LSAC RESEARCH REPORT SERIES

A Study of Structural Modeling Using Plausible Value Imputation

Cees A. W. Glas
Hanneke Geerlings
University of Twente, Enschede, The Netherlands

Law School Admission Council
Research Report 08-03
March 2008

A publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation whose members are more than 200 law schools in the United States, Canada, and Australia. Headquartered in Newtown, PA, USA, the Council was founded in 1947 to facilitate the law school admission process. The Council has grown to provide numerous products and services to law schools and to more than 85,000 law school applicants each year.

All law schools approved by the American Bar Association (ABA) are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also members. Accredited law schools outside of the United States and Canada are eligible for membership at the discretion of the LSAC Board of Trustees.

This study is published and distributed by LSAC. The opinions and conclusions contained in this report are those of the author(s) and do not necessarily reflect the position or policy of LSAC.

Table of Contents

Executive Summary

In computerized adaptive testing, the responses of test takers to test questions (items) are used to select subsequent items for administration that are tailored to the test taker's ability level. A mathematical model called item response theory (IRT) is commonly used to estimate both the characteristics of the items and the ability level of the test takers. Data from computerized adaptive tests can be used to evaluate hypotheses about student proficiency, such as hypotheses about subgroup differences (gender, racial/ethnic group, previous education, preparatory training) or the development of their proficiencies. The hypotheses can be evaluated by adding a structural model for test taker ability to the IRT model. Typical examples of such structural models are (linear) analysis of variance and regression models.

The goal of this study is to compare the power of several common but complex methods for such extended models (i.e., marginal maximum or Bayesian methods) to a simpler alternative (i.e., plausible value imputation). The power of the methods is evaluated for IRT models for dichotomously and polytomously scored items, and for a model for responses to dichotomous items combined with response times. Simulation studies show that a relatively simple version of the plausible value imputation method does not perform worse than more advanced methods.

Abstract

Data from computerized adaptive tests can be used to evaluate hypotheses about student proficiency, such as hypotheses about differences between groups of students (e.g., gender, racial/ethnic group, previous education, preparatory training) and hypotheses concerning the development of proficiency. Such hypotheses can be evaluated by analysis of variance and regression models for item response theory (IRT) proficiency parameters. Several methods for the estimation of such models are compared with respect to their power for the detection of effects: methods based on plausible value imputation and methods based on marginal maximum likelihood estimation. The power of the methods is evaluated for models for dichotomously scored items, polytomously scored items, and an IRT model for responses to dichotomous items combined with the RTs. Simulation studies show that a relatively simple plausible value imputation method that ignores the covariance between measurement occasions does not perform worse than more advanced methods.

Introduction

Data obtained using a computerized adaptive test (CAT) can be used to evaluate hypotheses about student proficiency, such as hypotheses about differences between groups of students (e.g., gender, racial/ethnic group, previous education, preparatory training) and hypotheses concerning the development of proficiency. Such hypotheses can be tested using analysis of variance and regression models for item response theory (IRT) proficiency parameters. Several methods are available to estimate the parameters of these models. The parameters of the IRT measurement model (item parameters) and the structural model (regression parameters) can be estimated concurrently using marginal maximum likelihood (MML, Mislevy & Bock, 1989). This estimation procedure can also be divided into two separate steps, where the parameters of the measurement model are estimated first, followed by MML estimation of the structural parameters treating the estimated item parameters as known constants. This procedure will be labeled two-step MML (MML2). As an alternative, Fox and Glas (2001, 2002, 2003) considered concurrent estimation in a fully Bayesian framework where they made computations using the Gibbs sampler. The advantage of MML and Bayesian methods is that they are based on a well-founded statistical framework. Disadvantages are the numerical complexity of the methods, the need to use specialized and not readily available software, and possible confounding of the fit of the measurement model and the structural model. A much used alternative is a method using multiple imputations, generally known as plausible value imputation (Mislevy, 1991). The method consists of three steps:

1. The IRT model is estimated and validated.

2. Values of the student parameters are drawn from their posterior distribution or their sample distribution.

3. These so-called plausible values are imputed into the structural model (e.g., analysis of variance model).

Plausible values rather than maximum likelihood estimates or Bayesian estimates are imputed to account for the estimation error of these parameters. An advantage of the method is that the last step can be performed using standard user-software, such as SPSS, SAS, or STATA.

The methods and simulation studies presented here are an extension of the work by Holman, Glas, and de Haan (2003). These authors examined the power of the MML2 method in a two-legged trial with the two-parameter logistic (2PL) model as measurement model. They concluded that the number of respondents in each arm of a randomized trial varies with the number of items used. They also concluded that as long as 20 dichotomously scored items are used, the

number of items barely affects the number of respondents needed to detect effect sizes of 0.5 and 0.8 with a power of 80%.

In the present paper, their research is generalized in several directions. With respect to estimation methods, concurrent MML estimation and two versions of plausible value imputation are considered. The design is generalized to a two-way longitudinal design. Further, three models for polytomously scored items (the generalized partial credit model, the sequential model, and the graded response model) and a combined IRT model for accuracy and speed are considered.

This report is organized as follows. First, the IRT models are presented and the estimation methods outlined. Then, the simulation studies are presented and some conclusions drawn.

Item Response Theory

Item Response Models for Dichotomously Scored Response Data

IRT models relate discrete dichotomously or polytomously scored responses to latent respondent variables. The family of IRT models is by now quite big (for an overview, see de Boeck & Wilson, 2004; Skrondal & Rabe-Hesketh, 2004; van der Linden & Hambleton, 1997). We will first focus here on the basic IRT model most used in CAT, the 2PL model (Birnbaum, 1968). The model pertains to dichotomously scored responses to items that will be indexed $i = 1, \ldots, K$. The respondents will be indexed $n = 1, \ldots, N$. The responses are coded $U_{ni} = 0$ and $U_{ni} = 1$ and are modeled by a unidimensional latent variable θ_n. Using the abbreviation for the logistic function given by

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

in the 2PL model, the probability of a correct response on item i is given by

$$p_i(\theta_n) = \Pr(U_{ni} = 1 \mid \theta_n) = \Psi(a_i(\theta_n - b_i)), \tag{2}$$

where $\theta_n \in (-\infty, \infty)$ is the parameter representing the trait level of the respondent and $b_i \in (-\infty, \infty)$ and $a_i \in [o, \infty)$ represent the location and the discriminating power of item i, respectively.

Item Response Models for Polytomously Scored Response Data

Consider a test with polytomously scored items labeled $i = 1, \ldots, K$. Every item has response categories labeled $j = 0, \ldots, m_i$. Item responses will be coded by stochastic variables U_{nij} ($n = 1, \ldots, N$, $i = 1, \ldots, K$; $j = 0, \ldots, m_i$; in the sequel the index i of m_i is dropped for convenience) with realizations u_{nij}. Assume that $u_{nij} = 1$ if a response was given in category j, and zero otherwise. It will be assumed that the response categories are ordered, and that there exists a latent ability variable θ_n such that a response in a higher category reflects a higher ability level than a response in a lower category. The probability of scoring in a response category j on item i is given by a response function $P_{ij}(\theta_n) = P(U_{nij} = 1 \mid \theta_n)$. In many measurement situations, such as in measurement of abilities, it is reasonable to assume that the response function of the category $j = 0$ decreases as a function of ability, the response function for $j = m$ increases as a function of ability, and the response functions of the intermediate categories are single peaked. Mellenbergh (1995) showed that IRT models with such response functions can be divided into three classes. Though the rationales underlying the models in these classes are very different, their response functions appear to be very close (Verhelst, Glas, & de Vries, 1997), so the models might be hard to distinguish on the basis of empirical data. We will now introduce three models from the three classes distinguished by Mellenbergh (1995).

The Graded Response Model

In the graded response model (GRM), the probability of a response in category j of item i, $P(u_{nij} = 1 \mid \theta_n)$, is given by

$$P_{ij}(\theta_n) = \begin{cases} 1 - \Psi(a_i\theta_n - b_{i1}) & \text{if } j = 0 \\ \Psi(a_i\theta_n - b_{ij}) - \Psi(a_i\theta_n - b_{i(j+1)}) & \text{if } 0 < j < m \\ \Psi(a_i\theta_n - b_{im}) & \text{if } j = m \end{cases} \tag{3}$$

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

The Sequential Model

In the sequential model (SEQM, Tutz, 1990, 1997) the probability of a response in category j of item i is given by

$$P_{ij}(\theta_n) = \begin{cases} 1-\Psi(a_i\theta_n - b_{i1}) & \text{if } j = 0 \\ \prod_{h=1}^{j}\Psi(a_i\theta_n - b_{ih})\left[1-(\Psi(a_i\theta_n - b_{i(j+1)}))\right] & \text{if } 0 < j < m \\ \prod_{h=1}^{m}\Psi(a_i\theta_n - b_{ih}) & \text{if } j = m\ . \end{cases} \tag{4}$$

Verhelst, Glas, and de Vries (1997) noted that in the SEQM, every polytomously scored item can be viewed as a sequence of virtual dichotomously scored items. These virtual items are considered to be presented to the respondent as long as a correct response is given, and the presentation stops when an incorrect response is given. An important consequence of this conceptualization of the response process is that estimation and testing procedures for the 2PL model with incomplete data can be directly applied to the SEQM.

