that, assuming asymptotic normality of the estimates, ϕ^*_{02} has a multivariate normal distribution with mean zero and dispersion matrix **W**. Hence, $\phi^{*t}_{02}\mathbf{W}^{-1}\phi^*_{02}$ has asymptotic χ^2-distribution with degrees of freedom equal to the number of parameters fixed in the null-model (Glas, 1999). The test based on this statistic is asymptotically equivalent to the LM test (see, for instance, Buse, 1982).

Note that the score test is based on performing one Newton-Raphson step, that is, the test is based on an estimate that improves the likelihood, but does not completely maximize it under the alternative model. The LR and Wald tests, on the other hand, are both based on actual maximization of the likelihood under the alternative model. Therefore, it must be expected that these tests will have greater power than the LM test and score test. The reason for considering the LM test, where the LR and Wald tests are available, is that in complicated models with many parameters, various possibilities are open for improvement. Instead of estimating all these alternatives and performing LR and Wald tests to evaluate the improvement, one can perform a number of LM tests using one estimate under the null model only. The outcome of the tests are then used as a diagnostic tool to direct further analyses. So the LM test derives its significance from the fact that it serves another purpose than the LR and Wald tests.

To apply the principle of the LM test to the linear regression for θ, we partition the regression parameters $\boldsymbol{\delta}$ as $(\boldsymbol{\delta}_0, \boldsymbol{\delta}_1)$ and we test the null-hypothesis $\boldsymbol{\delta}_1 = 0$. Further the matrix **Y** is partitioned $[\mathbf{Y}_0, \mathbf{Y}_1]$ analogously to the partition of $\boldsymbol{\delta}$. So $\boldsymbol{\delta}_0$, with $\boldsymbol{\mu} = \mathbf{Y}_0\boldsymbol{\delta}_0$ are the regression coefficients in the model under the null-hypothesis and $\boldsymbol{\delta}_1$, with $\boldsymbol{\mu} = \mathbf{Y}_1\boldsymbol{\delta}_1$ are the regression coefficients under the alternative hypothesis, respectively. The LM test statistic is calculated under the null hypothesis $\boldsymbol{\delta}_1 = 0$ using the MML estimates of the item parameters, $\boldsymbol{\delta}_0$, and the variances of the ability distributions σ in that order. Further, $\hat{\boldsymbol{\lambda}}$ stands for the MML estimates of the item parameters, and $\boldsymbol{\mu}$ and σ, where $\hat{\boldsymbol{\mu}} = \mathbf{Y}_0\hat{\boldsymbol{\delta}}_0$ and $\hat{\boldsymbol{\delta}}_0$ is the MML estimate of the regression coefficients under the null-model. The first order derivatives are calculated as

$$\mathbf{h}(\boldsymbol{\delta}_1) = \frac{\partial \log L}{\partial \boldsymbol{\delta}} = \mathbf{Y}_1^t \left[\frac{\partial \log L}{\partial \boldsymbol{\mu}}\right],$$

where the vector $\partial\log L/\partial\boldsymbol{\mu}$ has entries defined by Equation 8. The matrix of weights is calculated as

$$\begin{aligned}\mathbf{W} &= \mathbf{H}(\boldsymbol{\delta}_1,\boldsymbol{\delta}_1) - \mathbf{H}(\boldsymbol{\delta}_1,\boldsymbol{\delta}_0)\mathbf{H}(\boldsymbol{\delta}_0,\boldsymbol{\delta}_0)^{-1}\mathbf{H}(\boldsymbol{\delta}_0,\boldsymbol{\delta}_1) \\ &= \mathbf{B}_1^t\mathbf{H}(\boldsymbol{\lambda},\boldsymbol{\lambda})\mathbf{B}_1 - \mathbf{B}_1^t\mathbf{H}(\boldsymbol{\lambda},\boldsymbol{\lambda})\mathbf{B}_0\left[\mathbf{B}_0^t\mathbf{H}(\boldsymbol{\lambda},\boldsymbol{\lambda})\mathbf{B}_0\right]^{-1}\mathbf{B}_0^t\mathbf{H}(\boldsymbol{\lambda},\boldsymbol{\lambda})\mathbf{B}_1, \qquad (16)\end{aligned}$$

where $\mathbf{H}(\boldsymbol{\lambda}, \boldsymbol{\lambda})$ is the matrix of second order derivatives with respect to a parameter vector $\boldsymbol{\lambda}$ containing the item parameters, the expectations $\boldsymbol{\mu}$, and σ, in that order, and

$$\mathbf{B}_0 = \begin{bmatrix} \mathbf{I}_P & 0 & 0 \\ 0 & \mathbf{Y}_0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{B}_1 = \begin{bmatrix} 0 \\ \mathbf{Y}_1 \\ 0 \end{bmatrix}.$$

