



## Marginal likelihood inference for a model for item responses and response times

Cees A. W. Glas<sup>1\*</sup> and Wim J. van der Linden<sup>2</sup>

<sup>1</sup>University of Twente, Enschede, The Netherlands

<sup>2</sup>CTB/McGraw-Hill, Monterey, California, USA

Marginal maximum-likelihood procedures for parameter estimation and testing the fit of a hierarchical model for speed and accuracy on test items are presented. The model is a composition of two first-level models for dichotomous responses and response times along with multivariate normal models for their item and person parameters. It is shown how the item parameters can easily be estimated using Fisher's identity. To test the fit of the model, Lagrange multiplier tests of the assumptions of subpopulation invariance of the item parameters (i.e., no differential item functioning), the shape of the response functions, and three different types of conditional independence were derived. Simulation studies were used to show the feasibility of the estimation and testing procedures and to estimate the power and Type I error rate of the latter. In addition, the procedures were applied to an empirical data set from a computerized adaptive test of language comprehension.

### I. Introduction

Concurrent modelling of responses and response times (RTs) has a long tradition in psychological and educational assessment (Luce, 1986; Roskam, 1987; van Breukelen, 2005; Verhelst, Verstralen, & Jansen, 1997). An important motive for this tradition was the wish to allow for the speed-accuracy trade-off, that is, the trade-off between working fast with low accuracy slow and slowly with high accuracy.

The joint model for responses and RTs on test items considered in this paper was presented in van der Linden (2006, 2007, 2008). The model differs from previously proposed models in its hierarchical or multi-level structure. Previously proposed models either integrate speed parameters or RTs into traditional single-level response models (Verhelst *et al.*, 1997) or, reversely, response parameters into RT models (Thissen, 1983). The present model consists of two item response theory (IRT) measurement models – one for the responses and another for the RTs. Both are nested under a

\* Correspondence should be addressed to Cees A. W. Glas, Department of Measurement and Data Analysis, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands (e-mail: c.a.w.glas@gw.utwente.nl).

second-level model which represents the speed-accuracy distribution in the population of respondents.

Allowing for the speed-accuracy distribution on a second level of modelling supports the investigation of a phenomenon that was an important motivation for the development of multi-level models in the first place: the ecological fallacy (see, for instance, Snijders & Bosker, 1999). This fallacy emerges because the association between two variables on a lower level (say, an individual level) and a higher level (say, a group or population level) may be quite different or even opposite. Van der Linden (2007) argues that this phenomenon is also relevant in modelling the observed relationships between responses and RTs, and that conclusions that ignore the multi-level structure may be quite misleading.

Van der Linden (2006, 2007) presents fully Bayesian methods for estimation and testing of the model. In the present research, we adopted a frequentist, likelihood-based alternative. The motivation is that we recognize that Bayesian and frequentist approaches both have their advantages and disadvantages. The advantage of a Bayesian approach, particularly when implemented through Markov chain Monte Carlo (MCMC) sampling from the posterior distribution, is the easy calculation of the posterior distribution of any function of the estimates. However, the likelihood approach has a long-standing, more rigorously developed tradition of statistical tests for model fit. In the likelihood approach, van der Linden and Glas (2010) present a number of person-fit tests which assume that the item parameters and the level 2 parameters are known. In the present paper, a marginal maximum-likelihood (MML, Bock & Aitkin, 1981) estimation procedure of these parameters is outlined. Further in the MML framework, we follow the tradition of Lagrange multiplier (LM) testing to develop specific tests of the assumptions of subpopulation invariance (i.e., absence of differential item functioning, DIF), the shape of the response functions, and three types of conditional independence. Below, the MML framework will be outlined first and then the tests will be presented. Also, results from simulation studies of the feasibility of the estimation and testing procedures as well as the power Type I error rates of the tests are presented. The paper concludes with an application of the estimation and testing procedures to a data set from a computerized adaptive test of language comprehension.

## 2. The model

The model consists of a hierarchical framework of four different component models on two different levels. The two first-level models are for the distributions of the responses and RTs for a fixed person on a fixed item. The two second-level models are for the distribution of the person parameters in the first-level models in the population of test takers and the distribution of the item parameters in the domain of items that is sampled.

On the first level, the probability of a correct response is given by the three-parameter logistic (3PL) model (Birnbaum, 1968; Lord, 1980); that is,

$$\begin{aligned} \Pr \{U_{ni} = 1; \theta_n, a_i, b_i, c_i\} &= c_i + (1 - c_i)\phi(\theta_n, a_i, b_i), \\ &= c_i + (1 - c_i)[1 + \exp(-a_i(\theta_n - b_i))]^{-1}, \end{aligned} \quad (1)$$

where  $\theta_n \in \mathbb{R}$  is the ability of test taker  $n$  and  $b_i \in \mathbb{R}$ ,  $a_i \in \mathbb{R}^+$ , and  $c_i \in [0, 1]$  are the difficulty, discrimination, and guessing parameters for item  $i$ , respectively. If the guessing parameter is set to zero, the model specializes to the two-parameter logistic (2PL) model.

For the distribution of RT  $T_{ni}$  of test taker  $n$  on item  $i$ , we use the log-normal model

$$f(t_{ni}; \tau_n, \alpha_i, \beta_i) = \frac{\alpha_i}{t_{ni}\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} [\alpha_i(\log t_{ni} - (\beta_i - \tau_n))]^2 \right\}, \quad (2)$$

where  $\tau_n \in \mathbb{R}$  is the speed at which test taker  $n$  operates on the test,  $\beta_i \in \mathbb{R}$  is the time intensity of item  $i$ , and  $\alpha_i \in \mathbb{R}^+$  is its discrimination parameter. The rest of this paper uses the fact that the model is equivalent to that of a normal distribution for the logarithm of the RT,  $\log T_{ni}$ .

On the second level, it is assumed that the first-level person parameters are an independent and identically distributed (i.i.d.) sample from a bivariate normal distribution; that is,

$$\boldsymbol{\lambda}_n = (\theta_n, \tau_n) \sim \text{MVN}(\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P). \quad (3)$$

Likewise, the first-level item parameters are assumed to be an i.i.d. sample from a multivariate normal distribution,

$$\boldsymbol{\xi}_i = (a_i, b_i, c_i, \alpha_i, \beta_i) \sim \text{MVN}(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I). \quad (4)$$

The model is identified if  $\boldsymbol{\mu}_P = \mathbf{0}$  and  $\text{diag}(\boldsymbol{\Sigma}_P) = \mathbf{1}$ , where  $\boldsymbol{\mu}_P$  and  $\boldsymbol{\Sigma}_P$  are the mean and covariance matrix as defined in (3). The restriction that the mean of the ability parameters  $\theta_n$  is equal to zero is analogous to the restriction that is usually imposed in MML estimation (Bock & Aitkin, 1981). Note that the difference  $\beta_i - \tau_n$  is the expected value of  $\log t_{ni}$  in a normal distribution and the restriction that the mean of the speed parameters  $\tau_n$  is equal to zero removes the trade-off between  $\beta_i - \tau_n$  from (2).

### 3. MML estimation of the model parameters

In MML estimation, a distinction is made between structural and incidental parameters. The structural parameters are estimated from a log-likelihood marginalized with respect to the incidental parameters. In the present case, our interest is in the estimation of the item parameters  $\boldsymbol{\xi}_i$ , the mean item parameters  $\boldsymbol{\mu}_I$ , the diagonal and lower-diagonal elements of  $\boldsymbol{\Sigma}_I$ , and the lower-diagonal elements of  $\boldsymbol{\Sigma}_P$ . Hence, the ability parameters  $\boldsymbol{\lambda}_n$  are to be treated as the incidental parameters. The asymptotic case we consider is thus for the sample size of test takers, denoted by  $N$ , going to infinity.

The structural parameters are collected in a vector  $\boldsymbol{\eta}$ , that is,  $\boldsymbol{\eta}$  contains the item parameters  $\boldsymbol{\xi}_i$  ( $i = 1, \dots, K$ , where  $K$  is the number of items in the test), the mean item parameters  $\boldsymbol{\mu}_I$ , the diagonal and lower-diagonal elements of  $\boldsymbol{\Sigma}_I$ , and the lower-diagonal elements of  $\boldsymbol{\Sigma}_P$ . The log-likelihood of the parameters  $\boldsymbol{\eta}$  is given by

$$\log L(\boldsymbol{\eta}) = \sum_{n=1}^N \log \int \int p(\mathbf{u}_n, \mathbf{t}_n | \boldsymbol{\lambda}_n, \boldsymbol{\xi}) g(\boldsymbol{\lambda}_n | \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P) d\boldsymbol{\lambda}_n + \sum_{i=1}^K \log h(\boldsymbol{\xi}_i | \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I), \quad (5)$$

where  $p(\mathbf{u}_n, \mathbf{t}_n | \boldsymbol{\lambda}_n, \boldsymbol{\xi})$  is the probability of the response pattern of person  $n$ , and  $g(\boldsymbol{\lambda}_n | \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$  and  $h(\boldsymbol{\xi}_i | \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$  are the normal densities of the person and item parameters, respectively, and  $\boldsymbol{\xi}$  is the vector of all item parameters. Note that (5) assumes independence of  $\boldsymbol{\lambda}_n$  and  $\boldsymbol{\xi}_i$ .

