Educational and
Psychological Measurement
Volume 66 Number 1
February 2006 5-34
© 2006 Sage Publications
10.1177/0013164405282490
http://epm.sagepub.com
hosted at
http://online.sagepub.com

Application of Multidimensional Item Response Theory Models to Longitudinal Data

Janneke M. te Marvelde
Cees A. W. Glas
University of Twente, the Netherlands
Georges Van Landeghem
Jan Van Damme
Catholic University of Leuven, Belgium

The application of multidimensional item response theory (IRT) models to longitudinal educational surveys where students are repeatedly measured is discussed and exemplified. A marginal maximum likelihood (MML) method to estimate the parameters of a multidimensional generalized partial credit model for repeated measures is presented. It is shown that model fit can be evaluated using Lagrange multiplier tests. Two tests are presented: the first aims at evaluation of the fit of the item response functions and the second at the constancy of the item location parameters over time points. The outcome of the latter test is compared with an analysis using scatter plots and linear regression. An analysis of data from a school effectiveness study in Flanders (Belgium) is presented as an example of the application of these methods. In the example, it is evaluated whether the concepts "academic self-concept," "well-being at school," and "attentiveness in the classroom" were constant during the secondary school period.

Keywords: *longitudinal data; repeated measures; panel data; item response theory; generalized partial credit model; multidimensional IRT models; marginal maximum likelihood estimation*

In educational and psychological research, changes over time are often investigated by performing longitudinal analyses for observations collected at several time points. For example, in educational research a goal can be to determine the development of achievement in mathematics of pupils over time. To investigate such a development, pupils may be presented one or more mathematics tests at several time points. The item responses on these tests can be related to a latent variable math achievement via an item response theory (IRT) model (see, for instance, Lord, 1980). In longitudi-

Authors' Note: Address requests for information to Cees A. W. Glas, Department of Research Methodology, Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands; e-mail: C.A.W.Glas@gw.utwente.nl.

nal designs, it is usually assumed that the positions of the students on the latent scale change over time. However, repeated measures have the complication that the responses on different time points are not independent. Therefore, for the analysis of repeated measures Andersen, (1985) suggested an extended Rasch model for dichotomously scored items, where the item responses at each time point are modeled with a unidimensional IRT model, and the latent variables of each time point are correlated. Adams, Wilson, and Wang (1997; also see Wang & Wu, 2004) have discussed a more general multidimensional IRT model, which can also be applied to polytomous items. This model can also be used to analyze test data with items measuring more than one latent variable and test data consisting of several subscales each measuring one specific latent variable. With some minor adjustments, the latter case can be used to analyze repeated measures. For analyzing repeated measures, the multidimensional IRT model provides direct estimates of the relations between the latent variables at several time points, and the accuracy of the parameters estimates is enhanced by the often strong relationship between the latent variables.

In the present study, repeated measures are modeled using the Generalized Partial Credit Model (GPCM; Muraki, 1992), which is a model for polytomous items with ordered response categories. A marginal maximum likelihood (MML) estimation procedure, adapted to take into account the dependencies of the responses of different time points, is used to estimate the item and population parameters. It is shown that model fit can be evaluated using Lagrange multiplier (LM) tests. The first LM test presented evaluates the fit of the item response functions within time points, whereas a second LM test evaluates the constancy of the item location parameters over time points. The results of the latter test are compared with an analysis using linear regression and scatter plots.

In this article, change over time is modeled by assuming that the latent ability parameters have a multivariate normal distribution. This assumption pertains to the dependence between these parameters over time points. The change is estimated along with the item parameters from a likelihood function that is marginalizing over the ability parameters. There are several alternatives to this approach. For instance, Embretson (1991) proposed a two-step procedure based on the Rasch model where the item parameters are first estimated by conditional maximum likelihood and then the person parameters are estimated with the estimates of the item parameters imputed as fixed constants. The reasons for not pursuing this approach here is that it is difficult to compute standard errors for the change parameters in a two-step procedure (see Verhelst & Glas, 1995, pp. 185-186). Furthermore, in this approach the Rasch model must fit the data, and in educational settings this is seldom the case. The latter argument also holds for an approach by Fischer and Parzer (1991; also see Fischer & Ponocny, 1994), where the change is modeled by imposing linear restrictions on item parameters, and an approach by Fischer (2003), where conditional maximum likelihood estimates of item parameters and individual change parameters are computed simultaneously.