The Generalized Partial Credit Model

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

$$P_{ij}(\theta_n) = \frac{\exp(ja_i\theta_n - b_{ij})}{1+\sum_{h=1}^{m}\exp(ha_i\theta_n - b_{ih})}. \tag{5}$$

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

Multidimensional Generalizations

In many situations, the assumption that an individual's response behavior can be explained by a unidimensional student parameter θ_n does not hold. In that case the assumption of a unidimensional student parameter can be replaced by the assumption of a multidimensional student parameter $\theta_{n1},\ldots,\theta_{nq},\ldots,\theta_{nQ}$. The multidimensional versions of the models given by (3)–(5) are defined by replacing $a_i\theta_n$ by

$$\sum_{q=1}^{Q} a_{iq}\theta_{nq}\,.$$

Further, it is usually assumed that the parameters $\theta_{n1},\ldots,\theta_{nq},\ldots,\theta_{nQ}$ have a joint Q -variate normal distribution (McDonald, 1997; Reckase, 1997).

A Combined Model for Responses and Response Times

The model for accuracy and speed used in this report was developed by van der Linden (2005, 2006). It is a hierarchical model consisting of four different components on two levels. The first-level models are for the distributions of the responses and the response times (RTs) for a fixed student on a fixed item. The second-level models are for the joint distribution of the student parameters in the two first-level models in some student population and the distribution of the item parameters and the item domain.

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

$$\Pr\{U_{ni} = 1\} = c_i + (1-c_i)\psi(a_i(\theta_{n1} - b_i)), \tag{6}$$

where $\theta_{n1} \in \mathbb{R}$ is the ability of student n; and $b_i \in \mathbb{R}$, $a_i \in \mathbb{R}^+$, and $c_i \in [0,1]$ are the difficulty, discrimination, and guessing parameter, respectively, for item i.

For the distribution of RT T_{ni} of student n on item i, we use the lognormal model

$$f(t_{ni} \mid \theta_{n2}, d_i, e_i) = \frac{d_i}{t_{ni}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}[d_i(\log t_{ni} - (e_i - \theta_{n2}))]^2\right\}, \tag{7}$$

where $\theta_{n2} \in \mathbb{R}$ is the speed at which student n operates on the test, $e_i \in \mathbb{R}$ is the time intensity of item i, and $d_i \in \mathbb{R}^+$ is its discrimination parameter. 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 student parameters are independent and identically distributed (i.i.d.) samples from a bivariate normal distribution, that is,

$$\boldsymbol{\theta}_n = (\theta_{n1}, \theta_{n2}) \sim \text{MVN}(\boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}), \tag{8}$$

and the first-level item parameters are i.i.d. samples from a multivariate normal distribution

$$\boldsymbol{\xi}_i = (a_i, b_i, c_i, d_i, e_i) \sim \text{MVN}(\boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}). \tag{9}$$

The model is identified by setting the mean of the distribution of the latent trait variables equal to zero—that is, using $\boldsymbol{\mu}_{\mathcal{P}} = 0$—and the diagonal of the covariance matrix equal to one—that is, $diag(\boldsymbol{\Sigma}_{\mathcal{P}}) = \mathbf{1}$. So $\boldsymbol{\Sigma}_{\mathcal{P}}$ becomes a correlation matrix. Note that the latent variables $\boldsymbol{\theta}_n$ are independent between respondents but dependent within respondents.

Linear Models on the Latent Variables

Longitudinal data can be analyzed in the framework of multilevel models (Bryke & Raudenbush, 1992, Goldstein, 1987; Longford, 1993) where occasions are nested within respondents. Examples of applications of the multilevel paradigm in the field of IRT can be found in Mislevy and Bock (1989) and Fox and Glas (2001, 2002, 2003). Assume that θ_{nt} are latent variables of students $n = 1, \ldots, N$ related to measures $t = 1, \ldots, T$. The measures may, for example, relate to different time points, or different ability dimensions, or to a combination of the two. We impose a regression model on the latent variables given by

$$\theta_{nt} = \sum_{p=1}^{P} \beta_p x_{ntp} + \varepsilon_{nt},$$

where x_{ntp} are observations on P covariates. In matrix notation we have,

$$\boldsymbol{\theta}_n = \mathbf{X}_n \boldsymbol{\beta} + \boldsymbol{\varepsilon}_n.$$

The error terms have a multivariate normal distribution, that is,

$$\boldsymbol{\varepsilon}_n \sim N(\mathbf{0}, \boldsymbol{\Sigma}).$$

The covariance matrix $\boldsymbol{\Sigma}$ need not be diagonal; for instance, in longitudinal designs it is usually assumed that the errors are autocorrelated.

Estimation Methods

Marginal Maximum Likelihood

MML estimation is a much used technique for item calibration. For the 2PL and 3PL models, the theory was developed by Bock and Aitkin (1981). Under the label "Full Information Factor Analysis," a multidimensional version of the 2PL model and 3PL model was developed by Bock, Gibbons, and Muraki (1988). (See also Ackerman, 1996a, 1996b; and Reckase, 1985, 1997.) Glas and van der Linden (2006) presented an MML estimation procedure for the hierarchical model for accuracy and speed. Further, for this model, they developed Lagrange multiplier tests to test three different assumptions of conditional independence in the model (Glas & van der Linden, 2006; van der Linden & Glas, 2006).

MML estimates of the regression coefficients $\boldsymbol{\beta}$ and the covariance matrix $\boldsymbol{\Sigma}$ can be obtained either concurrently with the item parameters or by treating the item parameters as known constants. Both approaches are based on the following. In general, let $\mathbf{u}_n$ be the vector of the item responses and RTs of student n; that is, $\mathbf{u}_n = (u_{n11}, \ldots, u_{nij}, \ldots, u_{nKm}, t_{n1}, \ldots, t_{ni}, \ldots, t_{nK})$. Using the assumption of local independence, the probability of a response pattern $\mathbf{u}_n$ is given by

$$P(\mathbf{u}_n \mid \boldsymbol{\theta}_n, \boldsymbol{\gamma}) = \prod_{i=1}^{K} \prod_{j=0}^{m} P(U_{nij} = u_{nij} \mid \boldsymbol{\theta}_n, \boldsymbol{\gamma}_i) f(t_{ni} \mid \boldsymbol{\theta}_n, \boldsymbol{\gamma}_i), \tag{10}$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_i$ are vectors of item parameters.

The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ can be estimated using an alternating generalized least-squares algorithm with the steps

$$\boldsymbol{\beta} = (\sum_n \mathbf{X}_n' \boldsymbol{\Sigma}^{-1} \mathbf{X}_n)^{-1} \sum_n \mathbf{X}_n' \boldsymbol{\Sigma}^{-1} \boldsymbol{\eta}_n,$$

where we estimate the covariance matrix of the residuals as

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_n \mathbf{H}_n - (\mathbf{X}_n \boldsymbol{\beta})(\mathbf{X}_n \boldsymbol{\beta})'$$

with

$$\boldsymbol{\eta}_n = \int \cdots \int \boldsymbol{\theta}_n p(\boldsymbol{\theta}_n \mid \mathbf{u}_n, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X}) d\boldsymbol{\theta}_n,$$

and

$$\mathbf{H}_n = \int \cdots \int \boldsymbol{\theta}_n \boldsymbol{\theta}_n' p(\boldsymbol{\theta}_n \mid \mathbf{u}_n, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X}) d\boldsymbol{\theta}_n,$$

for $n = 1, \ldots, N$, where $p(\boldsymbol{\theta}_n \mid \mathbf{u}_n, \mathbf{b}, \mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X})$ is the posterior distribution of $\boldsymbol{\theta}_n$ given by

$$p(\boldsymbol{\theta}_n \mid \mathbf{u}_n, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X}) \propto P(\mathbf{u}_n \mid \boldsymbol{\theta}_n, \boldsymbol{\gamma}) p(\boldsymbol{\theta}_n \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}).$$

In MML, the item parameters $\boldsymbol{\gamma}$ are imputed as constants in the estimation equations for the regression coefficients $\boldsymbol{\beta}$ and the covariance matrix $\boldsymbol{\Sigma}$. For concurrent MML estimation, the estimation equations for the regression coefficients $\boldsymbol{\beta}$ and the covariance matrix $\boldsymbol{\Sigma}$ are solved concurrently with the MML estimation equations for the item parameters $\boldsymbol{\gamma}$.