The marginal likelihood equations for  $\boldsymbol{\eta}$  can be easily derived using Fisher's identity (Efron, 1977; Louis, 1982). This identity equates the first-order derivatives of the marginal likelihood in (5) with respect to  $\boldsymbol{\eta}$  to the expected first-order derivatives of

a so-called ‘complete data’ log-likelihood. In the present case, the complete data are the response data  $\mathbf{u}_n$  and  $\mathbf{t}_n$  and the latent incidental parameters  $\lambda_n$ . So, we define

$$\omega_n(\boldsymbol{\eta}) = \frac{\partial}{\partial \boldsymbol{\eta}} \log p(\mathbf{u}_n, \mathbf{t}_n, \lambda_n, \boldsymbol{\eta}).$$

(We will use  $\omega_n(\cdot)$  as a generic operator for a first-order derivative of  $\log p(\mathbf{u}_n, \mathbf{t}_n, \lambda_n, \boldsymbol{\eta})$  with respect to any subset of the parameters in  $\boldsymbol{\eta}$ ; so the dimension of  $\omega_n(\cdot)$  is equal to its argument.)

From Fisher’s identity, it follows that the first-order derivatives are given by

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\eta}} \log L(\boldsymbol{\eta}) &= \sum_{n=1}^N E[\omega_n(\boldsymbol{\eta}) | \mathbf{u}_n, \mathbf{t}_n, \boldsymbol{\eta}] \\ &= \sum_{n=1}^N E \left[ \frac{\partial}{\partial \boldsymbol{\eta}} \log p(\mathbf{u}_n, \mathbf{t}_n | \lambda_n, \boldsymbol{\xi}) \middle| \mathbf{u}_n, \mathbf{t}_n, \boldsymbol{\eta} \right] \\ &\quad + \sum_{n=1}^N E \left[ \frac{\partial}{\partial \boldsymbol{\eta}} \log g(\lambda_n | \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P) \middle| \mathbf{u}_n, \mathbf{t}_n, \boldsymbol{\eta} \right] + \sum_{i=1}^K \frac{\partial}{\partial \boldsymbol{\eta}} \log b(\xi_i | \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I); \quad (6) \end{aligned}$$

that is, the first-order derivatives are equal to the posterior expected first-order derivatives  $\omega_n(\boldsymbol{\eta})$  of the complete data likelihood where the complete data comprise  $\mathbf{u}_n$ ,  $\mathbf{t}_n$ , and  $\lambda_n$ , for  $n = 1, \dots, N$ .

The power of Fisher’s identity is that the derivatives of the complete-data likelihood are generally easy to derive, while the derivation of the first-order derivatives of the observed data likelihood can be quite cumbersome. For instance, it is straightforward to obtain the maximum-likelihood (ML) estimate of the covariance matrix of the person parameters as

$$\boldsymbol{\Sigma}_P = \frac{1}{N} \sum_{n=1}^N \lambda_n \lambda_n^t.$$

Inserting this into (6) results in

$$\boldsymbol{\Sigma}_P = \frac{1}{N} \sum_{n=1}^N E(\lambda_n \lambda_n^t | \mathbf{u}_n, \mathbf{t}_n, \boldsymbol{\eta}),$$

where

$$E(\lambda_n \lambda_n^t | \mathbf{u}_n, \mathbf{t}_n, \boldsymbol{\eta}) = \int \cdots \int \lambda_n \lambda_n^t f[\lambda_n | \mathbf{u}_n, \mathbf{t}_n, \boldsymbol{\eta}] d\lambda_n,$$

and the posterior density has the form

$$f[\lambda_n | \mathbf{u}_n, \mathbf{t}_n, \boldsymbol{\eta}] = \frac{\prod_i p(\mathbf{u}_{ni}, \mathbf{t}_{ni} | \xi_i, \lambda_n) g(\lambda_n | \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)}{\int \cdots \int \prod_i p(\mathbf{u}_{ni}, \mathbf{t}_{ni} | \xi_i, \lambda_n) g(\lambda_n | \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P) d\lambda_n}.$$

Further, the likelihood equations for  $\boldsymbol{\mu}_I$  and  $\boldsymbol{\Sigma}_I$  are

$$\boldsymbol{\mu}_I = \frac{1}{K} \sum_{i=1}^K \xi_i$$

and

$$\Sigma_{\mathcal{I}} = \frac{1}{K} \sum_{i=1}^K \xi_i \xi_i^t.$$

The equations for the item parameters are derived analogously. Let  $P_{ni}$  and  $\phi_{ni}$  stand for  $\Pr\{U_{ni} = 1; \theta_n, a_i, b_i, c_i\}$  and  $\phi(\theta_n, a_i, b_i)$  as defined in (1). Then, the expressions for the first-order derivatives of the item parameters, say,  $\omega_n(a_i)$ ,  $\omega_n(b_i)$ ,  $\omega_n(c_i)$ ,  $\omega_n(\alpha_i)$ , and  $\omega_n(\beta_i)$ , are given by

$$\omega_n(a_i) = \frac{(u_{ni} - P_{ni})(1 - c_i)\phi_{ni}(1 - \phi_{ni})(\theta_n - b_i)}{P_{ni}(1 - P_{ni})}, \quad (7)$$

$$\omega_n(b_i) = \frac{(P_{ni} - u_{ni})(1 - c_i)\phi_{ni}(1 - \phi_{ni})}{P_{ni}(1 - P_{ni})}, \quad (8)$$

$$\omega_n(c_i) = \frac{(u_{ni} - P_{ni})(1 - \phi_{ni})}{P_{ni}(1 - P_{ni})}, \quad (9)$$

$$\omega_n(\alpha_i) = \alpha_i^{-1} - \alpha_i[\log t_{ni} - (\beta_i - \tau_n)]^2, \quad (10)$$

and

$$\omega_n(\beta_i) = \alpha_i^2[\log t_{ni} - (\beta_i - \tau_n)]. \quad (11)$$

Inserting these expressions as elements of the vector  $\omega_n(\xi_i)$  into

$$\sum_{n=1}^N E[\omega_n(\xi_i)|\mathbf{u}_n, \mathbf{t}_n, \boldsymbol{\eta}] + \Sigma_{\mathcal{I}}^{-1}(\xi_i - \boldsymbol{\mu}_{\mathcal{I}}), \quad (12)$$

and equating to zero gives the likelihood equations for the item parameters. Note that in Bayesian terminology, the structure of (12) leads to a so-called shrinkage estimator: the first term,  $\sum_{n=1}^N E[\omega_n(\xi_i)|\mathbf{u}_n, \mathbf{t}_n, \boldsymbol{\eta}]$ , depends on the data, while the second term,  $\Sigma_{\mathcal{I}}^{-1}(\xi_i - \boldsymbol{\mu}_{\mathcal{I}})$ , depends on the distribution of  $\xi_i$ . Therefore, the usual ML estimate of the parameters  $\xi_i$  is shrunken towards the common mean of the parameters  $\boldsymbol{\mu}_{\mathcal{I}}$ .

Computation of the standard errors of the parameter estimates is a straightforward generalization of the method for the 3PL model presented in Glas (1999). These estimates are found upon inverting the approximate information matrix

$$\sum_{n=1}^N E[\omega_n(\boldsymbol{\eta})|\mathbf{u}_n, \mathbf{t}_n, \boldsymbol{\eta}] E[\omega_n(\boldsymbol{\eta})|\mathbf{u}_n, \mathbf{t}_n, \boldsymbol{\eta}]^t. \quad (13)$$

Note that the information matrix is a sum over persons of outer products of a vector of first-order derivatives and its transpose.

### 3.1. Computation

The estimation equations are solved simultaneously. In practice this is done by Newton-Raphson, the expectation-maximization (EM) algorithm, or a combination of both (Bock & Aitkin, 1981). The EM algorithm (Dempster, Laird, & Rubin, 1977) is a general iterative algorithm for ML estimation in incomplete data problems. It handles missing data, first, by replacing missing values by their distribution; second, by estimating new parameters given this distribution; and third, by re-estimating the distribution of the

missing values assuming the new parameter estimates are correct. This process is iterated until convergence is achieved. The multiple integrals above can be evaluated using adaptive Gauss-Hermite quadrature (Schilling & Bock, 2005). 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 analysed simultaneously. Wood *et al.* (2002) indicate that the maximum number is 10 for adaptive quadrature, 5 for non-adaptive quadrature, and 15 for Monte Carlo integration.