The estimation and testing procedure is exemplified using data of a Flemish school effectiveness study. In this survey, the achievement on mathematics and Dutch language ability were measured along with "Academic self-concept," "well-being at school," and "attentiveness in the classroom" (Van Damme, De Fraine, Van Landeghem, Opdenakker, & Onghena, 2002). The examples presented here pertain to the latter three scales. Using a varimax rotated factor analysis; the questionnaire measuring the noncognitive concepts was divided into several scales, each measuring a different concept. It is important to stress that the example serves a didactical purpose and has no pretension of being a contribution to the substantive theory of the topic. The main goal in the example is to show how to evaluate latent variables over time. In the next section, the multidimensional GPCM is presented. Then the MML estimation procedure is outlined and methods for testing model fit are discussed. Next, examples are given of the estimation and testing procedure. Finally, some conclusions are drawn.

Multidimensional IRT Models

Usually, in IRT models it is assumed that there is one (dominant) latent variable θ that explains test performance. However, it may be a priori clear that multiple latent variables are involved or the dimensionality of the latent variable structure might not even be clear at all. In these cases, multidimensional IRT models can serve confirmatory and explorative purposes. An example is the mathematical "story problem" (Reckase, 1985), where both mathematical and verbal skills are required to obtain a correct answer. A test with items related to more than one latent variable is often labeled a within-item-multidimensional test (Adams et al., 1997). Adams et al. (1997) examined another class of multidimensional IRT models, between-item-multidimensional models, in which one test can be divided into subtests or scales where the responses to the items of each scale can be described by a unidimensional IRT model. The latent variables measured by the different scales are assumed to correlate. Andersen (1985) applied a between-item-multidimensional IRT model to analyze longitudinal data. In this case, the division into subtests is easy, as the same test is used at different points in time.

These models are typified as follows. Let a test have k items (i =1, . . . , k) with item i having m_i + 1 response categories indexed as j = 0, . . . , m_i. The response of a person, indexed n, is presented by an m_i-dimensional stochastic vector $\mathbf{x}_{ni}$ with elements

$$X_{nij} = \begin{cases} 1 & \text{if person } n \text{ gives a response in category } j \text{ of item } i, \\ 0 & \text{otherwise,} \end{cases}$$

with $n = 1, \ldots, N$ and $j = 1, \ldots, m_i$. Note that if the person scores in category $j = 0$, the response to item i is given by the response vector $\mathbf{x}_{ni} = 0$.

A generalization of the GPCM to a Q-dimensional model (Andersen, 1985; Reckase, 1985, 1997) is given by

$$P(X_{nij} = 1 | \boldsymbol{\theta}_n, \mathbf{a}_i, \mathbf{b}_i) = \psi_{ij}(\boldsymbol{\theta}_n) = \frac{\exp\left(\sum_{q=1}^{Q} a_{iq} j\theta_{nq} - b_{ij}\right)}{1 + \sum_{h=1}^{m_i} \exp\left(\sum_{q=1}^{Q} a_{iq} h\theta_{nq} - b_{ih}\right)} \tag{1}$$

for $j = 1, \ldots, m_i$, where $\mathbf{a}_{\mathrm{i}} = (a_{i1}, \ldots, a_{iq}, \ldots, a_{iQ})$ is a Q-dimensional vector of discrimination parameters; $\mathbf{b}_i = (b_{i1}, \ldots, b_{iq}, \ldots, b_{im})$ is an m_i-dimensional vector of location parameters, and $\boldsymbol{\theta}_n = (\theta_{n1}, \ldots, \theta_{nq}, \ldots, \theta_{nQ})$ is a Q-dimensional vector of latent variable parameters. For $j = 0$, the nominator of Formula 1 is equal to one and the denominator remains the same. Note that if Q equals one, a unidimensional GPCM is obtained. The model described by Formula 1 is a within-item-multidimensional model, where Q latent variables are involved in the item response.