For solving the estimation equations, the expectation-maximization (EM) algorithm by Dempster, Laird, and Rubin (1977) can be used. The E-step consists of computing $\boldsymbol{\eta}_n$ and $\mathbf{H}_n$ for $n = 1, \ldots, N$, and the M-step consists of solving the equations for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. These new estimates are then inserted in the E-step and the whole process is iterated until convergence is achieved. The multiple integrals that appear 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 analyzed simultaneously. Wood et al. (2002) indicated that the maximum number of factors is 10 with adaptive quadrature, 5 with nonadaptive quadrature, and 15 with Monte Carlo integration. Recently, alternative procedures based on Monte Carlo integration and importance sampling were suggested by Fox (2003) and van Davier and Sinharay (2007).

Plausible Value Imputation

The use of the methods described in the previous section depends on specialized software that is not readily available to the practitioner. However, most data analysts only have standard software, such as SPSS, SAS, and STATA available. The problem is solved by using an alternative approach that is known as multiple imputation or plausible value imputation (Mislevy, 1991). Plausible values are random draws from a student's posterior distribution, $p(\boldsymbol{\theta}_n \mid \mathbf{u}_n, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X})$. Usually, three to five draws are taken from the posterior distribution for each student, and the regression model $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is estimated for each of the draws. The variance in the estimates of the regression parameters $\boldsymbol{\beta}$ gives an indication of the uncertainty attributable to the uncertainty of the estimates of $\boldsymbol{\theta}$. If an explicit estimate of this uncertainty is not needed, but the uncertainty only has to be taken into account in the estimation of $\boldsymbol{\beta}$, one draw for every student is sufficient. This approach is, for instance, used in the analyses in the National Assessment of Educational Progress (NAEP, Mislevy, Johnson, & Muraki, 1992).

In the case of between-items multidimensionality, the drawing of plausible values can be done using two methods. In the first method, the item and population parameters are estimated for each dimension separately, and the plausible values for every student and for every dimension are drawn separately. In the second method, the values of the item and population parameters are estimated concurrently under the assumption that the students' ability parameters are multivariate normally distributed, and the plausible values for every student are drawn concurrently for all dimensions, thereby taking the covariance between the dimensions into account. The plausible values can be drawn with a sampling-importance-resampling (SIR) algorithm (Thomas & Gan, 1997; also see Gelfan & Smith, 1990; and Tanner, 1993).

Simulation Studies

Dichotomously Scored Items, Two Groups

The first set of simulations pertains to replication of some of the simulations done by Holman, Glas, and de Haan (2003). These authors used the MML2 method, in which item parameters are fixed. In the present report, concurrent MML estimation and plausible value imputation are also considered. Holman et al. considered a design with 30, 40, 50, 100, 200, 300, 500, or 1,000 simulees in each of two groups and test lengths of 5, 10, 15, 20, 30, 50, 70, or 100 dichotomously scored items. It is beyond the scope of the present report to evaluate all these conditions, so here we restrict ourselves to 5, 10, 20, and 50 items, and sample sizes of 30, 50, 100, and 500 respondents in each group. Other aspects of the present study were analogous to the study by Holman et al. That is, the ability parameters for both groups were sampled from normal distributions with standard deviations equal to one. The means were equal to zero for the first group and equal to some effect size for the second group. These effect sizes were equal to 0.0, 0.2, and 0.5. Cohen (1988) labeled the latter two effects as a small and moderate effect, respectively.

TABLE 1
Type I error rate and power of a test for the difference between the means of two ability distributions for a test length of 5 and 10 items

			True Theta			Plausible Value			MML Estimate		
Significance Level:			0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
K	*N*	Effect Size									
5	30	0.0	0.11	0.05	0.01	0.11	0.07	0.01	0.10	0.06	0.01
		0.2	0.30	0.21	0.10	0.23	0.15	0.06	0.32	0.23	0.11
		0.5	0.85	0.78	0.57	0.68	0.58	0.37	0.87	0.81	0.57
	50	0.0	0.11	0.05	0.02	0.13	0.05	0.02	0.11	0.07	0.02
		0.2	0.42	0.31	0.14	0.30	0.23	0.11	0.44	0.31	0.14
		0.5	0.98	0.96	0.84	0.83	0.74	0.56	0.96	0.93	0.82
	100	0.0	0.10	0.06	0.02	0.13	0.06	0.01	0.09	0.05	0.01
		0.2	0.64	0.50	0.28	0.48	0.37	0.15	0.63	0.50	0.29
		0.5	1.00	1.00	0.99	0.98	0.94	0.87	1.00	1.00	0.99
	500	0.0	0.11	0.07	0.02	0.19	0.13	0.06	0.14	0.06	0.01
		0.2	1.00	0.99	0.96	0.91	0.86	0.73	1.00	1.00	0.98
		0.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10	30	0.0	0.11	0.06	0.02	0.11	0.05	0.01	0.10	0.06	0.01
		0.2	0.32	0.22	0.10	0.27	0.20	0.09	0.32	0.23	0.10
		0.5	0.86	0.80	0.57	0.77	0.67	0.46	0.87	0.79	0.57
	50	0.0	0.11	0.04	0.01	0.11	0.06	0.02	0.10	0.06	0.02
		0.2	0.48	0.37	0.19	0.40	0.30	0.14	0.47	0.36	0.18
		0.5	0.99	0.97	0.85	0.94	0.89	0.72	0.99	0.97	0.86
	100	0.0	0.10	0.04	0.01	0.12	0.06	0.01	0.13	0.07	0.01
		0.2	0.67	0.55	0.29	0.53	0.43	0.24	0.64	0.52	0.30
		0.5	1.00	1.00	1.00	1.00	0.99	0.95	1.00	1.00	0.99
	500	0.0	0.12	0.05	0.02	0.15	0.09	0.03	0.10	0.05	0.01
		0.2	1.00	1.00	0.95	0.99	0.95	0.87	1.00	0.99	0.98
		0.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

TABLE 2
Type I error rate and power of a test for the difference between the means of two ability distributions for a test length of 20 and 50 items

			True Theta			Plausible Value			MML Estimate		
Significance Level:			0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
K	*N*	Effect Size									
20	30	0.0	0.11	0.07	0.02	0.12	0.07	0.01	0.12	0.06	0.02
		0.2	0.32	0.22	0.09	0.30	0.20	0.09	0.31	0.20	0.09
		0.5	0.88	0.79	0.58	0.81	0.73	0.50	0.89	0.81	0.58
	50	0.0	0.10	0.04	0.01	0.09	0.04	0.01	0.09	0.04	0.01
		0.2	0.42	0.31	0.13	0.40	0.28	0.10	0.46	0.31	0.15
		0.5	0.97	0.95	0.85	0.95	0.91	0.77	0.98	0.95	0.85
	100	0.0	0.12	0.07	0.02	0.12	0.06	0.02	0.12	0.07	0.02
		0.2	0.67	0.55	0.32	0.62	0.51	0.28	0.67	0.55	0.34
		0.5	1.00	1.00	0.99	1.00	1.00	0.97	1.00	1.00	0.99
	500	0.0	0.11	0.05	0.01	0.10	0.06	0.02	0.10	0.05	0.01
		0.2	1.00	1.00	0.98	0.99	0.99	0.95	1.00	1.00	0.98
		0.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
50	30	0.0	0.11	0.06	0.02	0.11	0.06	0.02	0.11	0.06	0.01
		0.2	0.33	0.22	0.08	0.31	0.21	0.08	0.33	0.23	0.08
		0.5	0.88	0.80	0.60	0.87	0.77	0.56	0.88	0.80	0.61
	50	0.0	0.10	0.06	0.02	0.10	0.06	0.01	0.09	0.06	0.01
		0.2	0.43	0.28	0.14	0.41	0.26	0.11	0.42	0.28	0.13
		0.5	0.98	0.96	0.84	0.97	0.94	0.81	0.98	0.95	0.85
	100	0.0	0.10	0.04	0.01	0.12	0.06	0.02	0.11	0.06	0.01
		0.2	0.61	0.49	0.27	0.59	0.45	0.24	0.62	0.50	0.27
		0.5	1.00	1.00	0.99	1.00	1.00	0.99	1.00	1.00	0.99
	500	0.0	0.10	0.05	0.01	0.10	0.05	0.01	0.08	0.04	0.01
		0.2	1.00	1.00	0.97	1.00	0.99	0.96	1.00	1.00	0.97
		0.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

The responses were generated according to the 2PL model. In each replication, the item difficulty parameters were drawn from a standard normal distribution, and the item discrimination parameters were drawn from a lognormal distribution with a mean of 0.2 and a variance of 1.0.