## 4. Testing the model

### 4.1. Preliminaries

The LM test of Aitchison and Silvey (1958) has been used extensively to diagnose violations of various IRT models given the response data (Glas, 1999; Glas & Dagohoy, 2007; Glas & Suárez Falcón, 2003). For a general introduction to this type of test, which is also known as the Rao score test, see, for instance, Lehmann (1999, sect. 7.7) or Silvey (1975, sect. 7.4). The LM test is a locally most powerful test (see Cox & Hinkley, 1974). The arrangement of the LM test is the same as those 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. The tests are asymptotically equivalent. Further, using simulation studies, Glas and Hendrawan (2005) show that in an IRT framework they are also equivalent for small sample sizes. The reason for choosing the LM test rather than one of the two other tests is that the LM test only needs the estimates of the parameters of the IRT model, whereas the other two tests also need estimates of the parameters of alternative models associated with the model violations targeted. That is, the LM test supports the evaluation of several model violations for all individual items in one estimation run.

The test is defined as follows: consider some general parametric model as well as a special case of the model, the restricted model. This model is derived from the general model by imposing constraints on its parameter space. In many instances, this is accomplished by setting one or more of its parameters equal to a constant. The LM test is based on the evaluation of the first-order partial derivatives of the log-likelihood function of the general model at the ML estimates of the restricted model. The elements of the vector of first-order derivatives for the unrestricted parameters evaluated at the estimates are equal to zero because the estimates are solutions to the likelihood equations. Hence, the size of the elements of the vector of first-order partial derivatives for the restricted parameters determine the value of the LM 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 a general model with parameters  $\boldsymbol{\eta}$  by fixing one or more parameters to known constants. We partition  $\boldsymbol{\eta}_0$  as  $\boldsymbol{\eta}_0 = (\boldsymbol{\eta}_{01}^t, \boldsymbol{\eta}_{02}^t)^t$ , and postulate known constants  $\mathbf{c}$  for subvector  $\boldsymbol{\eta}_{02}$ . Thus, the null hypothesis is  $\boldsymbol{\eta}_{02} = \mathbf{c}$ . In the applications below, the restricted model is always of the type  $\mathbf{c} = \mathbf{0}$ . The first- and second-order partial derivatives of the log-likelihood function are  $\mathbf{h}(\boldsymbol{\eta}) = \partial \log L(\boldsymbol{\eta}) / \partial \boldsymbol{\eta}$  and  $\mathbf{H}(\boldsymbol{\eta}, \boldsymbol{\eta}) = -\partial^2 \log L(\boldsymbol{\eta}) / \partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^t$ , respectively.  $\mathbf{h}(\cdot)$  and  $\mathbf{H}(\cdot, \cdot)$  will be used as generic symbols for the vector of first-order derivatives and the matrix of the opposites of the second-order derivatives of the log-likelihood function. The variables with respect to which the derivatives are taken are the arguments of  $\mathbf{h}(\cdot)$  and  $\mathbf{H}(\cdot, \cdot)$  and can be any subset of  $\boldsymbol{\eta}$  or  $\boldsymbol{\eta}_0$ .

The LM statistic is given by

$$LM = \mathbf{h}(\boldsymbol{\eta}_0)^t \mathbf{H}(\boldsymbol{\eta}_0, \boldsymbol{\eta}_0)^{-1} \mathbf{h}(\boldsymbol{\eta}_0). \quad (14)$$

For the null hypothesis  $\boldsymbol{\eta}_0 = (\boldsymbol{\eta}_{01}^t, \mathbf{c}^t)^t$ , since the partial derivatives are evaluated at the ML estimates of the free parameters  $\boldsymbol{\eta}_{01}$ , we have that  $\mathbf{h}(\boldsymbol{\eta}_{01}) = 0$ . Hence, (14) simplifies to

$$LM(\mathbf{c}) = \mathbf{h}(\mathbf{c})^t \mathbf{W}^{-1} \mathbf{h}(\mathbf{c}), \quad (15)$$

where

$$\mathbf{W} = \mathbf{H}(\mathbf{c}, \mathbf{c}) - \mathbf{H}(\mathbf{c}, \boldsymbol{\eta}_{01}) \mathbf{H}(\boldsymbol{\eta}_{01}, \boldsymbol{\eta}_{01})^{-1} \mathbf{H}(\boldsymbol{\eta}_{01}, \mathbf{c}). \quad (16)$$

The LM statistic has an asymptotic chi-square distribution with degrees of freedom equal to the number of parameters in  $\boldsymbol{\eta}_2$  (Aitchison & Silvey, 1958). The LM test is equivalent to the efficient score test (Rao, 1947) as well as the modification index commonly used in structural equation modelling (Sörbom, 1989). From Sörbom (1989), it follows that the value of the LM statistic is proportional to the expected increase in the conditional likelihood should the additional parameters be estimated. In sections 4.2–4.11, we will introduce LM statistics targeted at the detection of DIF, violation of the postulated shape of the item response function or the RT distribution, and violations of conditional (or ‘local’) independence.

#### 4.2. Differential item functioning

DIF arises when equally proficient members of two or more groups show different response behaviour. As an example, assume that boys and girls show different behaviour, e.g., boys perform better on science and mathematical items but worse on language items. By itself, however, this finding need not indicate DIF. DIF arises when, for a given item, boys and girls equally proficient in science, mathematics, or language perform differently on the item, for instance, because the item refers to irrelevant knowledge ubiquitous among boys but less so among girls.

#### 4.3. DIF in item responses

There are several techniques for detecting DIF, and most of them are based on the evaluation of differences in the response probabilities between groups conditional on a measure of proficiency. In the framework of the LM test, the same idea can be implemented as follows.

First, we define an indicator variable to distinguish between the two groups, say, a reference and a focal group. (The generalization to more than two groups is straightforward.) Define  $y_n$  as

$$y_n = \begin{cases} 1 & \text{if } n \text{ belongs to the focal group,} \\ 0 & \text{if } n \text{ belongs to the reference group.} \end{cases} \quad (17)$$

Second, DIF is modelled to be a function of this variable. We use the approach in Glas (1998, 2001) by introducing the following more general alternative to the 3PL model:

$$\Pr \{U_{ni} = 1; \theta_n, \delta_i, Y_n\} = c_i + (1 - c_i)[1 + \exp(-a_i(\theta_n - b_i - y_n \delta_i))]^{-1}, \quad (18)$$

where  $\delta_i$  is a (positive or negative) shift in the difficulty of item  $i$  for the focal group. The regular 3PL model follows upon the restriction  $\delta_i = 0$ . So the LM statistic is used to test the hypothesis  $H_0 : \delta_i = 0$  against  $H_1 : \delta_i \neq 0$ .

Third, note that the alternative model is derived from the special model by adding only one parameter. Therefore, the statistic has an asymptotic chi-square distribution with one degree of freedom, and (15) specializes to

$$LM(\delta_i) = \frac{b(\delta_i)^2}{\mathbf{H}(\delta_i, \delta_i) - \mathbf{H}(\boldsymbol{\eta}_{01}, \delta_i) \mathbf{H}(\boldsymbol{\eta}_{01}, \boldsymbol{\eta}_{01})^{-1} \mathbf{H}(\boldsymbol{\eta}_{01}, \delta_i)} \Big|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}, \delta_i=0}, \tag{19}$$

where (using the definitions given above)  $b(\delta_i) = \partial \log L / \partial \delta_i$  is the first-order derivative of the log-likelihood of the model extended with respect to  $\delta_i$ ,  $\mathbf{H}(\boldsymbol{\eta}_{01}, \delta_i)$  is  $-\partial^2 \log L / \partial \boldsymbol{\eta}_{01} \partial \delta_i$ , and  $\mathbf{H}(\boldsymbol{\eta}_{01}, \boldsymbol{\eta}_{01})$  is defined as in (13). Note that  $b(\delta_i)$  and  $\mathbf{H}(\delta_i, \delta_i)$  are scalars, and  $\mathbf{H}(\boldsymbol{\eta}_{01}, \delta_i)$  is a vector because  $\delta_i$  is a scalar. Note further that  $\mathbf{H}(\boldsymbol{\eta}_{01}, \boldsymbol{\eta}_{01})$  is the observed information matrix, which is available when the MML estimation equations are solved.

The actual expression for  $b(\delta_i)$  can be derived easily as follows. Comparing the null and alternative models in (1) and (18), it can be seen that the roles of the parameters  $b_i$  and  $\delta_i$  are similar, except for the fact that the latter is multiplied by  $y_n$  while the former is not. Therefore,  $\omega_n(\delta_i) = y_n \omega_n(b_i)$ . Applying Fisher's identity, we obtain

$$b(\delta_i) = \sum_{n=1}^N y_n E[\omega_n(b_i) | \mathbf{u}_n, \mathbf{t}_n, \boldsymbol{\eta}], \tag{20}$$

that is, the first-order derivatives are similar to the first-order derivatives with respect to  $b_i$  (disregarding the prior distribution of  $b_i$ , of course), except that they are now summed only over the students in the focal group.