The within-item-multidimensional model can be adapted to a between-item-multidimensional model by imposing restrictions on the discrimination parameters. That is, the discrimination parameter corresponding to the latent variable that is relevant for the item response is a free parameter, and the other discrimination parameters are set equal to zero. Besides identifying the latent dimensions, also the location of the latent scale must be fixed. This can be done by setting the mean of the ability distribution equal to zero. Béguin and Glas (2001) and Holman and Glas (2005) showed that the ensemble of these restrictions suffices to identity the model.

Because the model is equivalent with a full-information factor analysis model (Takane & de Leeuw, 1987), the discrimination parameters are often called factor loadings. In case of repeated measures, Andersen (1985) used a between-items multidimensional model where each repeated administration of the test was associated with a separate latent variable. So it is assumed that the characteristics of the test do not change over time, which means that the item parameters should be constant over time. This can be realized using a between-item-multidimensional model with linear restrictions on the item parameters. Furthermore, the covariance matrix of the ability dimensions accounts for the dependence between the item responses over time points.

Marginal Maximum Likelihood Estimation

MML is probably the most used technique for estimation of the parameters of the GPCM. The theory was developed by Bock and Aitkin (1981), Thissen (1982), Rigdon and Tsutakawa (1983), and Mislevy (1984, 1986), among others. For dichotomously scored items, unidimensional models can be estimated by the software package Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 2002); for polytomous items, estimates can be computed using the software packages Multilog (Thissen, Chen, & Bock, 2002) or Parscale (Muraki & Bock, 2002). MML estimation procedures are also available for multidimensional IRT models (Bock, Gibbons, & Muraki, 1988) and implemented in TESTFACT (Wood et al., 2002) and ConQuest (Wu, Adams, & Wilson, 1997). All these packages compute concurrent MML estimates of all the structural parameters in the model, and this is the approach that is also pursued in the present article. However, it must be mentioned that MML estimates can also be obtained in two-step procedures such as implemented in Mplus (Muthén & Muthén, 2003).

One of the major advantages of IRT is the possibility to use incomplete designs to collect data (Kolen & Brennan, 1995; Lord, 1980; Trivellato, 1999). This gives the researcher the opportunity to cover a broad domain with a large number of items without burdening the students with the need to respond to very long tests. Furthermore, IRT also provides a straightforward way of handling missing item responses in the case that the missing item responses are missing at random (MAR), which means that the missing values are random within observed covariate classes (Rubin, 1976). An obvious example is accidentally skipped pages of items. However, also the unobserved item responses in data obtained via computerized adaptive testing and flexilevel testing, where selected items and tests depend on previous responses, are MAR (see, for instance, Lord, 1980). To adapt to incomplete designs and missing data, let a variable d_{ni} be defined as

$$d_{ni} = \begin{cases} 1 & \text{if a response is available for person } n \text{ on item } i, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

If $d_{ni} = 0$, X_{nij} and $P(X_{nij})$ assume arbitrary values that have no consequences for the computations. If $d_{ni} = 1$, the probability of person n responding in category j of item i is given by Formula 1. Let x_n be the response pattern of person n. Using the assumption of local independence, the probability of a response pattern $\mathbf{x}_n$, $\mathbf{x}_n = (x_{n1}, \ldots, x_{ni}, \ldots, x_{nk})$, is

$$P(\mathbf{x}_n \mid \boldsymbol{\theta}_n, \mathbf{a}, \mathbf{b}) = \prod_{i-1}^{k} \prod_{j=0}^{m_i} P(X_{nij} = x_{nij} \mid \boldsymbol{\theta}_n, \mathbf{a}_i, b_{ij})^{x_{nij} d_{ni}} \tag{3}$$

Note that $\mathbf{x}_n$ determines the values of d_{ni}, so these values do not appear on the left side of Equation 3.

Maximizing a likelihood function defined as the product over the probabilities of the individual response patterns given by Equation 3 with respect to θ, **a** and **b** will generally not produce consistent estimators because the number of person parameters goes to infinity if the sample size goes to infinity. Therefore, in MML estimation, the parameters of an IRT model marginalized with respect to θ are estimated.

It is assumed that the person parameters are sampled from a multivariate normal distribution with a Q-dimensional vector of means $\boldsymbol{\mu}$ and a $Q \times Q$ covariance matrix Σ. The density will be denoted by $g(\boldsymbol{\theta} | \boldsymbol{\mu}, \Sigma)$. In the sequel, the covariance matrix can be the covariance between scales at a certain time point, the covariance over time points for a specific scale, and the entire covariance matrix over scales and time points.