The results obtained using 1,000 replications are given in Tables 1 and 2. These results can be compared with the results obtained using MML2 reported by Holman et al. (2003) in their Table 3. The first three columns give the test length, sample sizes in each group, and effect sizes, respectively. The other columns give the proportion of replications in the 1,000 replications where the test statistic was significant at 10%, 5%, and 1% using a two-sided test. The statistic was the estimated difference of the means divided by the standard deviation of this difference. The significance probability was computed assuming that the statistic had a normal distribution. For the columns under the heading "True Theta," the true generating values of the abilities were used to compute least-squares estimates of the mean of the abilities of the second group and the standard deviations of the abilities of the two groups. The mean of the first group was fixed to zero to identify the model. For the next three columns, the same was done using plausible values. Finally, the three last columns give the significance proportions obtained using concurrent MML estimates.

In the rows for an effect size equal to 0.0, it can be seen that the control of Type I error rate was excellent for all conditions. Further, with respect to the power, there were the usual main effects of the effect size, the sample size, and the number of items. The power was highest when the true theta values were used as input for the estimation of the difference between the two means, followed by the power of the MML method and the power of the plausible value method. Note, however, that the differences in power were very small. On the other hand, the reader can verify in the article by Holman et al. that the powers obtained using the MML2 method (i.e., the two-step method, in which the item parameters are treated as fixed constants) were substantially lower. Some examples of the differences are displayed in Table 3.

TABLE 3
The difference in power between two versions of MML and plausible value imputation for an effect size of 0.2

K	*N*	Plausible Values	MML	MML2
5	50	0.23	0.31	0.08
	100	0.37	0.50	0.14
10	50	0.30	0.36	0.11
	100	0.43	0.52	0.17
20	50	0.28	0.31	0.13
	100	0.51	0.55	0.22
50	50	0.36	0.28	0.13
	100	0.45	0.50	0.23

The reason for this difference may have to do with the following. Holman et al. note that the estimate of the difference between the means and its standard error are complex functions of the item and population parameters. Treating the item parameters as fixed rather than using MML estimates of these parameters obviously leads to loss of power.

Dichotomously Scored Items, Two Groups and Two Time Points

In the next set of simulation studies, the design was expanded to two groups measured at two time points using the same set of items. Other features of the simulation design were analogous to the set-up in the previous paragraph. Table 4 shows the results for a model with a main group effect equal to 0.2 or 0.5 and a zero time effect, Table 5 shows the results for a model with a main time effect equal to 0.2 or 0.5 and a zero group effect, and in Table 6 both main effects are either equal to 0.2 or 0.5. The significance level of the two-sided test was 5%. Besides the usual effects of test length, sample size, and effect size, the following effects are of interest. First, the power of all three methods was very close, but the power of MML was slightly lower than the power for the two methods based on plausible values. Second, the control of the Type I error rate for the zero effect, that is, the time effect in Table 4 and the group effect in Table 5, was excellent: The proportions of significant tests in the replications were very close to the nominal significance level of 5%. Finally, overall, the power of the test for the time effect increased as the correlation between the two time points increased from 0.2 to 0.4.

TABLE 4

Type I error rate for time effect and power for group effect for different test lengths, sample sizes, effect sizes, autocorrelations, and estimation methods

				Plausible Value Unidimensional		Plausible Value Multidimensional		MML Estimate	
K	*N*	Effect Size	Autocorr.	Time Effect	Group Effect	Time Effect	Group Effect	Time Effect	Group Effect
10	50	0.2	0.2	0.06	0.41	0.06	0.37	0.07	0.32
			0.4	0.06	0.37	0.06	0.37	0.06	0.30
		0.5	0.2	0.06	0.98	0.06	0.97	0.06	0.93
			0.4	0.05	0.98	0.05	0.96	0.04	0.92
	100	0.2	0.2	0.06	0.64	0.06	0.63	0.06	0.46
			0.4	0.06	0.64	0.06	0.64	0.04	0.47
		0.5	0.2	0.06	1.00	0.05	1.00	0.04	1.00
			0.4	0.07	1.00	0.06	1.00	0.04	1.00
	200	0.2	0.2	0.05	0.93	0.06	0.90	0.05	0.81
			0.4	0.05	0.92	0.05	0.89	0.04	0.77
		0.5	0.2	0.05	1.00	0.05	1.00	0.04	1.00
			0.4	0.06	1.00	0.05	1.00	0.05	1.00
20	50	0.2	0.2	0.05	0.36	0.05	0.35	0.05	0.35
			0.4	0.06	0.37	0.07	0.36	0.04	0.29
		0.5	0.2	0.04	0.98	0.05	0.98	0.07	0.97
			0.4	0.06	0.97	0.05	0.97	0.06	0.96
	100	0.2	0.2	0.07	0.68	0.05	0.65	0.07	0.57
			0.4	0.04	0.64	0.05	0.62	0.06	0.52
		0.5	0.2	0.05	1.00	0.04	1.00	0.07	1.00
			0.4	0.06	1.00	0.05	1.00	0.05	1.00
	200	0.2	0.2	0.05	0.91	0.05	0.88	0.05	0.83
			0.4	0.04	0.89	0.05	0.85	0.05	0.82
		0.5	0.2	0.05	1.00	0.05	1.00	0.05	1.00
			0.4	0.06	1.00	0.05	1.00	0.05	1.00
40	50	0.2	0.2	0.06	0.39	0.05	0.37	0.04	0.36
			0.4	0.05	0.36	0.05	0.34	0.05	0.33
		0.5	0.2	0.06	0.99	0.06	0.98	0.04	0.98
			0.4	0.05	0.98	0.04	0.98	0.05	0.98
	100	0.2	0.2	0.07	0.64	0.05	0.65	0.04	0.59
			0.4	0.06	0.61	0.06	0.59	0.04	0.54
		0.5	0.2	0.06	1.00	0.05	1.00	0.04	1.00
			0.4	0.06	1.00	0.06	1.00	0.06	1.00
	200	0.2	0.2	0.05	0.90	0.05	0.89	0.04	0.88
			0.4	0.04	0.90	0.05	0.88	0.02	0.82
		0.5	0.2	0.05	1.00	0.04	1.00	0.03	1.00
			0.4	0.05	1.00	0.06	1.00	0.03	1.00

TABLE 5
Type I error rate for group effect and power for time effect for different test lengths, sample sizes, effect sizes, autocorrelations, and estimation methods

				Plausible Value Unidimensional		Plausible Value Multidimensional		MML Estimate	
K	*N*	Effect Size	Autocorr.	Time Effect	Group Effect	Time Effect	Group Effect	Time Effect	Group Effect
10	50	0.2	0.2	0.42	0.05	0.46	0.06	0.34	0.06
			0.4	0.45	0.06	0.54	0.05	0.36	0.07
		0.5	0.2	0.99	0.05	0.99	0.07	0.96	0.06
			0.4	0.99	0.06	1.00	0.05	0.98	0.05
	100	0.2	0.2	0.70	0.04	0.73	0.06	0.62	0.05
			0.4	0.74	0.05	0.80	0.04	0.62	0.04
		0.5	0.2	1.00	0.04	1.00	0.04	1.00	0.06
			0.4	1.00	0.04	1.00	0.04	1.00	0.04
	200	0.2	0.2	0.94	0.05	0.95	0.05	0.93	0.04
			0.4	0.95	0.06	0.98	0.06	0.90	0.05
		0.5	0.2	1.00	0.05	1.00	0.07	1.00	0.04
			0.4	1.00	0.05	1.00	0.05	1.00	0.05
20	50	0.2	0.2	0.50	0.06	0.50	0.07	0.32	0.06
			0.4	0.51	0.06	0.57	0.06	0.43	0.06
		0.5	0.2	0.99	0.06	1.00	0.06	0.99	0.04
			0.4	1.00	0.05	1.00	0.06	0.99	0.05
	100	0.2	0.2	0.73	0.07	0.75	0.06	0.61	0.07
			0.4	0.78	0.05	0.82	0.05	0.71	0.06
		0.5	0.2	1.00	0.05	1.00	0.05	1.00	0.05
			0.4	1.00	0.04	1.00	0.05	1.00	0.04
	200	0.2	0.2	0.96	0.05	0.96	0.04	0.92	0.05
			0.4	0.97	0.04	0.98	0.04	0.97	0.05
		0.5	0.2	1.00	0.05	1.00	0.05	1.00	0.05
			0.4	1.00	0.05	1.00	0.06	1.00	0.07
40	50	0.2	0.2	0.42	0.06	0.43	0.06	0.40	0.05
			0.4	0.48	0.07	0.54	0.05	0.51	0.04
		0.5	0.2	0.99	0.06	1.00	0.06	1.00	0.05
			0.4	1.00	0.06	1.00	0.06	1.00	0.04
	100	0.2	0.2	0.72	0.05	0.74	0.05	0.72	0.06
			0.4	0.77	0.04	0.80	0.06	0.81	0.05
		0.5	0.2	1.00	0.05	1.00	0.05	1.00	0.04
			0.4	1.00	0.04	1.00	0.05	1.00	0.04
	200	0.2	0.2	0.96	0.05	0.96	0.05	0.94	0.06
			0.4	0.98	0.06	0.99	0.05	0.97	0.04
		0.5	0.2	1.00	0.05	1.00	0.04	1.00	0.05
			0.4	1.00	0.05	1.00	0.05	1.00	0.06