#### 4.4. DIF in RTs

The second test is for DIF with respect to the RTs of an item by a focal and a reference group, the idea being that an item becomes more suspicious if it has different time distributions for test takers from the two groups who work at the same speed on the test. The approach is analogous to that for different item functioning with respect to the responses: an additional parameter  $\delta_i$  is introduced that gauges the shift in the time intensity parameter,  $\beta_i$ , for the focal group. Therefore, the model under the alternative hypothesis becomes

$$f(t_{ni}; \tau_n, \alpha_i, \beta_i, y_n) = \frac{\alpha_i}{t_{ni} \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} [\alpha_i (\log t_{ni} - (\beta_i - \tau_n - y_n \delta_i))]^2 \right\}. \tag{21}$$

In this alternative model, the expected log-time,  $\beta_i - \tau_n - \delta_i$ , is a linear function of the item and person, just as in (2), plus the additional parameter  $\delta_i$ . If  $\delta_i = 0$ , the null model holds. In the alternative model,  $\delta_k$  is a free parameter. Thus, the hypotheses to be tested are  $H_0 : \delta_i = 0$  against  $H_1 : \delta_i \neq 0$ . For this test, the LM statistic defined by (14) and (15) can be used with  $\boldsymbol{\eta}_{02} = \delta_i = \mathbf{0}$ .

The expression for  $\mathbf{h}(\delta_i) = \partial \log L(\boldsymbol{\eta}) / \partial \delta_i$  can be found using Fisher's identity analogous to the earlier derivation of the MML equations. It immediately follows that the

first-order derivatives of the log-likelihood function,  $\mathbf{h}(\delta_i)$  are given by

$$\mathbf{h}(\delta_i) = \sum_{n=1}^N (\log t_{ni}) y_n - \sum_{n=1}^N y_n E((\beta_i - \tau_n) | \mathbf{u}_n, \mathbf{t}_n, \boldsymbol{\eta}). \quad (22)$$

Substitution of this expression into (16) yields

$$\text{LM} = \frac{(\sum_{n=1}^N (\log t_{ni}) y_n - \sum_{n=1}^N y_n E((\beta_i - \tau_n) | \mathbf{u}_n, \mathbf{t}_n, \boldsymbol{\eta}))^2}{\mathbf{W}}, \quad (23)$$

as the LM statistic, where  $\mathbf{W}$  is defined by (16). Note that  $\mathbf{W}$  is now a scalar, which can be interpreted as the variance of  $\mathbf{h}(\delta_i)$  given the parameter estimates. The statistic has an asymptotic chi-square distribution with one degree of freedom. It is also interesting to note that the numerator of (23) is based on the differences between the observed log-times of the test takers in the focal group and their posterior expected means; that is, on their posterior residuals.

#### 4.5. Shape of the item response function

Traditionally, the shape of the response functions in IRT has been evaluated by pooling the ability estimates into a number of adjacent intervals and comparing the average observed and expected responses for the respondents in the same interval. However, a requirement for a pooled statistic to lead to a Pearson type chi-square test is that all parameters are estimated from the same pooled data (e.g., Lindgren, 1968, sect. 9.1.1), which clearly is not the case in this application. Orlando and Thissen (2000) suggest an alternative in which the pooling of the test takers is based on observed number-correct scores rather than on model parameter estimates. An analogous approach will also be pursued here.

For a test targeted at item  $k$ , we partition the sample of respondents using a set of intervals for the total score on all other items. So, let the item of interest be labelled  $k$  and the other items  $i = 1, 2, \dots, k-1, k+1, \dots, I$ . Let  $\mathbf{u}_n^{(k)}$  be the response data without item  $k$ . Further, let  $r(\mathbf{u}_n^{(k)})$  be the total score on this partial response pattern,

$$r(\mathbf{u}_n^{(k)}) = \sum_{\substack{i=1 \\ i \neq k}}^K u_{ni}. \quad (24)$$

The range of  $r(\mathbf{u}_n^{(k)})$  is partitioned into  $S_k$  intervals. Define

$$\mathbf{w}(s, \mathbf{u}_n^{(k)}) = \begin{cases} 1 & \text{if } r_{s-1} \leq r(\mathbf{u}_n^{(k)}) < r_s, \\ 0 & \text{otherwise,} \end{cases} \quad (25)$$

for  $s = 1, \dots, S_k$  with  $r_0 = -\infty$  and  $r_{S_k} = \infty$ . So  $\mathbf{w}(s, \mathbf{u}_n^{(k)})$  is an indicator function assuming a value equal to 1 if the number-correct score for response pattern  $\mathbf{u}^{(k)}$  is in interval  $s$ .

A more general alternative to the 3PL model is

$$\Pr \{U_{ni} = 1; \theta_n, \delta_i\} = c_i + (1 - c_i) \left[ 1 + \exp \left( -a_i \left( \theta_n - b_i - \sum_{s=1}^{S-1} \mathbf{w}(s, \mathbf{u}^{(k)}) \delta_{is} \right) \right) \right]^{-1}, \quad (26)$$

where  $\delta_i = (\delta_{i1}, \dots, \delta_{iS-1})$ . Note that  $\mathbf{w}(s, \mathbf{u}^{(k)})$  is equal to 1 only for one of the  $S$  intervals, so the summation in (26) selects precisely one of the parameters from  $\delta_i$ . Parameter  $\delta_i$  gauges the shift in item parameter  $\beta_i$  for score group  $s$ . Finally, note that there is no parameter  $\delta_{iS}$ ; that is, the highest score level is used as a baseline. (If  $\delta_{iS}$  was also present, the model defined by (26) would be overidentified.)

The application of the LM statistic to test the model is analogous to the application to DIF. If  $\delta_i = \mathbf{0}$ , the null model holds. In the alternative model,  $\delta_i$  is a free parameter that can be interpreted as a shift in the item parameter  $\beta_i$ . In order to test whether the parameter significantly differs from zero, the LM statistic in (14) and (15) can be used with  $\boldsymbol{\eta}_{02} = \mathbf{0} = \delta$ . The statistic has an asymptotic chi-square distribution with  $S - 1$  degrees of freedom.

**4.6. Shape of the response function for the RTs**

Analogous to the test of the shape of the response function, the test based on the RTs is of the shape of their distribution. An alternative to the RT model is

$$f(t_{ni}; \tau_n, \alpha_i, \beta_i, \delta_i) = \frac{\alpha_i}{t_{ni} \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left[ \alpha_i (\log t_{ni} - (\beta_i - \tau_n - \sum_{s=1}^{S-1} \mathbf{w}(s, \mathbf{t}^{(k)}) \delta_{is})) \right]^2 \right\}, \quad (27)$$

where  $\mathbf{w}(s, \mathbf{t}^{(k)})$  is defined analogously to  $\mathbf{w}(s, \mathbf{u}^{(k)})$ . Observe that  $\delta_i$  introduces a shift of the density but that otherwise its shape remains the same. The appropriate test is thus of the hypothesis  $\delta_i = \mathbf{0}$ , which can be performed using an LM statistic with  $S - 1$  degrees of freedom.

The expression for  $\mathbf{h}(\delta)$  is found just as (11) was derived. The expression for the first-order derivative with respect to  $\delta_{si}$  is

$$\mathbf{h}(\delta_{si}) = \sum_{n=1}^N \mathbf{w}(s, \mathbf{t}^{(k)}) (\log t_{nk}) - \sum_{n=1}^N \mathbf{w}(s, \mathbf{t}^{(k)}) E(\beta_i - \tau_n | \mathbf{t}_n, \boldsymbol{\eta}). \quad (28)$$

Note that the first-order derivative is the difference between the observed and expected log-time for the persons in interval  $s$ .

**4.7. Simple case**

The simplest form of the two tests emerges if only two interval levels are considered; that is, for  $S = 2$ . For example, for the test based on the responses, one could set a cut-off score somewhere in the middle of the total score range and test whether students with a high rest-score  $r(\mathbf{u}^{(k)})$  perform better or worse than expected on the target item  $k$ . The null distribution for this version of the test has one degree of freedom.

#### 4.8. Local independence

As pointed out in van der Linden and Glas (2010), the hierarchical structure in (1)–(4) is based on three assumptions of conditional independence, namely between (i) responses on different items given  $\theta$ , (ii) RTs on different items given  $\tau$ , and (iii) responses and RTs on the same item given  $\theta$  and  $\tau$ . They propose three statistical tests of model fit based on the assumption that the item parameters are known. Here, we derive analogous tests under the assumption that all structural parameters are unknown but estimated using the MML method.

#### 4.9. Independence between responses

The alternative model is identical to that for a test of conditional independence for the 3PL model with MML estimation of the item parameters in Glas and Suárez Falcón (2003). Suppose the test is for the pair of items  $(i, k)$ . The alternative model,

$$\Pr\{U_{ni} = 1; \theta_n, \delta_{ik}, u_{nk}\} = c_i + (1 - c_i)[1 + \exp(-a_i(\theta_n - b_i - u_{nk}\delta_{ik}))]^{-1}, \quad (29)$$

allows for dependence between  $U_{ni}$  and  $U_{nk}$ ; it implies different distributions for  $U_{ni}$  given  $U_{nk} = 0$  and 1, where the size of the differences between the two distributions depends on the value of  $\delta_{ik}$ . Since the regular 3PL model follows for  $\delta_{ik} = 0$ , we test the hypothesis  $H_0 : \delta_{ik} = 0$  against  $H_1 : \delta_{ik} \neq 0$ . Using (15) and (16), we derive the test statistic

$$LM(\delta_{ik}) = \frac{b(\delta_{ik})^2}{\mathbf{H}(\delta_{ik}, \delta_{ik}) - \mathbf{H}(\boldsymbol{\eta}_{01}, \delta_{ik})^t \mathbf{H}(\boldsymbol{\eta}_{01}, \boldsymbol{\eta}_{01})^{-1} \mathbf{H}(\boldsymbol{\eta}_{01}, \delta_{ik})}, \quad (30)$$

where  $b(\delta_{ik}) = \partial \log L(\boldsymbol{\eta}) / \partial \delta_{ik}$  is the first-order derivative of the log-likelihood of the model extended with  $\delta_{ik}$ ,  $\mathbf{H}(\boldsymbol{\eta}_{01}, \delta_{ik})$  is a vector  $-\partial^2 \log L(\boldsymbol{\eta}) / \partial \boldsymbol{\eta}_{01} \partial \delta_{ik}$ , and  $\mathbf{H}(\boldsymbol{\eta}_{01}, \boldsymbol{\eta}_{01})$  is defined by (13). The statistic is evaluated using the MML estimates of the parameters in the null model. The test has an asymptotic chi-square distribution with one degree of freedom.