The matrices $\mathbf{W}$ and $\mathbf{H}(\boldsymbol{\delta}_1, \boldsymbol{\delta}_1)$ in Equation 16 can be viewed as the asymptotic covariance matrices of $\mathbf{h}(\boldsymbol{\delta}_1)$ with $\boldsymbol{\lambda}$ estimated and known, respectively. Further, $\left[\mathbf{B}_0^t\mathbf{H}(\boldsymbol{\lambda},\boldsymbol{\lambda})\mathbf{B}_0\right]^{-1}$ is the asymptotic covariance matrix of the estimate of the parameters of the null-model, so the term $\mathbf{B}_1^t\mathbf{H}(\boldsymbol{\lambda},\boldsymbol{\lambda})\mathbf{B}_0\left[\mathbf{B}_0^t\mathbf{H}(\boldsymbol{\lambda},\boldsymbol{\lambda})\mathbf{B}_0\right]^{-1}\mathbf{B}_0^t\mathbf{H}(\boldsymbol{\lambda},\boldsymbol{\lambda})\mathbf{B}_1$ accounts for the influence of the estimation of the parameters on the covariance matrix of $\mathbf{h}(\boldsymbol{\delta}_1)$. So $\mathbf{W}$ is a conditional covariance matrix where the variance of the estimates of parameters under the null-model is explicitly taken into account. Below, the LM test will also be studied in a situation where the item parameters are considered known. In that case, these elements are removed from the matrix $\mathbf{H}(\boldsymbol{\lambda}, \boldsymbol{\lambda})$ and the first P rows are deleted from the matrices $\mathbf{B}_0$ and $\mathbf{B}_1$. The elements of $\mathbf{H}(\boldsymbol{\lambda}, \boldsymbol{\lambda})$ are of course still a function of the item parameters also, but $\mathbf{H}(\boldsymbol{\lambda}, \boldsymbol{\lambda})$ is then computed with the known values imputes as fixed constants.

A SIMULATED EXAMPLE

Before presenting the simulation studies pertaining to the Type I error rate and power of the tests, a simulated example will be presented to give the flavor of the method. We will use the model given by Equation 1, where δ_0 is an overall mean that is here assumed to equal to zero, δ_1 is the main effect of, say, gender, δ_2 is the main effect of, say, place of residence (urban or rural), and δ_{12} is the interaction effect between gender and place of residence. The main effects were chosen as $\delta_1 = 0.5$ and $\delta_2 = 1.0$, and there was no interaction effect, so $\delta_{12} = 0$. The variance of the ability distribution was chosen equal to one. The number of simulees for each of the four combinations of the main effects was 50.

To keep the example concise, the Rasch model was used to simulate the data. The parameters of 50 items were chosen equally spaced between –0.5 and 2.0, that is, $b_i = -0.5 + 2.5(i - 1)/(K - 1)$, for $i = 1, \ldots, K$, with $K = 50$.

Three test administration designs were considered. In the first design, labeled Design 1, all simulees responded to all items. The generated data in Design 1 are summarized in the third column of Table 1. The column labeled Treatment refers to the four combinations of the main effects. The rows labeled Items give the item numbers, the rows labeled Range give the range of the item parameters, the rows

TABLE 1
Information on Simulated Data Set

Treatment					
y_1	y_2		*Design 1*	*Design 2*	*Design 3*
0	0	Items	1 – 50	1 – 20	31 – 50
		Range	–0.5 – 2.0	–0.5 – 0.5	1.0 – 2.0
		Mean	18.0	10.1	4.3
		Sd	9.6	4.6	3.5
		Alpha	.90	.81	.76
1	0	Items	1 – 50	11 – 30	21 – 40
		Range	–0.5 – 2.0	0.0 – 1.0	0.5 – 1.5
		Mean	23.1	10.0	8.1
		Sd	10.1	4.6	4.5
		Alpha	.91	.81	.81
0	1	Items	1 – 50	21 – 40	11 – 30
		Range	–0.5 – 2.0	0.5 – 1.5	0.0 – 1.0
		Mean	26.6	9.4	11.3
		Sd	11.0	4.9	4.9
		Alpha	.92	.84	.84
1	1	Items	1 – 50	31 – 50	1 – 20
		Range	–0.5 – 2.0	1.0 – 2.0	–0.5 – 0.5
		Mean	30.8	9.6	15.0
		Sd	9.6	4.5	3.4
		Alpha	.90	.80	.72

Note. Sample size 50 simulees per test; test length Design 1 was 50 items; test length Design 2 and 3 was 20 items.

labeled Mean and Sd give the mean and the standard deviation of the distributions of the number-correct scores, and the rows labeled Alpha give Cronbach's Alpha indexing the reliability for the generated data. It should be noted that the item parameters were well matched to the ability distributions, so it was conjectured that the problems of bias in the estimates of interaction effects using number-correct scores reported by Embretson (1996) might not occur here. An analysis of variance supported this conjecture.

For the next two designs, subsamples of the original simulated data sets were used. The second design, labeled Design 2, was an optimal design given the chosen item parameters with 20 items per treatment. In the Rasch model, information is optimal if the item and ability parameters are as close as possible (see, for instance, Birnbaum, 1968). In the treatment $y_1 = 0$ and $y_2 = 0$, the expectation of θ is equal to zero, so this treatment is administered the 20 easiest items, which have item parameters ranging from –0.5 to 0.5, with a mean of 0.0. The items for the other three treatments were chosen in an analogous manner. The resulting tests are given in the rows labeled Items under the column labeled Design 2 in Table 1. Note that the test administration design is anchored by the common items of the tests. Since the means of the items administered to each treatment are now equal to the mean abilities, the mean probability of a correct response in each treatment is equal to 0.5, so the means observed scores of are all approximately equal to 10.0. This, of course, provides no further basis for evaluation of the mean and interaction effects using number-correct scores. Note that the reliability of the tests in the four treatments was smaller than the analogous reliability for Design 1. This is attributed to the decrease in test length. Still, in all four treatments the reliability is above the often-used benchmark of 0.80.