TABLE 6
Power for time effect and group effect for different test lengths, sample sizes, effect sizes, autocorrelations, and estimation methods

				Plausible Value Unidimensional		Plausible Value Multidimensional		MML Estimate	
K	*N*	Effect Size	Autocorr.	Time Effect	Group Effect	Time Effect	Group Effect	Time Effect	Group Effect
10	50	0.2	0.2	0.43	0.38	0.45	0.38	0.35	0.29
			0.4	0.44	0.39	0.52	0.38	0.35	0.26
		0.5	0.2	0.99	0.99	1.00	0.98	0.96	0.94
			0.4	0.99	0.97	1.00	0.95	0.98	0.93
	100	0.2	0.2	0.74	0.66	0.75	0.64	0.57	0.49
			0.4	0.71	0.65	0.79	0.63	0.68	0.43
		0.5	0.2	1.00	1.00	1.00	1.00	1.00	1.00
			0.4	1.00	1.00	1.00	1.00	1.00	1.00
	200	0.2	0.2	0.93	0.94	0.95	0.91	0.90	0.81
			0.4	0.95	0.91	0.98	0.90	0.94	0.79
		0.5	0.2	1.00	1.00	1.00	1.00	1.00	1.00
			0.4	1.00	1.00	1.00	1.00	1.00	1.00
20	50	0.2	0.2	0.46	0.38	0.46	0.38	0.39	0.28
			0.4	0.50	0.37	0.54	0.37	0.45	0.28
		0.5	0.2	0.99	0.98	1.00	0.98	0.99	0.96
			0.4	1.00	0.97	1.00	0.97	0.99	0.95
	100	0.2	0.2	0.72	0.67	0.74	0.63	0.68	0.57
			0.4	0.77	0.63	0.83	0.60	0.73	0.51
		0.5	0.2	1.00	1.00	1.00	1.00	1.00	1.00
			0.4	1.00	1.00	1.00	1.00	1.00	1.00
	200	0.2	0.2	0.96	0.90	0.96	0.89	0.91	0.84
			0.4	0.97	0.91	0.98	0.89	0.93	0.83
		0.5	0.2	1.00	1.00	1.00	1.00	1.00	1.00
			0.4	1.00	1.00	1.00	1.00	1.00	1.00
40	50	0.2	0.2	0.41	0.40	0.41	0.40	0.40	0.40
			0.4	0.50	0.36	0.55	0.38	0.51	0.33
		0.5	0.2	0.99	0.98	1.00	0.98	1.00	0.99
			0.4	1.00	0.97	1.00	0.97	1.00	0.97
	100	0.2	0.2	0.72	0.64	0.75	0.61	0.69	0.58
			0.4	0.79	0.61	0.84	0.58	0.80	0.51
		0.5	0.2	1.00	1.00	1.00	1.00	1.00	1.00
			0.4	1.00	1.00	1.00	1.00	1.00	1.00
	200	0.2	0.2	0.96	0.91	0.96	0.90	0.93	0.88
			0.4	0.98	0.90	0.98	0.89	0.97	0.86
		0.5	0.2	1.00	1.00	1.00	1.00	1.00	1.00
			0.4	1.00	1.00	1.00	1.00	1.00	1.00

To verify this phenomenon further, a set of simulations was run with the same setup, only the autocorrelations were varied as 0.2, 0.4, 0.6, and 0.8. Table 7 gives the results for a one-sided test at a 5% significance level for the two plausible value methods. The test length was equal to 5. Note that the power increased as a function of the autocorrelation, and the power of the test for the group effect actually decreased as a function of the autocorrelation. The explanation for the first phenomenon is that the variance of the difference of the latent student parameters decreases as the correlation goes up. The second phenomenon is due to the fact that within students, the variance of the ability estimates increases as the autocorrelation goes up, because of the increasing dependence between the responses. That is, the reliabilities of the estimates are highest when all responses are locally independent, and the reliability decreases when dependence due to the autocorrelation increases.

TABLE 7
Power for time and group effect for different sample sizes and estimation methods as a function of autocorrelation

		Plausible Value Unidimensional		Plausible Value Multidimensional	
N	Autocorr.	Time Effect	Group Effect	Time Effect	Group Effect
20	0.2	0.33	0.26	0.34	0.27
	0.4	0.35	0.27	0.38	0.26
	0.6	0.46	0.26	0.47	0.26
	0.8	0.67	0.22	0.68	0.23
50	0.2	0.57	0.49	0.58	0.52
	0.4	0.67	0.45	0.68	0.44
	0.6	0.76	0.46	0.76	0.46
	0.8	0.95	0.47	0.95	0.48
100	0.2	0.83	0.76	0.84	0.78
	0.4	0.89	0.74	0.88	0.73
	0.6	0.96	0.66	0.96	0.68
	0.8	1.00	0.63	1.00	0.64

Polytomously Scored Items, Two Groups and Two Time Points

The next set of simulation studies pertained to polytomously scored items, and it did generally have the same setup as the previous study. However, a different approach was chosen for the choice of the item parameters. The approach was analogous to the approach used in Glas and Dagohoy (in press). For the GPCM, the parameters a_i were drawn from a lognormal distribution with a mean equal to zero and a standard deviation of 0.25. Drawing the item parameters η_{ij} ($j = 1, \ldots, m_i$) was not considered, because the interrelation of these parameters may result in very unfavorable values with the consequence that some item categories may be without responses. Therefore the values of η_{ij} were fixed. The values chosen for Items 1–5 are given in Table 8. Note that the parameters of Item 3 are located in such a way that the category bounds are located symmetric with respect to the standard normal ability distribution. The first two items are shifted to the left on the latent scale; the last two items are shifted to the right. For simulation studies with 10 items, the item parameters were repeated for the second part of the test.

TABLE 8
Item parameter values for the generalized partial credit model (GPCM)

	Category			
Item	1	2	3	4
1	−2.0	−1.5	−0.5	0.0
2	−1.5	−1.0	0.0	0.5
3	−1.0	−0.5	0.5	1.0
4	−0.5	0.0	1.0	1.5
5	0.0	0.5	1.5	2.0

The item parameters for the SEQM and GRM were chosen in such a way that the item-category response functions were close to the response functions under the GPCM. To achieve this, data were generated under the GPCM, and using these data, the item parameters of the SEQM and GRM were estimated using MML. These estimated values were then used as generating values for the simulation of data following SEQM and GRM.

First, the Type I error rate for the two methods using plausible values (unidimensional and multidimensional) and the two MML methods (concurrent and MML2 [i.e., two-step MML]) were assessed for tests with a one-sided significance level of 5%. One of the questions addressed here is the robustness of the tests when the wrong model is used. So, for instance, the Type I error rate and power are assessed when the data are generated using the GPCM, after which the test statistic is computed using plausible values or MML estimates obtained under the SEQM or the GRM.

The results for the Type I error rate using concurrent MML estimates are displayed in Table 9. In all cases, the observed proportion of significant tests was close to the nominal significance level. The results for the other methods were similar.