Bock and Aitkin (1981) mentioned that $\boldsymbol{\theta}$ can be seen as stochastic variables that could, in principle, be observed. However, $\boldsymbol{\theta}$ is not observed and therefore the marginal likelihood function $L(\mathbf{a}, \mathbf{b}, \boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{X})$ is defined by

$$L(\mathbf{a}, \mathbf{b}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} P(\mathbf{x}_n \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{4}$$

where the response patterns of n persons are collected in an $N \times k$ matrix $\boldsymbol{X}$ and the probability of the observed response pattern for person n is given by

$$P(\mathbf{x}_n \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int \ldots \int P(\mathbf{x}_n \mid \boldsymbol{\theta}, \mathbf{a}, \mathbf{b}) g(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\boldsymbol{\theta}. \tag{5}$$

Glas (1992) showed that the estimation equations for maximization of Equation 4 can be derived using Fisher's identity. The application of this identity will be illustrated by deriving the estimation equation for the mean of the qth ability dimension, μ_q. If the person parameters $\boldsymbol{\theta}_n$ were observed, the maximum likelihood estimate would be

$$\mu_q = \frac{1}{N} \sum_{n=1}^{N} \theta_{nq}. \tag{6}$$

Because the person parameters are not observed, we can take the posterior expectation of Equation 6 to obtain

$$\mu_q = \frac{1}{N} \sum_{n=1}^{N} E(\theta_{nq} \mid \mathbf{x}_n, \mathbf{a}, \mathbf{b}, \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{7}$$

where the expectation is with respect to the posterior distribution of $\boldsymbol{\theta}_n$, which is a function of the response pattern $\mathbf{x}_n$, the item parameters $\mathbf{a}$ and $\mathbf{b}$, and the population parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Analogously, the estimation equations of the elements of $\boldsymbol{\Sigma}$ are given by

$$\sigma_q^2 = \frac{1}{N} \left[\sum_{n=1}^{N} E(\theta_{nq}^2 \mid \mathbf{x}_n, \mathbf{a}, \mathbf{b}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) - N\mu_q^2 \right] \tag{8}$$

and

$$\sigma_{qq'} = \frac{1}{N} \left[\sum_{n=1}^{N} E(\theta_{nq}\theta_{nq'} \mid \mathbf{x}_n, \mathbf{a}, \mathbf{b}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) - N\mu_q\mu_{q'} \right]. \tag{9}$$

So the estimate of the mean of the ability parameters is given by the sum of the posterior expected values of the person parameters divided by the number of persons and the variance and covariance are computed as the analogous posterior expected variance and covariance. Note that the covariance is a measure of the joint variance of the person parameters of two latent variables and reflects the relation between these two latent variables. In the context of longitudinal data, a covariance expresses the relationship between the same latent variable measured at two different time points and denotes the stability of the measured latent variable over time.

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

Testing the Model

Inferences from IRT analyses are only valid if the model holds. In principle, the validity of the model can be tested by estimating the entire model over all scales and time points simultaneously and computing fit statistics for this encompassing model. However, evaluation of model fit in such a large encompassing model with many parameters has the drawback that the sources of misfit may be hard to identify. Therefore, we propose a bottom-up procedure where model fit is first evaluated for separate scales and time points and then, if the separate models prove acceptable, proceed with fitting larger models that include covariance matrices between scales, covariance matrices between time points, and finally, the entire covariance matrix between all latent dimensions.

The approach to testing fit to the GPCM presented here is an adaptation of a method proposed by Glas (1998, 1999). The tests that will be used are based on the LM statistic. The LM statistic is used to test a special model against a more general alternative. The special model is derived from the general model by fixing one or more parameters to known constants. Contrary to the likelihood ratio statistic and the Wald statistic, the LM statistic can be evaluated using the ML estimates (for IRT, the MML estimates) of the special model only; the parameters of the more general alternative need not be estimated. The statistic is asymptotically chi-square distributed with degrees of freedom equal to the number of fixed parameters. Glas and Suarez-Falcon (2003) showed that the Type I error rate and the power of the LM tests are good. For a general description of the LM test, the reader is referred to Buse (1982). In the present study, the model fit within time points is evaluated using a LM test to evaluate fit of item response functions. This LM test is based on MML estimates of the unidimensional IRT model. A second LM test is performed to evaluate stability of the location parameters over time. This LM test is based on MML estimates of the multidimensional IRT model. The analysis using linear regression and scatter plots will be explained as well.