The actual expression for  $b(\delta_{ik})$  can be derived very easily. Comparing the null and alternative models in (1) and (29), it can be seen that the roles of the parameters  $b_i$  and  $\delta_{ik}$  are similar, except for the multiplication of the latter by  $u_{nk}$ . Therefore,  $\omega_n(\delta_{ik}) = u_{nk}\omega_n(b_i)$ . Applying Fisher's identity, we obtain

$$b(\delta_{ik}) = \sum_{n=1}^N u_{nk} E[\omega_n(b_i) | \mathbf{u}_n, \mathbf{t}_n, \boldsymbol{\eta}]. \quad (31)$$

Observe that the first-order derivatives are similar to those with respect to  $b_i$  (disregarding the prior distribution of  $b_i$ , of course) but that they are now summed only over the test takers who responded correctly to item  $k$ .

#### 4.10. Independence between RTs

As already noted, the RT model in (2) can be viewed as a normal density for  $\log T_{ni}$ . This fact suggests a bivariate normal distribution for the log-times on the pair of items  $(i, k)$  as

an alternative model:

$$f(\log t_{ni}, \log t_{nk} | \tau_n, \rho_{ik}) = \frac{\alpha_i \alpha_k}{2\pi \sqrt{1 - \rho_{ik}^2}} \exp \left\{ \frac{-1}{2(1 - \rho_{ik}^2)} (\psi_{ni}^2 - 2\rho_{nik} \psi_{ni} \psi_{nk} + \psi_{nk}^2) \right\}, \tag{32}$$

with additional parameters  $|\rho_{ik}| \leq 1$ , where

$$\psi_{ni} = \alpha_i (\log t_{ni} - (\beta_i - \tau_n)). \tag{33}$$

The hypotheses to be tested are  $H_0 : \rho_{ik} = 0$  against  $H_1 : \rho_{ik} \neq 0$ . Under the null hypothesis, (32) factorizes into the product of the two densities for the RTs on item  $i$  and item  $k$  in (2). For test taker  $n$ , the complete data log-likelihood of the log RTs on items  $i$  and  $k$  can be written as

$$\ell(\rho_{ik}) = \text{const.} - \frac{1}{2} \log(1 - \rho_{ik}^2) - \frac{1}{2(1 - \rho_{ik}^2)} (\psi_{ni}^2 - 2\rho_{ik} \psi_{ni} \psi_{nk} + \psi_{nk}^2). \tag{34}$$

The first-order derivatives with respect to  $\rho_{ik}$  are given by

$$\frac{\partial \ell(\rho_{ik})}{\partial \rho_{ik}} = \frac{\rho_{ik} + \psi_{ni} \psi_{nk}}{1 - \rho_{ik}^2} - \frac{\rho_{ik} (\psi_{ni}^2 - 2\rho_{ik} \psi_{ni} \psi_{nk} + \psi_{nk}^2)}{(1 - \rho_{ik}^2)^2}. \tag{35}$$

Setting  $\rho_{ik} = 0$  and applying Fisher's identity results in

$$b(\rho_{ik}) = \sum_{n=1}^N E[\psi_{ni} \psi_{nk} | \mathbf{u}_n, \mathbf{t}_n, \boldsymbol{\eta}]. \tag{36}$$

The test statistic is given by

$$LM(\rho_{ik}) = \frac{b(\rho_{ik})^2}{\mathbf{H}(\rho_{ik}, \rho_{ik}) - \mathbf{H}(\boldsymbol{\eta}_{01}, \rho_{ik})^t \mathbf{H}(\boldsymbol{\eta}_{01}, \boldsymbol{\eta}_{01})^{-1} \mathbf{H}(\boldsymbol{\eta}_{01}, \rho_{ik})}, \tag{37}$$

where the matrices  $\mathbf{H}(\rho_{ik}, \rho_{ik})$ ,  $\mathbf{H}(\boldsymbol{\eta}_{01}, \rho_{ik})$ , and  $\mathbf{H}(\boldsymbol{\eta}_{01}, \boldsymbol{\eta}_{01})$  follow from (13) as before. Again, the statistic is evaluated using the MML estimates of the parameters in the null model only. It has an asymptotic chi-square distribution with one degree of freedom.

**4.11. Independence between response and RT**

The alternative model is

$$f(t_{ni}; \tau_n, \delta_i) = \frac{\alpha_i}{t_{ni} \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} [\alpha_i (\log t_{ni} - (\beta_i - \tau_n - u_{ni} \delta_i))]^2 \right\}. \tag{38}$$

In this model, the parameter  $\delta_i$  introduces a shift in the distribution of  $T_{ni}$  given  $U_{ni} = 1$  for test taker  $n$  relative to that given  $U_{ni} = 0$ . The hypotheses to be tested are  $H_0 : \delta_i = 0$  against  $H_1 : \delta_i \neq 0$ .

The complete data log-likelihood can be written as

$$\ell(\delta_i) = \log p(\mathbf{u}, \mathbf{t}; \delta_i, \boldsymbol{\tau}) = \text{const.} - \frac{1}{2} \sum_{n=1}^N \xi_{ni}^2, \tag{39}$$

with

$$\xi_{ni} = \alpha_i(\log t_{ni} - (\beta_i - \tau_n - u_{ni}\delta_i)). \quad (40)$$

The first-order derivatives with respect to  $\delta_i$  are

$$\frac{\partial \ell(\delta_i)}{\partial \delta_i} = -\alpha_i u_{ni} \xi_{ni}. \quad (41)$$

Setting  $\delta_i = 0$  and applying Fisher's identity results in

$$b(\delta_i) = -\alpha_i^2 \sum_{n=1}^N u_{ni} \{\log t_{ni} - \beta_i + E[\tau_n | \mathbf{u}_n, \mathbf{t}_n, \boldsymbol{\eta}]\}. \quad (42)$$

The test statistic is given by

$$\text{LM}(\delta_i) = \frac{b(\delta_i)^2}{\mathbf{H}(\delta_i, \delta_i) - \mathbf{H}(\boldsymbol{\eta}_{01}, \delta_i)^t \mathbf{H}(\boldsymbol{\eta}_{01}, \boldsymbol{\eta}_{01})^{-1} \mathbf{H}(\boldsymbol{\eta}_{01}, \delta_i)^t} \quad (43)$$

where  $b(\delta_i) = \partial \log L(\boldsymbol{\eta}) / \partial \delta_i$  is the first-order derivative of the log-likelihood of the model extended with  $\delta_i$ ,  $\mathbf{H}(\boldsymbol{\eta}_{01}, \delta_i)$  is  $-\partial^2 \log L(\boldsymbol{\eta}) / \partial \boldsymbol{\eta}_{01} \partial \delta_i$ , and  $\mathbf{H}(\boldsymbol{\eta}_{01}, \boldsymbol{\eta}_{01})$  is defined by (13). The statistic is evaluated using the MML estimates of the parameters in the null model. The statistic has an asymptotic chi-square distribution with one degree of freedom.

## 5. An example with simulated data

First, a small example with simulated data will be presented to show the feasibility of the estimation and testing procedures. An important limitation of the procedures is that the model may contain a large number of parameters so that the information matrix in (13) quickly becomes large when the number of items increases. This happens because every item contributes five item parameters. For instance, for a test of 20 items, the number of item parameters is already 100. In addition, we have 5 free parameters in  $\boldsymbol{\mu}_{\mathcal{I}}$ , 15 in  $\boldsymbol{\Sigma}_{\mathcal{I}}$ , and 1 in  $\boldsymbol{\Sigma}_{\mathcal{P}}$ . Thus in all, we would have 121 parameters, and the inversion of the information matrix would approach the limit of what is possible with reasonable precision.