TABLE 9
Type I error rate for group effect and time effect for different models, test lengths, sample sizes, and autocorrelations obtained using concurrent MML estimates of item and population parameters

	Computation Model:			GPCM		SEQM		GRM	
Generating Model	*K*	*N*	Autocorr.	Time Effect	Group Effect	Time Effect	Group Effect	Time Effect	Group Effect
GPCM	5	50	0.2	0.058	0.055	0.062	0.057	0.061	0.057
			0.4	0.060	0.063	0.052	0.068	0.053	0.064
		100	0.2	0.059	0.044	0.052	0.050	0.050	0.050
			0.4	0.055	0.040	0.057	0.040	0.057	0.053
	10	50	0.2	0.039	0.043	0.034	0.042	0.038	0.058
			0.4	0.049	0.051	0.049	0.042	0.045	0.039
		100	0.2	0.044	0.043	0.050	0.051	0.050	0.042
			0.4	0.037	0.052	0.043	0.053	0.033	0.048
SEQM	5	50	0.2	0.061	0.049	0.057	0.052	0.059	0.059
			0.4	0.061	0.063	0.056	0.066	0.051	0.068
		100	0.2	0.053	0.057	0.053	0.055	0.056	0.045
			0.4	0.055	0.061	0.053	0.056	0.045	0.062
	10	50	0.2	0.053	0.057	0.052	0.058	0.057	0.045
			0.4	0.051	0.071	0.051	0.071	0.050	0.060
		100	0.2	0.047	0.044	0.043	0.043	0.046	0.042
			0.4	0.039	0.049	0.041	0.051	0.045	0.043
GRM	5	50	0.2	0.060	0.061	0.057	0.064	0.056	0.066
			0.4	0.059	0.047	0.058	0.058	0.056	0.044
		100	0.2	0.055	0.059	0.052	0.056	0.058	0.066
			0.4	0.056	0.050	0.061	0.050	0.055	0.061
	10	50	0.2	0.053	0.058	0.049	0.057	0.054	0.050
			0.4	0.061	0.060	0.059	0.064	0.056	0.050
		100	0.2	0.048	0.043	0.048	0.041	0.050	0.041
			0.4	0.050	0.037	0.048	0.037	0.053	0.028

Next, the power was investigated using a number of simulations with the same setup as above: simulations with a group effect only, simulations with a time effect only, and simulation studies where both effects were present. The results for the two plausible value methods for the case with an autocorrelation of 0.4 are reported in Tables 10–12. The results for an autocorrelation of 0.2 are not reported here, but they had the same pattern as above; that is, the power for the time effect was slightly lower, and the power for the group effect was slightly higher. Note that the unidimensional plausible value method seems to have a slightly higher power than the multidimensional method. There is no clear explanation for this phenomenon. The magnitude of the power obtained using the concurrent MML method (not shown here) was very close to the power obtained using the multidimensional plausible value method, even if the wrong model was used.

TABLE 10
Type I error rate for time effect and power for group effect for different models, test lengths, sample sizes, and effect sizes obtained using plausible values

Generating Model	*K*	*N*	Effect Size	Plausible Value Unidimensional		Plausible Value Multidimensional	
				Time Effect	Group Effect	Time Effect	Group Effect
GPCM	5	50	0.2	0.055	0.497	0.062	0.455
			0.5	0.061	0.995	0.064	0.988
		100	0.2	0.046	0.792	0.058	0.711
			0.5	0.047	1.000	0.052	1.000
	10	50	0.2	0.049	0.499	0.052	0.455
			0.5	0.055	0.997	0.059	0.989
		100	0.2	0.035	0.794	0.046	0.715
			0.5	0.054	1.000	0.036	1.000
SEQM	5	50	0.2	0.074	0.579	0.067	0.457
			0.5	0.072	1.000	0.056	0.994
		100	0.2	0.056	0.808	0.060	0.694
			0.5	0.058	1.000	0.045	1.000
	10	50	0.2	0.049	0.585	0.035	0.466
			0.5	0.053	0.994	0.047	0.991
		100	0.2	0.047	0.808	0.046	0.694
			0.5	0.046	1.000	0.060	1.000
GRM	5	50	0.2	0.043	0.613	0.044	0.467
			0.5	0.049	1.000	0.038	0.988
		100	0.2	0.054	0.851	0.052	0.681
			0.5	0.061	1.000	0.066	1.000
	10	50	0.2	0.045	0.647	0.045	0.479
			0.5	0.051	0.997	0.059	0.984
		100	0.2	0.061	0.899	0.061	0.714
			0.5	0.060	1.000	0.054	1.000

TABLE 11
Type I error rate for group effect and power for time effect for different models, test lengths, sample sizes, and effect sizes obtained using plausible values

Generating Model	*K*	*N*	Effect Size	Plausible Value Unidimensional		Plausible Value Multidimensional	
				Time Effect	Group Effect	Time Effect	Group Effect
GPCM	5	50	0.2	0.769	0.050	0.637	0.052
			0.5	1.000	0.049	1.000	0.061
		100	0.2	0.954	0.047	0.878	0.046
			0.5	1.000	0.051	1.000	0.058
	10	50	0.2	0.786	0.051	0.634	0.052
			0.5	1.000	0.043	1.000	0.043
		100	0.2	0.956	0.040	0.876	0.030
			0.5	1.000	0.049	1.000	0.040
SEQM	5	50	0.2	0.865	0.052	0.638	0.053
			0.5	1.000	0.038	0.997	0.055
		100	0.2	0.979	0.046	0.880	0.048
			0.5	1.000	0.043	1.000	0.051
	10	50	0.2	0.875	0.055	0.664	0.055
			0.5	1.000	0.044	1.000	0.035
		100	0.2	0.957	0.045	0.885	0.039
			0.5	1.000	0.045	1.000	0.043
GRM	5	50	0.2	0.871	0.063	0.632	0.069
			0.5	1.000	0.053	1.000	0.043
		100	0.2	0.993	0.044	0.894	0.060
			0.5	1.000	0.046	1.000	0.052
	10	50	0.2	0.896	0.059	0.654	0.037
			0.5	1.000	0.052	1.000	0.039
		100	0.2	0.994	0.035	0.881	0.029
			0.5	1.000	0.043	1.000	0.047

TABLE 12
Power for time and group effect for different models, test lengths, sample sizes, and effect sizes obtained using plausible values

Generating Model	*K*	*N*	Effect Size	Plausible Value Unidimensional		Plausible Value Multidimensional	
				Time Effect	Group Effect	Time Effect	Group Effect
GPCM	5	50	0.2	0.769	0.503	0.649	0.450
			0.5	1.000	0.995	0.999	0.989
		100	0.2	0.957	0.770	0.893	0.710
			0.5	1.000	1.000	1.000	1.000
	10	50	0.2	0.788	0.503	0.682	0.496
			0.5	1.000	0.993	0.999	0.994
		100	0.2	0.981	0.756	0.905	0.721
			0.5	1.000	1.000	1.000	1.000
SEQM	5	50	0.2	0.862	0.563	0.640	0.458
			0.5	1.000	1.000	1.000	0.988
		100	0.2	0.993	0.819	0.893	0.707
			0.5	1.000	1.000	1.000	1.000
	10	50	0.2	0.877	0.583	0.663	0.467
			0.5	1.000	0.997	0.999	0.990
		100	0.2	0.987	0.883	0.901	0.722
			0.5	1.000	1.000	1.000	1.000
GRM	5	50	0.2	0.886	0.607	0.691	0.442
			0.5	1.000	0.999	0.999	0.984
		100	0.2	0.992	0.857	0.878	0.694
			0.5	1.000	1.000	1.000	1.000
	10	50	0.2	0.881	0.668	0.697	0.479
			0.5	1.000	0.998	1.000	0.989
		100	0.2	0.976	0.888	0.896	0.723
			0.5	1.000	1.000	1.000	1.000

However, this did not hold for the MML2 method (Tables 13–15). The columns labeled “True Model” give the power obtained when using the generating model in which the related generating values of the item parameters were used as fixed constants. The other columns give the power obtained when the item parameters were re-estimated, under both the correct model and the two wrong models. Note that in these three last cases, the power goes down further.

TABLE 13
Type I error rate for time effect and power for group effect for different models, test lengths, sample sizes, and effect sizes obtained using MML estimates of population parameters