Design 3 was a mirror-image of Design 2, in the sense that the most proficient group of simulees were administered the easiest test and the least proficient group of simulees were administered the most difficult test. The choice of Design 2 and Design 3 is such that they have no data in common. Details are given in Table 1 under the heading Design 3. Note that since Design 3 provides far less information than the optimal Design 2, the reliability decreases further, with the reliability in two treatments falling below 0.80.

Next, concurrent MML estimates of the item and regression parameters and the standard deviation of the ability distribution were made. The latent scale was identified by setting the mean of the item parameters to zero. This restriction also held for the simulating values. For each of the three designs, two estimates were made, one with and one without an interaction term. Further, the three test statistics outlined above were computed. The results are given in Table 2. The true and estimated item parameter values are given only for six items, the results for the other items are analogous. Note that the standard error of the estimates increased from Design 1 to Design 2 (due to the decrease of the data set) and from Design 2 to Design 3 (due to the difference in optimality of the two designs). Still, in all three designs the estimates of the main effects differed significantly from zero, both in the model with and the model without an interaction effect. The outcomes of the three tests for the null hypothesis of zero interaction effect are shown in the last three rows of the table. All nine tests lead to the conclusion that the null hypothesis was not rejected.

SIMULATION STUDY WITH ALL PARAMETERS ESTIMATED

The first set of simulation studies was made to compare the Type I error rate and the power of the LR test, the LM or efficient score test and the Wald test in a situation where the item, regression and variance parameters were estimated concurrently. In all simulation studies, the linear model for θ given by Equation 1 was used. So the design consisted of four treatment groups, and the model for the

TABLE 2
Results of the IRT Analyses

		Design 1		*Design 2*		*Design 3*	
Parameter	*True Value*	*Estimate*	*(Se)*	*Estimate*	*(Se)*	*Estimate*	*(Se)*
b_1	–0.5	–0.49	(.16)	–0.68	(.26)	–0.73	(.37)
b_{10}	0.0	–0.07	(.15)	–0.09	(.28)	–0.11	(.36)
b_{20}	0.5	0.64	(.15)	0.54	(.20)	0.50	(.28)
b_{30}	1.0	1.18	(.15)	0.99	(.25)	1.11	(.24)
b_{40}	1.5	1.35	(.15)	1.12	(.25)	1.33	(.22)
b_{50}	2.0	1.96	(.16)	1.84	(.33)	2.06	(.36)
δ_1[a]	0.5	0.52	(.11)	0.59	(.16)	0.54	(.18)
δ_2	1.0	0.87	(.11)	0.95	(.23)	0.88	(.26)
δ_1[b]	0.5	0.63	(.15)	0.58	(.19)	0.63	(.21)
δ_2	1.0	0.97	(.17)	0.95	(.26)	0.98	(.28)
δ_{12}	0.0	–0.17	(.24)	0.03	(.31)	–0.27	(.34)
Fit Test		*Outcome*	*(Pr)*	*Outcome*	*(Pr)*	*Outcome*	*(Pr)*
LR		0.15	(.70)	0.03	(.85)	0.96	(.33)
Wald		0.52	(.47)	0.01	(.91)	0.83	(.36)
LM		0.90	(.34)	0.02	(.88)	1.04	(.31)

[a]Model with main effects only. [b]Model including interaction effect.

means of the four groups included two main effects, δ_1 and δ_2 and an interaction effect δ_{12}. In the simulations, the null hypothesis $\delta_{12} = 0$ was tested with a significance level of .05. The number of replications was equal to 100 for every condition in the simulation design. The numbers of simulees in the four treatment groups (N_g, $g = 1, 2, 3, 4$) were equal. In this study, the factor N_g had three levels: 50, 250, and 500 simulees, respectively. Test length was varied as $K = 5$, 15, and 30. The item responses were generated with the 2PLM. The item discrimination parameters were drawn from a log-normal distribution with a mean 0.0 and a standard deviation 0.25. The item difficulty parameters were drawn from a standard normal distribution. These item parameter distributions can be considered realistic, in fact, they are the standard item parameter priors in Bilog-MG (Zimowski et al., 1996).

Finally, all simulations were made using two setups for the regression models. In the first setup, in the null model, all regression coefficients were zero, that is $\delta = 0$, and in the alternative model, the interaction effect was varied as $\delta_{12} = .0$, .2, and .5. So in this setup, the null model was constant for all effect sizes of δ_{12}. In the second setup, all regression parameters varied along with the effect size, except δ_0, which was always equal to zero. So in this case, the model is $\delta_1 = \delta_2 = \delta_{12} = \delta_s$, for effect sizes $\delta_s = .0$, .2, and .5. The latter two are usually labeled as minimal and moderate (see Cohen, 1988). The second setup was added as a replication of the first setup with different parameter values and to make comparisons with the simulations reported in the next section.

Results

The empirical Type I error rate and the power of LR test, Wald test, and LM test for 2PLM are shown in Tables 3 and 4, for the first and second setup, respectively. The first, second and third column give the effect size, the test length and the sample size, respectively. The three last columns give the proportion of tests significant at the .05 level in 100 replications, for the LR test, Wald test, and LM test, respectively. Note that the bold numbers pertain to the power of the tests and other numbers pertain to

TABLE 3
Type I Error Rate and Power of the LR, Wald, and LM Test for the 2PLM, Setup 1, All Parameters Estimated (Power in Bold)