The estimation procedure itself could be based on the EM algorithm, and matrix inversion would then play no role. However, computation of the standard errors would still require inverting the information matrix. Further, as can be verified from (16), the inverse of the information matrix also plays a role in the LM statistic. In fact, the term  $\mathbf{H}(\boldsymbol{\eta}_{01}, \boldsymbol{\eta}_{02})^t \mathbf{H}(\boldsymbol{\eta}_{01}, \boldsymbol{\eta}_{01})^{-1} \mathbf{H}(\boldsymbol{\eta}_{01}, \boldsymbol{\eta}_{02})$  is the variance in  $\mathbf{h}(\boldsymbol{\eta}_{02})$  lost by the estimation of  $\boldsymbol{\eta}_{01}$  (Glas & Suárez Falcón, 2003). Therefore, one of the topics investigated in this simulation was to simplify the estimation of this variance component, which was achieved by ignoring the covariances between the items, assuming that the matrix  $\mathbf{H}(\boldsymbol{\eta}_{01}, \boldsymbol{\eta}_{01})$  had a block-diagonal form with  $5 \times 5$  submatrices for every item with each submatrix a covariance matrix for five item parameters.

Although data sets of many different sizes were generated, for lack of space we only report the results for an example of 10 items and 1,000 respondents. The covariance between the latent person parameters  $\theta_n$  and  $\tau_n$  was set equal to 0.40. In order to keep the tables small, we considered an example with the guessing parameter fixed at 0.20

**Table 1.** Parameter recovery: true values and MML parameter estimates

Item	True values				Estimated values and standard errors							
	<i>a</i>	<i>b</i>	$\alpha$	$\beta$	<i>a</i>	<i>SE(a)</i>	<i>b</i>	<i>SE(b)</i>	$\alpha$	<i>SE(α)</i>	$\beta$	<i>SE(β)</i>
1	1.90	1.92	0.89	0.68	1.75	0.39	1.83	0.31	0.88	0.04	0.66	0.04
2	1.06	0.34	1.15	0.76	1.11	0.19	0.44	0.11	1.08	0.04	0.71	0.05
3	1.19	1.81	0.66	1.82	1.21	0.30	2.14	0.30	0.64	0.04	1.73	0.04
4	0.95	1.95	0.89	-0.16	1.03	0.24	2.02	0.23	0.89	0.04	-0.17	0.04
5	0.87	1.08	0.50	-0.11	1.10	0.16	1.23	0.12	0.46	0.04	-0.14	0.04
6	0.95	0.85	0.69	1.75	0.83	0.16	0.74	0.11	0.73	0.04	1.72	0.04
7	0.70	0.08	0.99	1.67	0.80	0.14	0.04	0.09	1.00	0.04	1.66	0.05
8	1.12	0.98	0.92	-0.42	1.03	0.17	0.92	0.12	0.98	0.04	-0.45	0.04
9	1.04	-1.65	0.99	0.47	0.98	0.19	-1.74	0.14	0.98	0.04	0.37	0.04
10	1.06	-0.13	1.31	1.41	1.33	0.20	-0.19	0.10	1.30	0.05	1.34	0.05

for all items. This parameter was not estimated but imputed as a constant during the estimation. The  $\alpha$ , *a*,  $\beta$ , and *b* parameters of all items were drawn from a four-variate normal distribution. The means and standard deviations of the  $\alpha$  and *a* parameters were 1.00 and 0.30 whereas those of the  $\beta$  and *b* parameters were 0.00 and 1.00, respectively. The results for one randomly chosen replication are shown in Tables 1-3; the results for the other replications were entirely comparable.

The generating values of the item parameters are shown in the four columns under the label ‘True values’ in Table 1. The estimates and their standard errors are given in the remaining columns. Note that the estimates are quite close to their generating values, and that the generating values are well within their confidence bands. Table 2 gives the values of the LM statistics for the violations of local independence between the responses in (30) and the RTs in (37), respectively. Both statistics were calculated for all pairs of consecutive items. The columns labelled ‘LM’ give the exact values of the statistics, the columns labelled ‘LM\*’ their approximate values computed using a block-diagonal information matrix.

**Table 2.** LM tests for local independence

Item pair	Between item responses				Between response times			
	LM	<i>p</i>	LM*	<i>p</i>	LM	<i>p</i>	LM*	<i>p</i>
1, 2	0.89	.35	0.37	.55	1.17	.28	1.08	.30
2, 3	1.24	.26	0.25	.61	0.96	.33	0.91	.34
3, 4	0.86	.35	0.39	.53	0.22	.64	0.21	.65
4, 5	0.10	.76	0.05	.82	1.79	.18	1.69	.19
5, 6	0.18	.67	0.08	.78	0.07	.79	0.07	.79
6, 7	0.63	.43	0.29	.59	0.00	.95	0.00	.95
7, 8	0.83	.36	0.21	.65	0.10	.76	0.09	.76
8, 9	0.12	.73	0.06	.81	0.00	.96	0.00	.96
9, 10	0.70	.40	0.05	.82	7.79	.01	7.30	.01

*Note.* LM is the statistic for the whole information matrix, LM\* for the block-diagonal matrix. Both statistics have one degree of freedom.

**Table 3.** LM tests for local independence between RTs and item responses with violation of local independence introduced for item 10

Item	LM	$p$	Subgroup $U_i = 0$		Subgroup $U_i = 1$	
			Observed	Expected	Observed	Expected
1	0.59	.44	-0.11	-0.12	-0.30	-0.26
2	0.88	.35	-1.16	-1.14	-1.12	-1.15
3	0.39	.53	1.61	1.66	1.61	1.60
4	0.76	.38	-1.08	-1.11	-1.27	-1.25
5	0.58	.45	0.12	0.13	0.21	0.18
6	0.60	.44	-1.88	-1.90	-1.96	-1.93
7	1.11	.29	1.21	1.17	1.15	1.17
8	1.60	.21	-0.42	-0.38	-0.28	-0.31
9	0.06	.81	-0.10	-0.11	-0.11	-0.09
10	15.97	.00	-1.80	-2.02	-2.00	-1.95

*Note.* The LM statistics have one degree of freedom.

For the tests of local independence between the item responses, ignoring the covariances between the parameter estimates of different items resulted in inflated significance probabilities and, as a result, a loss of power. However, such effects were hardly noticeable for the tests of local independence between the RTs. More importantly, in all the examples of more than 20 items, for which no results are reported here, the differences between LM and LM\* were negligible for all test statistics introduced in this paper.

Table 3 gives an example of the results for the LM tests of local independence between responses and RTs in (43). The set-up of the simulation was similar to the two previous examples, except that a violation of local independence with an effect size of  $\delta = .50$  was created for item 10. The third column of Table 3 shows that the LM test for item 10 was highly significant, while the other nine tests were not significant at all. Further, columns 4-7 of the same table present information about the estimated size of the violations. Formula (42) shows that the first-order derivative can be interpreted as a difference between observed and expected log RTs. The differences for the part of the sample with an incorrect response are in columns 4-5, those for the correct response in columns 6-7. In order to evaluate these differences, they should be weighted by their standard errors. This is exactly what is done in the expression of the LM statistics.

## 6. A simulation study of Type I error rate and power

The Type I error rate of a test is the probability of rejecting the null hypothesis of model fit when the null model actually holds. In the present study, the error was controlled at the 5% significance level. On the other hand, the power of a test is the probability of rejecting the null hypothesis when the model is violated. One could call this power also a detection or hit rate. For all tests introduced above, both the actual Type I error rate and the power were studied using simulated data. Hence, for both error rates the data were generated according to the null model and the alternative model (i.e., the null model with the alternative parameter added). In all studies, the sample sizes were equal to 500, 1,000, or 4,000. The item and person parameters were drawn as

in section 5. The number of replications was 100 for each combination of sample size and test length.

In the simulations with the null model as the generating model, the Type I error rates were computed as the number of tests significant at the 5% level aggregated over all items. The results showed that the control of the 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 versions of the statistic based on the complete and block-diagonal information matrices. The results are not reported in detail because the simulations below also show excellent control of the Type I error.

In this section, only the results for the tests of DIF in RTs, the shape of the RT distributions, and local independence between RTs are reported because the results for the other tests were entirely comparable. For the tests of the shape of the RT distributions, two score groups were formed (i.e.,  $S_k = 2$  for all  $k$ ) and the cut-off score was always equal to zero. As a result, the sizes of the two groups were approximately equal. The simulations for the tests of DIF were based on equal group sizes, too.

### 6.1. Differential item functioning

Three different values were chosen for the effect size:  $\delta = .1, .2$ , and  $.5$ . Following the terminology of Cohen (1988), these effect sizes can be labelled as minimal, small, and large. The item and person parameters were the same as in the study of the Type I error rate. For each of the 100 replications, model violation was created for one randomly chosen item. The results are given in Table 4, where the labels 'DIF', 'Shape', and 'LID' refer to the tests of DIF, shape of the RT distributions, and local independence between the RTs, respectively.

The columns labelled 'Hits' contain the proportion of replications for which the tests of DIF were significant at the 5% level when there was actual DIF. So, these columns give an estimate of the power of the test. The columns labelled 'False alarms' contain the proportions of significant results for the items conforming to the null model aggregated over replications. These columns give estimates of the Type I error rate.