Generating Model	Computation Model			True Model		GPCM		SEQM		GRM	
	K	*N*	Effect Size	Time Effect	Group Effect	Time Effect	Group Effect	Time Effect	Group Effect	Time Effect	Group Effect
GPCM	5	50	0.2	0.064	0.069	0.058	0.058	0.052	0.052	0.058	0.052
			0.5	0.057	0.303	0.063	0.269	0.057	0.274	0.063	0.286
		100	0.2	0.054	0.166	0.054	0.152	0.054	0.126	0.054	0.148
			0.5	0.078	0.654	0.097	0.682	0.097	0.673	0.111	0.673
	10	50	0.2	0.043	0.086	0.037	0.093	0.049	0.105	0.049	0.086
			0.5	0.065	0.319	0.087	0.275	0.072	0.284	0.094	0.288
		100	0.2	0.055	0.142	0.059	0.188	0.068	0.128	0.078	0.137
			0.5	0.037	0.689	0.050	0.721	0.050	0.685	0.059	0.721
SEQM	5	50	0.2	0.043	0.052	0.052	0.034	0.052	0.026	0.052	0.026
			0.5	0.034	0.293	0.052	0.190	0.060	0.155	0.043	0.181
		100	0.2	0.023	0.169	0.147	0.056	0.141	0.062	0.141	0.051
			0.5	0.073	0.606	0.121	0.327	0.115	0.255	0.115	0.327
	10	50	0.2	0.019	0.151	0.057	0.057	0.075	0.075	0.038	0.057
			0.5	0.094	0.328	0.109	0.188	0.062	0.188	0.078	0.094
		100	0.2	0.095	0.202	0.137	0.048	0.155	0.048	0.155	0.048
			0.5	0.054	0.631	0.113	0.344	0.155	0.284	0.125	0.273
GRM	5	50	0.2	0.029	0.095	0.058	0.051	0.036	0.058	0.036	0.051
			0.5	0.040	0.240	0.040	0.107	0.027	0.160	0.040	0.187
		100	0.2	0.017	0.106	0.095	0.039	0.106	0.056	0.101	0.045
			0.5	0.032	0.468	0.077	0.186	0.090	0.244	0.064	0.378
	10	50	0.2	0.109	0.065	0.043	0.065	0.000	0.109	0.000	0.065
			0.5	0.056	0.296	0.074	0.037	0.056	0.074	0.056	0.074
		100	0.2	0.048	0.126	0.186	0.060	0.204	0.078	0.204	0.072
			0.5	0.019	0.604	0.130	0.240	0.130	0.262	0.117	0.162

TABLE 14
Type I error rate for group effect and power for time effect for different models, test lengths, sample sizes, and effect sizes obtained using MML estimates of population parameters

Generating Model	Computation Model			True Model		GPCM		SEQM		GRM	
	K	*N*	Effect Size	Time Effect	Group Effect	Time Effect	Group Effect	Time Effect	Group Effect	Time Effect	Group Effect
GPCM	5	50	0.2	0.163	0.023	0.186	0.017	0.174	0.012	0.169	0.017
			0.5	0.543	0.051	0.526	0.040	0.503	0.034	0.509	0.029
		100	0.2	0.253	0.054	0.290	0.041	0.258	0.036	0.267	0.041
			0.5	0.824	0.050	0.811	0.027	0.806	0.027	0.824	0.032
	10	50	0.2	0.151	0.033	0.151	0.020	0.164	0.020	0.151	0.013
			0.5	0.633	0.030	0.615	0.024	0.609	0.018	0.633	0.024
		100	0.2	0.282	0.045	0.288	0.014	0.266	0.009	0.260	0.014
			0.5	0.847	0.037	0.838	0.028	0.819	0.037	0.838	0.042
SEQM	5	50	0.2	0.138	0.046	0.119	0.064	0.083	0.064	0.064	0.046
			0.5	0.494	0.052	0.455	0.052	0.494	0.078	0.545	0.065
		100	0.2	0.162	0.060	0.090	0.132	0.066	0.186	0.096	0.150
			0.5	0.782	0.083	0.590	0.051	0.692	0.090	0.744	0.045
	10	50	0.2	0.159	0.079	0.143	0.048	0.143	0.048	0.095	0.063
			0.5	0.485	0.023	0.42	0.045	0.477	0.068	0.555	0.091
		100	0.2	0.230	0.026	0.145	0.145	0.118	0.224	0.125	0.197
			0.5	0.854	0.024	0.691	0.057	0.693	0.089	0.685	0.089
GRM	5	50	0.2	0.066	0.008	0.049	0.066	0.033	0.074	0.033	0.057
			0.5	0.467	0.017	0.400	0.067	0.517	0.033	0.633	0.017
		100	0.2	0.172	0.040	0.075	0.109	0.052	0.184	0.040	0.121
			0.5	0.809	0.034	0.897	0.079	0.833	0.101	0.844	0.034
	10	50	0.2	0.164	0.000	0.115	0.180	0.082	0.131	0.098	0.164
			0.5	0.606	0.061	0.455	0.121	0.606	0.121	0.576	0.061
		100	0.2	0.191	0.038	0.045	0.191	0.038	0.229	0.045	0.217
			0.5	0.810	0.017	0.892	0.103	0.821	0.103	0.803	0.103

TABLE 15
Power for time and group effect for different models, test lengths, sample sizes, and effect sizes obtained using MML estimates of population parameters

	Computation Model			True Model		GPCM		SEQM		GRM	
Generating Model	*K*	*N*	Effect Size	Time Effect	Group Effect	Time Effect	Group Effect	Time Effect	Group Effect	Time Effect	Group Effect
GPCM	5	50	0.2	0.110	0.069	0.098	0.064	0.104	0.087	0.098	0.064
			0.5	0.579	0.360	0.573	0.354	0.524	0.335	0.543	0.354
		100	0.2	0.248	0.158	0.270	0.167	0.270	0.167	0.252	0.167
			0.5	0.805	0.686	0.818	0.705	0.814	0.668	0.809	0.691
	10	50	0.2	0.141	0.092	0.169	0.106	0.176	0.099	0.169	0.099
			0.5	0.541	0.389	0.580	0.414	0.573	0.369	0.567	0.427
		100	0.2	0.281	0.167	0.299	0.136	0.303	0.154	0.290	0.154
			0.5	0.908	0.693	0.876	0.702	0.876	0.665	0.881	0.688
SEQM	5	50	0.2	0.064	0.074	0.032	0.011	0.043	0.021	0.032	0.032
			0.5	0.333	0.394	0.273	0.333	0.515	0.455	0.455	0.424
		100	0.2	0.168	0.173	0.087	0.064	0.046	0.052	0.069	0.046
			0.5	0.727	0.662	0.636	0.442	0.870	0.818	0.857	0.766
	10	50	0.2	0.145	0.127	0.109	0.055	0.109	0.073	0.109	0.073
			0.5	0.788	0.477	0.801	0.400	0.901	0.422	0.889	0.412
		100	0.2	0.287	0.188	0.147	0.047	0.133	0.047	0.113	0.040
			0.5	0.933	0.600	0.899	0.488	0.933	0.533	0.933	0.647
GRM	5	50	0.2	0.065	0.093	0.047	0.019	0.028	0.028	0.028	0.019
			0.5	0.467	0.200	0.667	0.333	0.800	0.600	0.933	0.667
		100	0.2	0.188	0.175	0.081	0.037	0.056	0.031	0.062	0.031
			0.5	0.800	0.450	0.800	0.550	1.000	0.950	1.000	0.950
	10	50	0.2	0.204	0.037	0.130	0.074	0.093	0.074	0.074	0.056
			0.5	0.750	0.500	0.588	0.350	0.650	0.650	0.900	0.850
		100	0.2	0.212	0.176	0.099	0.020	0.086	0.046	0.079	0.040
			0.5	1.000	0.477	1.000	0.677	1.000	0.888	1.000	0.688

Dichotomously Scored Items With Response Times, Two Groups and Two Time Points

The last set of simulation studies referred to the combined model for accuracy and speed. In this setup, the parameters of the model were fixed to the values obtained in the real data example reported by van der Linden and Glas (2006). It was assumed that two groups of respondents were administered either the first 10 or 40 items of the item bank used in the example by van der Linden and Glas on two occasions. The autocorrelation was either 0.2 or 0.4. Both the mean of the accuracy parameters (the ability parameters in the 3PL model) and the mean of the speed parameters (the parameters of the model for speed) could differ across groups and time points.

The results in Table 16 pertain to a simulation where a group effect on the accuracy parameters was induced. Again, the magnitude of this group effect was either 0.2 or 0.5. There was no main effect for speed. The powers of the tests for the speed effect had the usual main effects of test length and sample size. The choice of the estimation methods, which were the two plausible value methods and the concurrent MML method, had little impact. The Type I error rate for the test targeted at group differences in speed was close to the nominal value. The same held for the (not displayed) results for the power of the tests targeted at time effects.

The results in Table 17 pertain to a simulation with a group effect on both the accuracy and speed parameters. As in the other tables, the magnitudes of these group effects were either 0.2 or 0.5. Again, the powers of the tests for the speed effect had the usual main effects of test length and sample size and, again, the choice of the estimation methods made little difference. The (not displayed) Type I error rate for time effects was close to the nominal value.

Finally, Table 18 gives the results of a power study for time effects both in accuracy and speed. The results are analogous to the results above. Note that the effects of the autocorrelation on a test targeted at a time effect reported above are also clearly visible in this last study.