			Statistics		
Effect Size δ_s	*Test Length K*	*Group Size* N_g	*LR*	*Wald*	*LM*
.0	5	50	.05	.04	.06
		250	.06	.06	.06
		500	.03	.02	.03
	15	50	.08	.02	.10
		250	.04	.03	.05
		500	.06	.05	.06
	30	50	.08	.02	.15
		250	.04	.03	.06
		500	.05	.04	.06
.2	5	50	**.12**	**.09**	**.12**
		250	**.25**	**.25**	**.25**
		500	**.44**	**.44**	**.44**
	15	50	**.12**	**.04**	**.14**
		250	**.34**	**.33**	**.34**
		500	**.67**	**.66**	**.68**
	30	50	**.13**	**.04**	**.26**
		250	**.43**	**.38**	**.47**
		500	**.78**	**.76**	**.79**
.5	5	50	**.22**	**.13**	**.21**
		250	**.70**	**.69**	**.70**
		500	**.97**	**.97**	**.97**
	15	50	**.31**	**.17**	**.35**
		250	**.90**	**.90**	**.91**
		500	**1.00**	**1.00**	**1.00**
	30	50	**.39**	**.17**	**.54**
		250	**.96**	**.96**	**.96**
		500	**1.00**	**1.00**	**1.00**

TABLE 4
Type I Error Rate and Power of the LR, Wald, and LM Test for the 2PLM, Setup 2, All Parameters Estimated (Power in Bold)

				Statistics	
Effect Size δ_s	*Test Length K*	*Group Size* N_g	*LR*	*Wald*	*LM*
.0	5	50	.07	.04	.06
		250	.08	.06	.08
		500	.05	.05	.05
	15	50	.05	.02	.08
		250	.06	.06	.08
		500	.03	.03	.03
	30	50	.06	.01	.12
		250	.07	.05	.08
		500	.04	.04	.05
.2	5	50	**.10**	**.06**	**.11**
		250	**.30**	**.28**	**.30**
		500	**.47**	**.46**	**.47**
	15	50	**.11**	**.08**	**.14**
		250	**.40**	**.39**	**.41**
		500	**.70**	**.69**	**.71**
	30	50	**.12**	**.04**	**.20**
		250	**.46**	**.41**	**.49**
		500	**.88**	**.86**	**.92**
.5	5	50	**.22**	**.06**	**.28**
		250	**.64**	**.59**	**.64**
		500	**.96**	**.96**	**.96**
	15	50	**.35**	**.25**	**.42**
		250	**.91**	**.91**	**.92**
		500	**1.00**	**1.00**	**1.00**
	30	50	**.38**	**.15**	**.49**
		250	**.91**	**.88**	**.93**
		500	**1.00**	**1.00**	**1.00**

the Type I error rate. Overall, the results in Table 3 and Table 4 are comparable, so the difference in the choice of the parameters δ_1 and δ_2 had no remarkable effect. In the rows pertaining to an effect size $\delta_{12} = .0$, it can be seen that, the Type I error rate was close to its nominal value. For some cases (the LR and LM tests with test lengths 15 and 30), the Type I error rate was slightly inflated. This is because the derivations of distributions of the statistics are based on asymptotic arguments.

In the rows with an effect size $\delta_{12} > .0$, it can be seen that the power characteristics of the tests are as expected: There are main effects of sample size and test length. The explanation is that long tests contain more information on the ability parameters. The expectation that the power of the LM tests would be inferior to the power of the LR and Wald tests was not corroborated. In fact, in both setups, the power of the Wald test is inferior for a small group size $N_g = 50$, and the effect diminishes for $N_g = 250$, and it vanishes for $N_g = 500$. Unlike the LR statistic, the W_l statistic is based on an approximation of the matrix of second-order derivatives, but this also holds for the LM statistic, so there is no clear-cut explanation for this finding.

In general, it must be noted that in order to have acceptable power characteristics, there must at least be a moderate effect size (in the terminology by Cohen, 1988) and a sample size of 50 is clearly too small. The reason for the relatively less than favorable power characteristics is that the tests are model based and a large number of parameters needs to be estimated. This leads to inflation of uncertainty and standard errors, which, in turn, lowers the power of the tests.

SIMULATION STUDY WITH ITEM PARAMETERS FIXED

A second set of simulation studies pertains to the situation where the item parameters are considered known and only the regression parameters and the variances are estimated. Roughly speaking, the tests performed analogously in the previous study, so it was decided to focus on one test only, the Wald test. The model, the test lengths and sample sizes and the choice of item parameters were analogous to the previous study. However, in this case, a condition using the 3PLM was added; the guessing parameter was equal to .25. Further, to be able to increase the number of factors in the simulation design and to be able to increase the number of replications in each setting to 1,000, true values rather than estimates of the item parameters were used. So only the regression coefficients δ and the variances of the distributions of θ were estimated. In this setup, two factors were varied. The first was an effect size labeled δ_s, which has two levels: $\delta_s = .2$ and $.5$. The second factor consisted of four variations of the basic model shown by Equation 1:

Model 0: $\delta_0 = \delta_1 = \delta_2 = \delta_{12} = .0$;
Model 1: $\delta_0 = \delta_2 = \delta_{12} = .0$ and $\delta_1 = \delta_s$;
Model 2: $\delta_0 = \delta_{12} = .0$ and $\delta_1 = \delta_2 = \delta_s$;
Model 3: $\delta_0 = .0$ and $\delta_1 = \delta_2 = \delta_{12} = \delta_s$.

Note that Model 3 was equal to the model used in the second setup of the previous study.