Model Fit Within Time Points

Glas (1999) proposed an LM test to evaluate the appropriateness of the item response functions ψ_{ij}. This could be done by estimating θ for each person, dividing the students into subgroups based on the estimates of θ, and evaluating for each subgroup the difference between the observed item response proportions and the estimated response probabilities. However, Glas and Suarez-Falcon (2003) showed that test statistics based on partitioning the sample of persons using estimates rather than observable statistics have very poor properties (see also Orlando & Thissen, 2000). Therefore, a test is considered which is based on a partitioning of the persons using their partial sum score. Usually, the sum score correlates highly with the estimate of θ. Therefore, it is reasonable to evaluate whether the item response functions ψ_{ij} match the analogous observed proportions in subgroups formed using these sum scores.

Let the item of interest be labeled i, while the other items are labeled $g = 1, \ldots, i-1, i+1, \ldots, k$. Let $\mathbf{x}_n^{(i)}$ be the response pattern of person n without item i, and let $r(\mathbf{x}_n^{(i)})$ be the unweighted sum score on this partial response pattern, that is,

$$r(\mathbf{x}_n^{(i)}) = \sum_{g \neq 1}^{k} \sum_{j=1}^{m_g} jx_{ngj}.$$

Based on the partial sum score $r(\mathbf{x}_n^{(i)})$, the sample is divided into S_i subgroups $s_i = 1, \ldots, S_i$. As an alternative model to the null model, which is the unidimensional GPCM, a model is considered where the probability of responding in category j of item i conditional on the subgroup s is given by

$$P(X_{nij} = 1 \mid s, \theta_n \mathbf{a}_i, \mathbf{b}_i \delta_{is}) = \frac{\exp(a_i j\theta_n + j\delta_{is} - b_{ij})}{1 + \sum_{h=1}^{m_i} \exp(a_i h\theta_n + h\delta_{is} - b_{ih})} \tag{10}$$

for $j = 1, \ldots, m_i$. Under the null model, the additional parameter δ_{is} is equal to zero. Notice that parameter δ_{is} is multiplied by θ, so in the model its role is analogous to that of θ. The alternative model entails that the latent parameter θ is insufficient to describe the response behavior, and some shift related to the response level must be incorporated. Following Glas (1999), it can be inferred that the first-order derivative of the likelihood with respect to δ_{is} is given by

$$-\sum_{n \mid s} \sum_{j=1}^{m} jx_{nij} + \sum_{n \mid s} \sum_{j=1}^{m} jE(\psi_{ij}(\theta_n) \mid \mathbf{x}_n, \mathbf{a}, \mathbf{b}, \mu, \sigma), \tag{11}$$

where $\psi_{ij}(\theta_n)$ is the GPCM specified by Equation 1 with $Q = 1$. Dividing both terms in Equation 11 by the number of persons in subgroup s, we obtain a test based on the difference between the observed average score on item i in score level group s and its posterior expectation. The expected value is computed using the GPCM without the additional parameters, the null model, given the MML estimates. If for item i the difference between the observed and expected values is large, the GPCM model does not fit the data for that item. That is, the additional parameter δ_{is} is required to obtain model fit, so the null hypothesis $\delta_{is} = 0$ is rejected.

It must be remarked that the alternative model has to be identified. This is accomplished by setting the last additional parameter δ_{is} equal to zero. Therefore, the LM statistic is based on $S_i - 1$ residuals, and it has an asymptotic χ^2 distribution with $S_i - 1$ degrees of freedom. It should be noted that this LM statistic gives insight in model fit for moderate sample sizes. For large sample sizes, this statistic becomes less interesting because its power becomes so large that even the smallest deviations from the model become significant. In that case, the effect size becomes more important than the significance probability of the test. As effect size, we considered the difference between the observed and expected average score on an item in a score level group.