TABLE 16
Type I error rate for group effect for speed and power for group effect for accuracy for different test lengths, sample sizes, effect sizes, autocorrelations, and estimation methods

				Plausible Value Unidimensional		Plausible Value Multidimensional		MML Estimate	
K	*N*	Effect Size	Autocorr.	Speed Effect	Accuracy Effect	Speed Effect	Accuracy Effect	Speed Effect	Accuracy Effect
10	100	0.2	0.2	0.06	0.66	0.05	0.63	0.06	0.66
			0.4	0.07	0.67	0.05	0.64	0.06	0.67
		0.5	0.2	0.06	1.00	0.05	1.00	0.04	1.00
			0.4	0.07	1.00	0.06	1.00	0.05	1.00
	200	0.2	0.2	0.05	0.98	0.06	0.92	0.05	0.87
			0.4	0.06	0.98	0.05	0.92	0.04	0.78
		0.5	0.2	0.05	1.00	0.05	1.00	0.04	1.00
			0.4	0.06	1.00	0.04	1.00	0.05	1.00
40	100	0.2	0.2	0.05	0.67	0.05	0.65	0.05	0.66
			0.4	0.06	0.67	0.06	0.66	0.04	0.66
		0.5	0.2	0.06	1.00	0.05	1.00	0.06	1.00
			0.4	0.06	1.00	0.06	1.00	0.06	1.00
	200	0.2	0.2	0.05	0.99	0.05	0.91	0.04	0.87
			0.4	0.04	0.99	0.05	0.91	0.05	0.89
		0.5	0.2	0.05	1.00	0.06	1.00	0.06	1.00
			0.4	0.04	1.00	0.04	1.00	0.06	1.00

TABLE 17
Power for group effects for speed and accuracy for different test lengths, sample sizes, effect sizes, autocorrelations, and estimation methods

				Plausible Value Unidimensional		Plausible Value Multidimensional		MML Estimate	
K	*N*	Effect Size	Autocorr.	Speed Effect	Accuracy Effect	Speed Effect	Accuracy Effect	Speed Effect	Accuracy Effect
10	100	0.2	0.2	0.73	0.67	0.75	0.63	0.56	0.49
			0.4	0.70	0.67	0.83	0.63	0.68	0.43
		0.5	0.2	1.00	1.00	1.00	1.00	1.00	1.00
			0.4	1.00	1.00	1.00	1.00	1.00	1.00
	200	0.2	0.2	0.90	0.95	0.94	0.93	0.88	0.81
			0.4	0.92	0.96	0.97	0.88	0.95	0.79
		0.5	0.2	1.00	1.00	1.00	1.00	1.00	1.00
			0.4	1.00	1.00	1.00	1.00	1.00	1.00
40	100	0.2	0.2	0.72	0.67	0.77	0.68	0.70	0.68
			0.4	0.77	0.68	0.87	0.68	0.84	0.66
		0.5	0.2	1.00	1.00	1.00	1.00	1.00	1.00
			0.4	1.00	1.00	1.00	1.00	1.00	1.00
	200	0.2	0.2	0.96	0.94	0.97	0.91	0.95	0.87
			0.4	0.98	0.94	0.97	0.99	0.95	0.87
		0.5	0.2	1.00	1.00	1.00	1.00	1.00	1.00
			0.4	1.00	1.00	1.00	1.00	1.00	1.00

TABLE 18
Power for time effects for speed and accuracy for different test lengths, sample sizes, effect sizes, autocorrelations, and estimation methods

				Plausible Value Unidimensional		Plausible Value Multidimensional		MML Estimate	
K	*N*	Effect Size	Autocorr.	Speed Effect	Accuracy Effect	Speed Effect	Accuracy Effect	Speed Effect	Accuracy Effect
10	100	0.2	0.2	0.74	0.66	0.72	0.67	0.60	0.51
			0.4	0.79	0.67	0.77	0.69	0.68	0.55
		0.5	0.2	1.00	1.00	1.00	1.00	1.00	1.00
			0.4	1.00	1.00	1.00	1.00	1.00	1.00
	200	0.2	0.2	0.90	0.94	0.96	0.91	0.90	0.91
			0.4	0.98	0.99	0.98	0.95	0.97	0.99
		0.5	0.2	1.00	1.00	1.00	1.00	1.00	1.00
			0.4	1.00	1.00	1.00	1.00	1.00	1.00
40	100	0.2	0.2	0.73	0.66	0.76	0.66	0.76	0.68
			0.4	0.78	0.69	0.86	0.71	0.84	0.72
		0.5	0.2	1.00	1.00	1.00	1.00	1.00	1.00
			0.4	1.00	1.00	1.00	1.00	1.00	1.00
	200	0.2	0.2	0.96	0.92	0.95	0.90	0.95	0.90
			0.4	0.98	0.96	0.99	0.99	0.95	0.98
		0.5	0.2	1.00	1.00	1.00	1.00	1.00	1.00
			0.4	1.00	1.00	1.00	1.00	1.00	1.00

Conclusions

The essential problem with the estimation of linear models on ability parameters is that the estimates of these parameters are not observations, but they are estimates with sampling variance. In CAT and in educational surveys such as PISA, TIMSS, and NAEP, students do not respond to the same items, or sometimes not even to the same number of items. This leads to differences in the uncertainty in their parameter estimates, and this uncertainty and these differences in uncertainty have to be taken into account. The theoretically ideal way to do this is to obtain a concurrent estimate of all the parameters in the model. Unfortunately, this is often quite complicated. Several alternatives have been developed. In the studies presented here, we assessed the performance with respect to the power to detect main effects in an analysis of variance model with a repeated measure of one alternative, plausible value imputation. It turns out that a relatively simple method, a unidimensional plausible value method where the plausible values are drawn from separate unidimensional posteriors, has a Type I error rate and power that is at least comparable to the Type I error rate and power obtained using concurrent MML estimation. So the more complicated approach of drawing from a multidimensional posterior seems to be unnecessary. It must be noted that Rubin and Thomas (2001) proposed an alternative method for introducing multidimensional ability estimates into regression equations. Their method is a two-step approach where the measurement model is estimated first, and in their method the dependence between a student's abilities must be taken into account in the first step. The present simulations seem to indicate that this approach is unnecessary.

Finally, we found that a two-step MML approach where the second step consists of MML estimation of the parameters of the linear model on the abilities, treating the item parameters as constants, results in a substantial loss of power. Therefore this method is not advisable.

References

Ackerman, T. A. (1996a). Developments in multidimensional item response theory. *Applied Psychological Measurement 20,* 309–310.

Ackerman, T. A. (1996b). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 20,* 311–329.

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.

Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement, 12,* 261–280.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models.* Newbury Park: Sage.

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

de Boeck, P., & Wilson, M. (2004). *Explanatory item response models*. New York: Springer.

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

Fox, J. P. (2003). Stochastic EM for estimating the parameters of a multilevel IRT model. *British Journal of Mathematical and Statistical Psychology, 56,* 65–81.

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: Laurence Erlbaum.

Fox, J.-P. & Glas, C. A. W. (2003), Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika, 68*, 169–191.

Glas, C. A. W., & Dagohoy, A. V. T. (in press). A student fit test for IRT models for polytomous items. *Psychometrika.*

Glas, C. A. W., & van der Linden, W. J. (2006). *Likelihood-based estimation methods for models for concurrent continuous and discrete responses with a structure for the item and person parameters* (LSAC Research Report 06-06). Newtown: PA, Law School Admission Council, Inc.

Goldstein, H. (1986). Multilevel mixed linear models analysis using iterative generalized least squares. *Biometrika, 73,* 43–56.

Holman, R., Glas, C. A. W., & de Haan, R. J. (2003). Power analysis in randomized clinical trials based on item response theory. *Controlled Clinical Trials, 24,* 390–410.

Longford, N. T. (1993). Random coefficient models*. Oxford Statistical Science Series*, Vol. 11. Oxford: Clarendon.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

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

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257–269). New York: Springer.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19,* 91–100.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56,* 177–196.

Mislevy, R. J., & Bock, R. D. (1989). A hierarchical item-response model for educational testing. In R. D. Bock (Ed.), *Multilevel analysis of educational data.* San Diego: Academic Press.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17,* 131–154.

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

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401–412.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer.

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, *17*, 1–100.

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*, 533–555.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling*. London: Chapman & Hall/CRC.

Tanner, M. A. (1993). *Tools for statistical inference*. New York: Springer.

Thomas, N., & Gan, N. (1997). Generating multiple imputations for matrix sampling data with item response models. *Journal of Educational and Behavioral Statistics*, *22*, 425–445.

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

Tutz, G. (1997). Sequential models for ordered responses. In W. J.van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 139–152). New York: Springer.

van der Linden, W. J. (2005). *A hierarchical framework for modeling speed and accuracy on test items*. (LSAC Research Report 05-02). Newtown: PA, Law School Admission Council, Inc.

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., & Glas, C. A. W. (2006). *Statistical tests for conditional independence in a hierarchical model for speed and accuracy on test items* (LSAC Research Report 06-02). Newtown: PA, Law School Admission Council, Inc.

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

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 and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 123–138). 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, IL: Scientific Software International, Inc.