Results

The empirical Type I error rate and the power of the W_l-test for 2PLM with effect size $\delta_s = .2$ are shown in Table 5. The first, second and third column of Table 5 are the test length, the sample size, and regression parameter tested, respectively. The four last columns in Table 5, pertain to the data generation models, for example,

TABLE 5
Type I Error Rate and Power of Wald Test for 2PLM With Effect Size δ_S = .2
(Power in Bold)

			Model			
Test Length K	*Sample Size N_g*	*Parameter*	*0*	*1*	*2*	*3*
5	50	δ_0	.07	.05	.06	.05
		δ_1	.05	**.15**	**.18**	**.15**
		δ_2	.03	.04	**.16**	**.14**
		δ_{12}	.04	.05	.04	**.09**
	250	δ_0	.06	.07	.06	.06
		δ_1	.06	**.51**	**.48**	**.51**
		δ_2	.04	.04	**.47**	**.51**
		δ_{12}	.06	.05	.04	**.22**
	500	δ_0	.06	.06	.05	.05
		δ_1	.06	**.74**	**.73**	**.76**
		δ_2	.06	.08	**.75**	**.74**
		δ_{12}	.07	.06	.07	**.42**
15	50	δ_0	.07	.08	.07	.06
		δ_1	.07	**.21**	**.21**	**.20**
		δ_2	.08	.09	**.21**	**.20**
		δ_{12}	.06	.07	.08	**.14**
	250	δ_0	.04	.06	.04	.04
		δ_1	.04	**.65**	**.63**	**.68**
		δ_2	.05	.05	**.66**	**.71**
		δ_{12}	.04	.04	.04	**.37**
	500	δ_0	.05	.05	.05	.06
		δ_1	.05	**.93**	**.93**	**.93**
		δ_2	.05	.05	**.92**	**.93**
		δ_{12}	.04	.05	.06	**.66**
30	50	δ_0	.07	.05	.06	.08
		δ_1	.06	**.20**	**.20**	**.20**
		δ_2	.09	.07	**.20**	**.24**
		δ_{12}	.08	.08	.08	**.13**
	250	δ_0	.04	.04	.05	.04
		δ_1	.06	**.70**	**.71**	**.72**
		δ_2	.04	.06	**.71**	**.71**
		δ_{12}	.04	.06	.06	**.46**
	500	δ_0	.06	.06	.05	.05
		δ_1	.06	**.95**	**.95**	**.95**
		δ_2	.05	.04	**.97**	**.97**
		δ_{12}	.04	.04	.03	**.74**

Model 0, 1, 2, and 3. For Model 0, all regression parameters were equal to .0, so this column shows the Type I error rate. For the other three models, the bold numbers pertain to non-zero-effects, and the other numbers to zero-effects. So the former are estimates of the power and the latter estimates of the Type I error rate. It can be seen that the presence of non-zero regression coefficients did not interfere

TABLE 6
Type I Error Rate and Power of Wald Test for 2PLM With Effect Size $\delta_s = .5$
(Power in Bold)

			Model			
Test Length K	*Sample Size N_g*	*Parameter*	*0*	*1*	*2*	*3*
5	50	δ_0	.04	.05	.05	.05
		δ_1	.04	**.39**	**.38**	**.37**
		δ_2	.04	.05	**.41**	**.40**
		δ_{12}	.04	.04	.06	**.18**
	250	δ_0	.07	.06	.06	.04
		δ_1	.06	**.96**	**.95**	**.96**
		δ_2	.04	.04	**.95**	**.96**
		δ_{12}	.05	.04	.05	**.69**
	500	δ_0	.05	.05	.06	.04
		δ_1	.06	**1.00**	**1.00**	**1.00**
		δ_2	.05	.05	**.99**	**1.00**
		δ_{12}	.06	.06	.06	**.93**
15	50	δ_0	.07	.06	.06	.06
		δ_1	.06	**.53**	**.54**	**.54**
		δ_2	.06	.05	**.56**	**.53**
		δ_{12}	.07	.08	.07	**.34**
	250	δ_0	.04	.04	.05	.06
		δ_1	.06	**1.00**	**1.00**	**1.00**
		δ_2	.05	.05	**1.00**	**1.00**
		δ_{12}	.06	.04	.05	**.92**
	500	δ_0	.05	.05	.05	.05
		δ_1	.04	**1.00**	**1.00**	**1.00**
		δ_2	.04	.04	**1.00**	**1.00**
		δ_{12}	.05	.05	.05	**1.00**
30	50	δ_0	.08	.07	.07	.07
		δ_1	.08	**.64**	**.60**	**.59**
		δ_2	.09	.08	**.62**	**.59**
		δ_{12}	.09	.09	.09	**.35**
	250	δ_0	.05	.05	.04	.04
		δ_1	.05	**1.00**	**1.00**	**1.00**
		δ_2	.05	.05	**1.00**	**1.00**
		δ_{12}	.06	.04	.05	**.94**
	500	δ_0	.05	.05	.04	.05
		δ_1	.05	**1.00**	**1.00**	**1.00**
		δ_2	.04	.04	**1.00**	**1.00**
		δ_{12}	.04	.06	.05	**1.00**

with the Type I error rates of the tests for the zero regression coefficients. In Table 6, the Type I error rate and the power across the different conditions is depicted for the effect size $\delta_s = .5$. The format of this table is the same as Table 5. As in the previous studies, the Type I error rate was approximately .05 for all conditions, and the power of the test increased with test length and group size. The power of the tests for the hypotheses $\delta_1 = 0$ and $\delta_2 = 0$ is larger than the power of the tests for the hypothesis $\delta_{12} = 0$.