Model Fit Over Time Points

One of the purposes of the methods discussed here is to evaluate trends over time, for instance by testing the null hypothesis that the means of the ability distributions of a scale over time points are equal. Testing this hypothesis is only meaningful if the measurement instrument does not change over time—more specifically, if the item parameters are constant over time.

For if the position of the item location parameters on the latent scale change, the reference point for evaluation of change in the ability parameters becomes illusive. The fact that the item and population parameters are concurrently estimated using MML makes it possible to test this hypothesis. We present two methods to evaluate the stability of the scales.

It should be noted that items that show parameter drift should not necessarily be removed from the analysis. Keeping them in the analysis as different items at different time points may still support the precision of the estimate of the trend in the ability parameters. However, both from a conceptual viewpoint (the concept to which the trend refers) as from the viewpoint of reliability it is essential to have a substantial number of anchor items that remain stable.

The LM Test

The LM test to evaluate the fit of the item response functions can be adapted to evaluate the stability of the item parameters over time. We will first discuss a test for the stability of the location parameters b_{ij}. The location parameters are estimated using MML with the responses of all time points. The null model is the multidimensional

GPCM with linear restrictions on the item parameters that impose the constancy of the parameters over time points. As alternative to the null model, a model is considered where the probability of responding in category j of item i conditional on time point q is given by

$$P(X_{nij} = 1 | q, \theta_{nq}, \mathbf{a}_i, \mathbf{b}_i, \delta_{iq}) = \frac{\exp(a_i j\theta_{nq} + j\delta_{iq} - b_{ij})}{1 + \sum_{h=1}^{m_i} \exp(a_i h\theta_{nq} + h\delta_{iq} - b_{ih})}$$

for $j = 1, \ldots, m_i$. Under the null model, the additional parameter δ_{iq} is equal to zero. The alternative model, where $\delta_{iq} \neq 0$, entails that the latent parameter θ_{nq} is insufficient to describe the response behavior on time point q. For this LM test, the first-order derivative of the likelihood with respect to δ_{iq} is defined as

$$-\sum_{n|q}\sum_{j=1}^{m} jx_{nij} + \sum_{n|q}\sum_{j=1}^{m} jE(\psi_{ij}(\theta_{nq}))|\mathbf{x}_n, \mathbf{a}, \mathbf{b}, \mu, \sigma). \tag{12}$$

If we divide both terms of Equation 12 by the sample size at time point q, we obtain a test based on the difference between the observed average score on item i for time point q and its posterior expected value computed by the multidimensional GPCM without the additional parameter. If for item i the difference between the observed and expected values is large, it can be concluded that the location parameters of item i were not constant over time, and the additional parameter δ_{iq} was necessary to fit the multidimensional GPCM model. The model is identified by setting the shift parameter of the last time point, δ_{iQ}, equal to zero. Therefore, the LM statistic is based on $Q - 1$ residuals, and is asymptotically χ^2 distributed with $Q - 1$ degrees of freedom.

The test discussed here is a generalization of the test for differential item functioning (DIF) by Glas (1998) to longitudinal data. The test is a generalization in the sense that DIF is a difference in response behavior between groups, while the present test focuses on differences in response behavior between time points within the same group. Besides a test for uniform DIF (constancy of b_{ij}), Glas also proposed a test for nonuniform DIF (constancy of a_i). Using the approach outlined here, it is also possible to generalize this test to the present framework. The alternative model then implies adding a shift parameter to the discrimination parameter, that is, the parameter a_i would be replaced by $a_i + \xi_{iq}$, where ξ_{iq} is the shift on time point q, for instance with time point Q as a base line. Because testing for nonuniform DIF is not essentially different from testing for uniform DIF, the test for nonuniform DIF will not be detailed here further.