Note that the tests of 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 tests of DIF had main effects for the test length and sample size. Finally, the control of the Type I error rate for the other two tests (i.e., their proportions of false alarms) remained generally close to their nominal significance level. The main exceptions occurred for the large effect size in combination with a short test. Our explanation is that for such cases the imposed model violation was so strong that it led to a global violation affecting all items. The two other statistics had both the proportion of hits and false alarms at the nominal significance level. From a model-diagnostic perspective, it is desirable that the statistical tests have power against specific model violations, so this is a positive result.

### 6.2. 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 are shown in Table 5. The power is reported in the columns labelled 'Hits'. In the present case, the test of DIF had no power but the test of the fit of the items response function had the highest power. The test of local independence had also power to detect violation, although, of course, its power was less than that of the specific test. In both cases, there were clear main effects of the

**Table 4.** Detection of DIF

<i>N</i>	<i>K</i>	$\delta$	DIF test		IRF test		LID test	
			Hits	False alarms	Hits	False alarms	Hits	False alarms
500	10	.1	.69	.06	.05	.04	.05	.03
		.2	1.00	.07	.05	.04	.05	.03
		.5	1.00	.15	.07	.04	.05	.04
	20	.1	.68	.06	.05	.05	.04	.06
		.2	1.00	.07	.08	.05	.05	.06
		.5	1.00	.09	.05	.05	.07	.06
	40	.1	.74	.07	.10	.06	.08	.07
		.2	1.00	.07	.07	.07	.08	.07
		.5	1.00	.08	.06	.08	.09	.08
1,000	10	.1	.90	.06	.05	.04	.05	.04
		.2	1.00	.10	.05	.04	.03	.04
		.5	1.00	.22	.12	.04	.05	.04
	20	.1	.94	.06	.09	.04	.06	.06
		.2	1.00	.07	.07	.05	.07	.06
		.5	1.00	.10	.07	.05	.05	.07
	40	.1	.96	.06	.05	.07	.06	.07
		.2	1.00	.06	.06	.08	.09	.07
		.5	1.00	.07	.07	.08	.06	.08
4,000	10	.1	1.00	.08	.07	.05	.07	.06
		.2	1.00	.23	.12	.05	.05	.07
		.5	1.00	.45	.31	.05	.05	.08
	20	.1	1.00	.05	.07	.05	.12	.11
		.2	1.00	.10	.05	.05	.11	.11
		.5	1.00	.23	.16	.06	.08	.12
	40	.1	1.00	.06	.07	.06	.13	.14
		.2	1.00	.06	.06	.06	.13	.14
		.5	1.00	.10	.07	.07	.08	.13

effect size  $\delta$ , sample size, and test length. Further, it can be seen that the Type I error rate was well under control.

### 6.3. Local independence

The results for the detection of violations of local independence are shown in Table 6. The test of violation of local independence now attained the highest power. Again, there were clear main effects of the effect size  $\delta$ , the sample size, and the test length. The test of the item response functions also had considerable power, but that of the test of DIF hardly exceeded its nominal significance level. For all three tests, the Type I errors were virtually similar to their nominal levels.

## 7. An empirical example

An empirical example is given to show how the procedure works in practice. The data were a small part of a large data set collected in a pre-test of a computerized adaptive test of Dutch language comprehension for non-native speakers. The test had no time limit.

**Table 5.** Detection of violation of the item response function

<i>N</i>	<i>K</i>	$\delta$	DIF test		IRF test		LID test	
			Hits	False alarms	Hits	False alarms	Hits	False alarms
500	10	.1	.05	.06	.24	.06	.09	.03
		.2	.06	.06	.71	.07	.12	.04
		.5	.08	.06	1.00	.08	.23	.04
	20	.1	.05	.06	.27	.05	.14	.06
		.2	.06	.06	.86	.05	.18	.05
		.5	.05	.06	1.00	.06	.29	.05
	40	.1	.09	.07	.49	.08	.19	.08
		.2	.08	.07	.96	.07	.20	.08
		.5	.09	.07	1.00	.07	.29	.09
1,000	10	.1	.06	.10	.26	.05	.14	.03
		.2	.05	.07	.94	.05	.24	.04
		.5	.07	.05	1.00	.06	.42	.05
	20	.1	.05	.06	.37	.05	.20	.06
		.2	.05	.05	.97	.06	.23	.06
		.5	.05	.06	1.00	.04	.37	.06
	40	.1	.06	.06	.60	.07	.18	.07
		.2	.05	.06	1.00	.07	.29	.08
		.5	.07	.06	1.00	.06	.43	.07
4,000	10	.1	.05	.05	.69	.05	.21	.07
		.2	.05	.05	1.00	.09	.53	.04
		.5	.05	.05	1.00	.05	.90	.04
	20	.1	.07	.05	.91	.06	.34	.11
		.2	.08	.05	1.00	.06	.59	.21
		.5	.05	.05	1.00	.07	.88	.10
	40	.1	.07	.05	.97	.06	.44	.12
		.2	.03	.05	1.00	.05	.60	.12
		.5	.05	.05	1.00	.06	.86	.12

We present the results of 317 test takers on 12 items (the complete data set consisted of 2,000 test takers responding to more than 200 items). The item administration design consisted of 11 tests, with between 7 and 11 items. Each item was administered to approximately 200 test takers.

The analyses were performed both for the IRT model with (3PL) and without the guessing parameter (2PL). Because the results were quite close, we only present those for the analysis without the guessing parameter. The parameter estimates and the standard errors are given in Table 7. The second panel of the table gives the estimate of the correlations between the item parameters. Note that the correlations between the item parameters in the response and RT models were small. The last panel gives the correlation between the two latent dimensions. The correlation was negative, so test takers of higher ability tended to have shorter RTs. The correlation was low ( $-.15$ ) but the absolute magnitude was substantially higher than the point biserial correlation between item responses and RTs, which was  $-.05$ . This result is in line with the often found phenomenon that latent correlations are higher than manifest correlations because they suffer less from the attenuation effect caused by the unreliability in the variables.

**Table 6.** Detection of violation of local independence

<i>N</i>	<i>K</i>	$\delta$	DIF test		IRF test		IID test	
			Hits	False alarms	Hits	False alarms	Hits	False alarms
500	10	.1	.06	.05	.09	.04	.11	.04
		.2	.07	.06	.13	.05	.41	.04
		.5	.05	.06	.23	.05	.95	.04
	20	.1	.07	.06	.11	.05	.17	.05
		.2	.05	.06	.12	.06	.40	.06
		.5	.05	.06	.14	.05	.93	.06
	40	.1	.07	.07	.14	.08	.17	.07
		.2	.06	.07	.17	.08	.38	.07
		.5	.09	.07	.18	.08	.90	.07
1,000	10	.1	.05	.05	.11	.05	.12	.04
		.2	.06	.05	.12	.04	.69	.04
		.5	.05	.05	.40	.04	1.00	.04
	20	.1	.05	.06	.14	.05	.13	.06
		.2	.06	.06	.12	.05	.64	.06
		.5	.05	.06	.26	.05	.98	.06
	40	.1	.05	.06	.10	.07	.11	.08
		.2	.07	.06	.12	.07	.60	.07
		.5	.06	.06	.14	.07	1.00	.07
4,000	10	.1	.05	.05	.19	.05	.38	.06
		.2	.05	.06	.49	.05	1.00	.06
		.5	.06	.05	.91	.07	1.00	.05
	20	.1	.05	.05	.12	.05	.18	.12
		.2	.05	.05	.29	.05	.99	.12
		.5	.07	.05	.57	.05	1.00	.11
	40	.1	.05	.05	.12	.06	.20	.13
		.2	.06	.05	.19	.06	.95	.13
		.5	.06	.05	.27	.06	1.00	.13

Table 8 gives the outcomes of the tests of local independence. The first panel gives those for the tests of local independence between the item responses. The last four columns give the average item scores given an incorrect response (label  $U_{i-1} = 0$ ) and a correct response on the previous item (label  $U_{i-1} = 1$ ). Note that the observed and expected averages are very close. To assess the magnitude of these differences, the LM statistics and their significance probabilities in columns 2 and 3 should be used. None of the tests was significant at the 5% level. The second panel of Table 8 gives the outcomes of the tests of local independence between the responses and RTs. The last four columns give the mean log-time given an incorrect response (label  $U_i = 0$ ) and a correct response on the item (label  $U_i = 1$ ). The observed and expected average RTs are close. However, in the second and third column, it can be seen that 5 out of 12 tests are significant at the 5% level. In particular, the first and last item have highly significant LM tests. Finally, the last panel of Table 8 gives the outcomes of the tests for local independence of the RTs. The last column gives the correlations computed using (33) and (36). In particular, the correlation between the items 1 and 2 and between items 10 and 11 was relatively high.