The effect of fixing the item parameters at their true values can be assessed by comparing the power of the test for the hypothesis $\delta_{12} = 0$ in Table 4 with the power of the analogous test for the hypothesis $\delta_{12} = 0$ for Model 3 in the tables 5 and 6. It can be verified that the results are comparable, so there is no systematic effect here.

The simulation studies using the 3PLM are reported in the Table 7 and 8, for effect sizes .2 and .5, respectively. Comparing these results with the results in the previous tables, it can be seen that the power under the 2PLM is larger than under the 3PLM. This result is as expected because adding parameters while keeping the amount of data constant increases uncertainty, which in turn decreases the power. In the limiting case where the number of parameters is equal to the number of free observations, the power is zero.

SIMULATION STUDY OF ROBUSTNESS

Inferences made using IRT models are only valid if these models fit the data. However, it may be expected that some violations of the model detriment the inferences more than others. Consider the case of differential item functioning (DIF, for an overview refer to Holland & Wainer, 1993, and Camilli & Shepard, 1994). DIF is a difference in item response probabilities between equally proficient members of two or more groups. One might think of a test of foreign language comprehension, where girls are impeded by items referring to a football setting. The poor performance of the girls on the football-related items must not be attributed to their poor level of comprehension of the foreign language but to their lack of knowledge of football. If gender is entered as a main effect in a latent analysis of variance model as given in Formula 1, and the item difficulties b_i are systematically higher for one treatment group, this interaction will obviously bias the inferences made. On the other hand, other assumptions of the IRT models, such as constancy of discrimination parameters a_i over subgroups, or local stochastic independence may not systematically bias the inferences. The reason why varying item discrimination parameters might not affect the inferences is because they can be viewed as a noise factor. That is the item difficulties define the latent scale on which the abilities θ are positioned and compared. Discrimination parameters varying over groups affect the precision of the evaluation of θ but may have a relatively small effect on the positioning of θ on the latent scale. With respect to the violation of local independence the expectation is based on the fact that this violation can be imposed as a random shift of difficulty parameters that does not depend on the treatment group. Therefore, the effects of these two violations were the subject of a simulation study.

The simulation study had the same setup as the simulation studies reported in Tables 5 and 6, that is, the Type 1 error rate and power of the Wald test was investi-

TABLE 7
Type I Error Rate and Power of Wald Test for 3PLM With Effect Size $\delta_S = .2$
(Power in Bold)

			Model			
Test Length K	*Sample Size N_g*	*Parameter*	*0*	*1*	*2*	*3*
5	50	δ_0	.06	.06	.07	.07
		δ_1	.07	**.10**	**.11**	**.12**
		δ_2	.05	.04	**.11**	**.12**
		δ_{12}	.07	.04	.06	**.07**
	250	δ_0	.05	.05	.06	.05
		δ_1	.05	**.38**	**.39**	**.39**
		δ_2	.06	.04	**.38**	**.36**
		δ_{12}	.05	.05	.05	**.15**
	500	δ_0	.05	.05	.06	.05
		δ_1	.04	**.58**	**.58**	**.59**
		δ_2	.05	.04	**.57**	**.63**
		δ_{12}	.05	.05	.06	**.32**
15	50	δ_0	.06	.06	.05	.06
		δ_1	.08	**.18**	**.15**	**.13**
		δ_2	.07	.06	**.17**	**.15**
		δ_{12}	.07	.07	.06	**.16**
	250	δ_0	.04	.04	.04	.06
		δ_1	.04	**.60**	**.60**	**.60**
		δ_2	.06	.05	**.59**	**.59**
		δ_{12}	.05	.03	.04	**.33**
	500	δ_0	.04	.04	.04	.04
		δ_1	.06	**.84**	**.86**	**.87**
		δ_2	.04	.04	**.89**	**.87**
		δ_{12}	.06	.06	.06	**.58**
30	50	δ_0	.07	.09	.07	.07
		δ_1	.09	**.22**	**.19**	**.20**
		δ_2	.07	.07	**.19**	**.21**
		δ_{12}	.10	.09	.08	**.14**
	250	δ_0	.05	.07	.06	.04
		δ_1	.05	**.67**	**.68**	**.71**
		δ_2	.06	.06	**.66**	**.68**
		δ_{12}	.05	.05	.05	**.39**
	500	δ_0	.04	.04	.05	.05
		δ_1	.05	**.92**	**.94**	**.93**
		δ_2	.05	.04	**.95**	**.94**
		δ_{12}	.05	.04	.04	**.69**

gated for the 2PLM with fixed item parameters. Test lengths, sample sizes and effect sizes remained unchanged. However, only Model 3, the model with non-zero main effects, and the null-hypothesis $\delta_{12} = 0$ were studied.

The constancy of discrimination parameters a_i over subgroups was violated by drawing the item discrimination parameters from the log-normal distribution with

TABLE 8
Type I Error Rate and Power of Wald Test for 3PLM With Effect Size δ_S = .5
(Power in Bold)

			Model			
Test Length K	*Sample Size N_g*	*Parameter*	*0*	*1*	*2*	*3*
5	50	δ_0	.04	.06	.06	.05
		δ_1	.08	**.26**	**.24**	**.29**
		δ_2	.07	.03	**.27**	**.28**
		δ_{12}	.07	.06	.05	**.11**
	250	δ_0	.07	.06	.05	.03
		δ_1	.06	**.90**	**.88**	**.90**
		δ_2	.05	.05	**.86**	**.89**
		δ_{12}	.06	.05	.06	**.52**
	500	δ_0	.05	.05	.06	.05
		δ_1	.04	**.58**	**.58**	**.59**
		δ_2	.05	.04	**.57**	**.63**
		δ_{12}	.05	.05	.06	**.32**
15	50	δ_0	.05	.06	.06	.06
		δ_1	.05	**.99**	**.99**	**.99**
		δ_2	.04	.06	**.99**	**.99**
		δ_{12}	.05	.05	.05	**.84**
	250	δ_0	.05	.05	.04	.05
		δ_1	.05	**.99**	**.99**	**.99**
		δ_2	.05	.04	**.99**	**.99**
		δ_{12}	.05	.05	.05	**.84**
	500	δ_0	.06	.05	.05	.05
		δ_1	.06	**1.00**	**1.00**	**1.00**
		δ_2	.04	.04	**1.00**	**1.00**
		δ_{12}	.04	.05	.04	**.99**
30	50	δ_0	.09	.07	.09	.09
		δ_1	.09	**.54**	**.53**	**.57**
		δ_2	.07	.07	**.55**	**.57**
		δ_{12}	.09	.06	.07	**.35**
	250	δ_0	.05	.05	.06	.04
		δ_1	.05	**1.00**	**1.00**	**1.00**
		δ_2	.05	.06	**1.00**	**1.00**
		δ_{12}	.06	.07	.06	**.93**
	500	δ_0	.03	.04	.04	.04
		δ_1	.05	**1.00**	**1.00**	**1.00**
		δ_2	.03	.05	**1.00**	**1.00**
		δ_{12}	.05	.05	.05	**1.00**

a mean 0.0 and a standard deviation 0.25 for each of the four treatment groups separately. The item difficulty parameters, that were drawn from a standard normal distribution remained the same for each group.

Local stochastic independence was violated using a model proposed by Kelderman (1984) and Jannarone (1986) in the framework of Rasch model. In the application considered here, the dependence between the response on item *i* and the response on item *i* – 1 is modeled by the introduction of a parameter γ. The model is given by

$$P\left[X_{ni}=1 \mid X_{n(i-1)}=x_{n(i-1)}\right]=\frac{\exp\left[a_i\left(\theta_n-b_i\right)+x_{n(i-1)}\gamma\right]}{1+\exp\left[a_i\left(\theta_n-b_i\right)+x_{n(i-1)}\gamma\right]}.$$

Note that if $\gamma = 0$, there is no dependence, so then the 2PLM holds. Apart from $\gamma = 0$, the values $\gamma = .2$ and $\gamma = .5$ were used. The latter two conditions were crossed with a condition where the model was violated in all four treatment groups and a condition were the model was only violated in the fourth treatment group. The number of replications for every combination in the simulation design was 1,000.

Results

The results are shown in Table 9. The proportions in the last six columns of Table 9 are the proportions of significant W_l-tests in 1,000 replications. The column labeled No Violation pertains to the condition without a model violation, that is, the condition was $\gamma = 0$. The results in this are the same as the results in the last column of Table 5 and Table 6, in the rows pertaining to the test for the hypothesis $\delta_{12} = 0$. These results are included in Table 9 for reference purposes. Note that introduction of a violation of local independence only leads to an inflation of the Type I error rate and the power in the condition where the violation is only applied to treatment group 4. In the conditions where the violation is applied to all four treatment groups, the Type I error rate and the power are similar to the analogous values found in the simulations without a model violation. So the fact that the 2PLM did not hold here had no effect. From the comparison of the last column of Table 9 with the column labeled No Violation, it can be seen that the same held for the conditions where the constancy of the discrimination parameters a_i was violated. So the model violation introduced by redrawing a_i in every treatment group had no effect.

DISCUSSION

In this article, methods for testing hypotheses on regression parameters in linear models for θ in IRT models were presented. Three tests were outlined: A LR test, a LM test and a Wald test. Simulation studies were conducted to assess the Type I error rate and power of the tests. In the first set of simulation studies, all model parameters were estimated concurrently. The results showed that there were no marked systematic differences between the three tests. The expectation that the power of the LM test would be inferior to the power of the LR and Wald tests was

TABLE 9
Type I Error Rate and Power of Wald Test 2PLM
(Item Parameters Are Fixed; Power in Bold)

				Treatment 4		*All Treatments*		
Effect Size δ_s	*Test Length K*	*Group Size* N_g	*No Violation*	$\gamma = .2$	$\gamma = .5$	$\gamma = .2$	$\gamma = .5$	*DIF on* a_i
.0	5	50	.05	.07	.07	.05	.06	.07
		250	.04	.09	.22	.06	.05	.07
		500	.04	.16	.40	.06	.06	.07
	15	50	.07	.07	.14	.06	.06	.07
		250	.05	.12	.38	.03	.06	.05
		500	.06	.23	.65	.04	.05	.06
	30	50	.10	.10	.15	.08	.07	.08
		250	.05	.15	.46	.06	.05	.07
		500	.05	.23	.70	.03	.05	.07
.2	5	50	**.09**	**.15**	**.21**	**.07**	**.09**	**.11**
		250	**.22**	**.44**	**.68**	**.24**	**.25**	**.25**
		500	**.42**	**.79**	**.94**	**.49**	**.46**	**.46**
	15	50	**.14**	**.21**	**.33**	**.14**	**.13**	**.15**
		250	**.37**	**.70**	**.93**	**.41**	**.38**	**.42**
		500	**.66**	**.95**	**1.00**	**.69**	**.65**	**.64**
	30	50	**.13**	**.25**	**.38**	**.14**	**.13**	**.14**
		250	**.46**	**.78**	**.96**	**.44**	**.45**	**.43**
		500	**.74**	**.97**	**1.00**	**.73**	**.73**	**.68**
.5	5	50	**.18**	**.27**	**.39**	**.20**	**.18**	**.21**
		250	**.69**	**.90**	**.97**	**.72**	**.72**	**.72**
		500	**.93**	**1.00**	**1.00**	**.95**	**.94**	**.92**
	15	50	**.34**	**.45**	**.67**	**.36**	**.32**	**.31**
		250	**.92**	**.98**	**.99**	**.92**	**.92**	**.90**
		500	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**.99**
	30	50	**.35**	**.51**	**.69**	**.38**	**.35**	**.29**
		250	**.94**	**1.00**	**1.00**	**.95**	**.94**	**.95**
		500	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**.99**