**Table 7.** MML parameter estimates

Item	$a$	$SE(a)$	$b$	$SE(b)$	$\alpha$	$SE(\alpha)$	$\beta$	$SE(\beta)$
<i>Estimates of the item parameters</i>								
1	0.75	0.39	-1.46	0.28	1.76	0.02	-4.17	0.06
2	0.95	0.38	-0.99	0.25	1.60	0.06	-3.83	0.06
3	0.67	0.39	-1.87	0.33	0.88	0.07	-2.11	0.08
4	0.82	0.47	-1.62	0.39	1.06	0.14	-2.36	0.10
5	1.21	0.46	-2.02	0.37	2.13	0.02	-2.92	0.08
6	1.10	0.49	-1.79	0.34	1.11	0.04	-2.76	0.06
7	1.45	0.58	-1.73	0.41	1.65	0.02	-2.74	0.07
8	0.69	0.59	-1.42	0.31	1.49	0.03	-2.38	0.10
9	0.36	0.45	-1.87	0.35	1.12	0.03	-1.77	0.11
10	2.22	0.98	0.29	0.40	0.64	0.05	-1.80	0.10
11	0.86	0.54	-1.55	0.30	0.59	0.04	-2.03	0.07
12	0.66	0.70	-1.93	0.46	0.14	0.14	-1.50	0.21
<i>Estimates of correlations between item parameters</i>								
$a$	1.00							
$b$	.72	0.15	1.00					
$\alpha$	.03	0.18	-.17	0.18	1.00			
$\beta$	.01	0.11	.00	0.12	-.76	0.11	1.00	
<i>Estimates of correlation between person parameters</i>								
$\rho(\theta, \tau) = -.15, SE(\rho(\theta, \tau)) = 0.09$								

## 8. Discussion

The model considered in this paper implies some definite assumptions about the relation between speed and accuracy in a data set. It consists of two separate IRT measurement models, and represents a speed-accuracy distribution on a second level of modelling. The IRT measurement models imply the assumption of local independence; that is, given the IRT model parameters, all responses and RTs are assumed to be independent. The present paper presents a general framework for testing these explicit assumptions, not only the assumptions regarding local independence, but also the assumptions regarding parameter invariance and the shapes of the response functions. The estimation and testing procedure is established in a framework that has been well proven for other IRT models: MML estimation and LM tests. Besides parameter estimates, the MML estimation procedure also provides standard errors of the parameter estimates, and these can be used to test hypothesis concerning the level 2 distributions, for instance, the hypothesis that the covariances are zero.

As an alternative for the generally used MML procedure, Bayesian procedures in combination with MCMC computational methods have come into prominence. Bayesian estimation procedures for IRT models were first developed by Albert (1992). Simulation studies have shown that estimates obtained by the Bayesian approach for the standard IRT models (2PL, 3PL) are generally not superior to estimates obtained by the MML procedure (see, for instance, Baker, 1998). However, the Bayesian approach also applies to complicated IRT models, where the MML approach poses serious problems. Recently, a fully Bayesian approach has been adopted for the estimation of IRT models with multiple raters, multiple item types, and missing data (Patz & Junker, 1999), testlet structures (Bradlow, Wainer, & Wang, 1999), models with multi-level structure on the

**Table 8.** LM tests for local independence

Item	LM	$p$	$U_{i-1} = 0$		$U_{i-1} = 1$	
			Observed	Expected	Observed	Expected
<i>Responses</i>						
2	1.88	.17	0.61	0.66	0.79	0.77
3	0.49	.49	0.80	0.82	0.90	0.89
4	2.67	.10	0.71	0.77	0.87	0.84
5	0.00	.95	0.71	0.74	0.86	0.86
6	1.56	.21	0.72	0.72	0.88	0.87
7	0.46	.50	0.62	0.67	0.85	0.85
8	0.01	.93	0.80	0.75	0.86	0.86
9	3.51	.06	0.89	0.88	0.87	0.90
10	0.88	.35	0.69	0.60	0.64	0.64
11	0.17	.68	0.82	0.81	0.89	0.89
12	0.11	.74	0.90	0.86	0.92	0.91
			$U_i = 0$		$U_i = 1$	
<i>Responses and response times</i>						
1	449.03	.00	4.20	4.44	4.76	3.90
2	0.74	.39	4.24	4.08	3.40	3.44
3	7.02	.01	2.97	2.13	1.81	1.93
4	3.24	.07	2.60	2.17	1.94	2.03
5	6.82	.01	2.47	3.05	3.17	2.79
6	1.79	.18	3.45	2.81	2.50	2.61
7	0.09	.77	2.36	2.47	2.63	2.59
8	0.39	.53	3.17	2.31	4.08	2.72
9	2.07	.15	2.34	1.91	1.95	2.00
10	8.19	.00	2.30	2.02	1.75	1.90
11	0.02	.88	2.32	2.20	2.16	2.17
12	30.96	.00	1.81	1.60	1.37	1.55
<i>Items</i>			Residual correlation			
<i>Response times</i>						
1, 2	22.86	.00	-.34			
2, 3	0.14	.71	-.08			
3, 4	6.64	.01	-.15			
4, 5	4.27	.04	-.12			
5, 6	10.62	.00	.35			
6, 7	0.28	.60	-.14			
7, 8	0.33	.57	-.08			
8, 9	0.30	.59	-.07			
9, 10	0.00	.97	-.05			
10, 11	22.70	.00	-.32			
11, 12	9.40	.00	-.18			

*Note.* The LM statistics have one degree of freedom.

ability parameters (Fox & Glas, 2001) and the item parameters (Janssen, Tuerlinckx, Meulders, & de Boeck, 2000), and multidimensional IRT models (Béguin & Glas, 2001). The motivation for the recent interest in Bayesian inference using MCMC procedures is that the complex dependency structures in the models mentioned require

the evaluation of multiple integrals to solve the estimation equations in an MML framework. These problems are avoided in an MCMC framework. However, the integrals needed to estimate and test the model as it is considered here are only two-dimensional, so MCMC methods are not inevitable. Future research using simulation studies can shed more light on the relative performance of the two competing approaches in terms of the precision of item parameter estimation, the associated standard error estimates, and the computation time.

More important, however, is the circumstance that testing of model fit in a Bayesian framework is less developed than in a frequentist framework. A much used Bayesian diagnostic tool is the posterior predictive check, which is also applied to IRT models (Glas & Meijer, 2003; Hoijtink, 2001; Sinharay, 2005). However, Bayarri and Berger (2000) have shown that posterior predictive checks have less than adequate behaviour of posterior  $p$  values and often fail to detect model violations. Bayesian analogues of the tests presented here are therefore much needed.

## Acknowledgements

This study received funding from the Law School Admission Council (LSAC). We thank Dr Henk Kuyper of the National Institute for Educational Measurement (CITO) in The Netherlands for the use of his data. The opinions and conclusions contained in this paper are those of the authors and do not necessarily reflect the policy and position of LSAC or CITO.

## References

- Aitchison, J., & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics*, *29*, 813–828.
- Albert, J. H. (1992). Bayesian estimation of normal ogive item response functions using Gibbs sampling. *Journal of Educational Statistics*, *17*, 251–269.
- Baker, F. B. (1998). An investigation of item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement*, *22*, 153–169.
- Bayarri, M. J., & Berger, J. O. (2000).  $P$ -values for composite null models. *Journal of the American Statistical Association*, *95*, 1127–1142.
- Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541–562.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM-algorithm. *Psychometrika*, *46*, 443–459.
- Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman & Hall.
- 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.
- Efron, B. (1977). Discussion on 'Maximum likelihood from incomplete data via the EM algorithm' (by A. Dempster, N. Laird, and D. Rubin). *Journal of the Royal Statistical Society, Series B*, *39*, 29.
- Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 271–288.

- 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. (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., & Dagohey, A. V. T. (2007). Person fit tests for IRT models for polytomous items with estimated person and item parameters. *Psychometrika*, 72, 159-180.
- Glas, C. A. W., & Hendrawan, I. (2005). Testing linear models for ability parameters in item response models. *Multivariate Behavioral Research*, 40, 25-51.
- Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, 27, 217-233.
- Glas, C. A. W., & Suárez Falcón, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, 27, 87-106.
- Hojitink, 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.
- Janssen, R., Tuerlinckx, F., Meulders, M., & de Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics*, 25, 285-306.
- Lehmann, E. L. (1999). *Elements of large-sample theory*. New York: Springer.
- Lindgren, B. W. (1968). *Statistical theory* (3rd ed.). New York: Macmillan.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- 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.
- Luce, R. D. (1986). *Response times: Their roles in inferring elementary mental organization*. Oxford: Oxford University Press.
- Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24, 50-64.
- Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, 24, 342-366.
- Rao, C. R. (1947). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society*, 44, 50-57.
- Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In E. E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (pp. 151-171). Amsterdam: North-Holland.
- Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, 70, 533-555.
- Silvey, S. D. (1975). *Statistical inference*. London: Chapman & Hall.
- Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, 42, 375-394.
- Snijders, T. A. B., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.
- Sörbom, D. (1989). Model modification. *Psychometrika*, 54, 371-384.
- Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.
- van Breukelen, G. J. P. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika*, 70, 359-376.

- van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181–204.
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308.
- van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*, 5–20.
- van der Linden, W. J., & Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and response times on test items. *Psychometrika*, *75*(1), 120–139. doi:10.1007/s11336-009-9129-9
- Verhelst, N. D., Verstralen, H. H. F. M., & Jansen, M. G. (1997). A logistic model for time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169–185). New York: Springer.
- 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.

Received 17 February 2009; revised version received 30 September 2009