not confirmed. For a small sample size, the power of the Wald test was inferior to that power of the other two tests. The power curve of all tests was as expected: Power increased with the effect size, test length and sample size. This was also found in the next set of simulation studies. These simulations were conducted to assess the effect of fixing the item parameters at their true values. The results showed no marked differences with the previous simulations.

Above, the testing procedure was outlined in detail and investigated in simulation studies for the 2PLM and 3PLM. However, the procedure can be easily generalized in several directions. First, incomplete designs, such as in multiple matrix sampling, flexilevel testing and computerized adaptive testing, can be accommodated by restricting the product over items in Equation 3 to the items that were actually responded to. The likelihood function shown by Equation 4 and all other derivations change analogously. Second, models for polytomous items (Bock, 1972; Masters, 1982; Muraki, 1992; Samejima, 1969, 1972, 1973; Tutz, 1990; Verhelst, Glas, & de Vries, 1997) are also easily accommodated. Consider polytomous items with m_i response categories, and define $x_{nih}(h = 0, \dots, m_i)$ as equal to one if the response was in category h, and zero otherwise. Then a response pattern can be coded as a vector $\mathbf{x}_n$ with entries $x_{nih}(h = 0, \dots, m_i)$, and the probability of a response pattern as given in Equation 3 can be redefined as

$$p(\mathbf{x}_n \mid \mathbf{d}_n, \theta_n) = \prod_i \prod_h P_{ih}(\theta_n)^{d_{ni} x_{nih}},$$

where $P_{ih}(\theta_n)$ is the probability of scoring in category h as defined by the IRT model considered, and all other derivations remain valid. Third, the tests are derived in an MML framework. Glas (1999, also see, Mislevy, 1984, 1985, 1986) shows that this framework is straightforwardly adapted to a Bayes modal framework, where prior distributions on item and population parameters are introduced.

A final remark concerns the robustness studies presented here. It was argued that the effects of model violations might be most serious if they pertained to the constancy of the location parameters b_i in combination with interaction with the covariates. The simulation studies showed that the tests were also biased if local independence was violated for one of the treatment groups. This bias did not occur in the case were local independence was uniformly violated and in the case where the constancy of the discrimination parameters a_i over treatment groups was violated. Still, in practical applications, it remains important to start an analysis by assessing item fit (for an overview, see Glas & Suarez Falcon, 2003) and person fit (for an overview, see Meijer & Sijtsma, 2001) in the separate treatment groups, and applying the techniques described above only to the subset of the data where the model fits, of course, as far as this does not substantially harm the validity of the research set up.

REFERENCES

Adams, R., & Wu, M. (2002). *Pisa 2000 technical report.* Paris: OECD.

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

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

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.

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *The American Statistician*, *36*, 153–157.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

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

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

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*, 201–212.

Fischer, G. H., & Scheiblechner, H. H. (1970). Algorithmen und programme für das probabilistische testmodell von Rasch. *Psychologische Beiträge*, *12*, 23–51.

Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs Sampling. *Psychometrika*, *66*, 271–288.

Fox, J. P., & Glas, C. A. W. (2002). Modelling measurement error in structural multilevel models. In G. A. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure models* (pp. 245–269). Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Glas, C. A. W. (2001). Differential item functioning depending on general covariates. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 131–148). New York: Springer.

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*, 87–106.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, *27*, 887–903.

Lehmann, E. L. (1986). *Testing statistical hypotheses* (*2nd ed.*). New York: Springer.

Lindgren, B. W. (1993). *Statistical theory* (*4th ed.*). London: Chapman & Hall.

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

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.

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

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

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359–381.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, *80*, 993–997.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Neyman, J., & Scott, E. L. (1948). Consistent estimates, based on partially consistent observations. *Econometrica*, *16,* 1–32.

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.

Rigdon S. E., & Tsutakawa, R. K. (1983). Parameter estimation in latent trait models. *Psychometrika*, *48*, 567–574.

Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. *Psychometrika*, *Monograph Supplement*, *No. 17.*

Samejima, F. (1972). A general model for free response data. *Psychometrika*, *Monograph Supplement*, *No. 18.*

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika*, *38*, 203–219.

Thissen D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, *47*, 175–186.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*, 39–55.

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

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.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*, 426–482.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, *29*, 23–48.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *ConQuest: Generalized item response modelling software.* Camberwell, Victoria: Australian Council for Educational Research.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *Bilog MG: Multiple-group IRT analysis and test maintenance for binary items.* Chicago: Scientific Software International.

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, *56*, 589–600.

Zwinderman, A. H. (1997). Response models with manifest predictors. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 245–256). New York: Springer.

Accepted June 2004