Linear Regression and Scatter Plots

The stability of the item parameters over time can also be investigated by comparing the estimates emanating from a unidimensional GPCM for each scale at each time point. This has the complication that the separate models are identified by setting the mean of the population distribution equal to zero. So if the population mean actually changes over time, the estimates of the difficulty parameters are subject to a linear transformation. Furthermore, if the variance of the population distribution changes over time, the estimates of the discrimination parameters would only be identical up to a multiplicative constant. So even though the population parameters will probably not be equal over time, a linear relation between the item parameter estimates must be expected when the model holds. Therefore, a linear regression line between the estimates of the item parameters obtained at pairs of time points will be calculated. In addition, a 95% confidence interval will be calculated for each parameter at each time point. For each confidence interval, it is investigated whether it intersects the appropriate regression line. If a confidence interval intersects the regression line, it can be concluded that the deviation of the estimate of the item parameter from the regression line was due to estimation error. Otherwise, it is evidence for structural change over time.

An Example

The method is illustrated by an example, which comprises a part of a school effectiveness study where systematical differences between pupils of Flemish secondary schools on achievement and attitudes were evaluated. At the end of the 1st, 2nd, 4th, and 6th school year, several cognitive and noncognitive variables of 2,207 pupils who passed all end-of-year examinations were measured (Van Landeghem & Van Damme, 2002). The cognitive variables were the achievement on Dutch language and mathematics. The noncognitive variables were measured by a questionnaire, which contained 104 items. A varimax rotated factor analysis indicated that these variables comprised eight scales. The scales “Well-Being at School” (4 items), “Academic Self-Concept” (9 items), and “Attentiveness in the Classroom” (10 items) were used in the present example. Questions of these scales are, for instance, “If the choice was mine, I would rather go to another school”; “My classmates are better at learning than me”; and “In class I am often thinking about things that have nothing to do with the lesson.” Each item of the three scales had five response categories: *strongly agree, agree, neutral, disagree*, and *strongly disagree*. Some items were recoded to give them the proper orientation. Categories with less than 20 observations were combined with categories above, except for the highest category, which was combined with the category below. Because of the limitation of the estimation procedure, only the first three time points were examined. Thus, Time Point 1 is the 1st school year. Of the total data set, only the responses of pupils who answered the questionnaire at all three time points were considered, which resulted in a data set of responses of 1,942 pupils. A number of questionnaires were filled out incompletely, so item nonresponse occurred.

An overview of some classical test theory indices is given in Table 1. These indices are based on the responses of the 1,749 complete cases only. The first panel relates to “Well-Being at School,”, the second to “Academic Self-Concept,” and the third to “Attentiveness in the Classroom.” For all items, the number of response categories is given in the column with heading “Cat.” For each item at each time point, the mean score, the standard deviation, and the item-total correlations are given in the columns labeled “Mean Score,” “*SD*,”, and “Rit,” respectively. The item-total correlations indicate the contribution of the item scores to the reliability of the scale. Note that for most of the items, the mean score and the item-total correlation show a small increase over time, and the standard deviation remains almost the same. For each scale at each time point, the mean of the scale and the standard deviation are given in the rows denoted as "Mean" and "*SD*." For all scales, the means show a small increase over time. To obtain a measure of reliability of the scale at each time point Cronbach's alpha was computed. The results are presented in the row labeled "Alpha." The lower bound for reliability of a scale of .80 is reached for the scales "Well-Being at School" and "Attentiveness in the Classroom" for all time points but not for the scale "Academic Self-Concept." However, at all three time points, Cronbach's alpha was close to .80.

Model Fit Within Time Points

First, MML estimates were computed for the three scales and all time points separately and LM tests for the response functions of the items were computed. The results are presented in Tables 2, 3, and 4 for the scales "Well-Being at School," "Academic Self-Concept,", and "Attentiveness in the Classroom," respectively.

Consider first the scale "Well-Being at School," which consisted of four items having five answer categories each. The LM statistics are based on a partitioning of the sample of students in three subgroups based on the students' partial sum scores $r(\boldsymbol{x}_n^{(i)})$. The cutoff scores were chosen in such a way that the numbers of students in the subgroups were approximately equal.

In Table 2, it can be seen that 9 of the 12 LM tests were significant at a 5% significance level. The observed and expected average item scores in the subgroups are shown under the headings "Obs" and "Exp," respectively. Note that the observed average scores increased with the score level of the subgroup. An indication of the seriousness of the model violation can be obtained by computing the absolute difference between the observed and expected average scores in the subgroups. It can be seen in the column labeled "Abs Dif" that these differences were quite small: The largest absolute difference was .09 and the mean absolute difference was approximately .03. Therefore, it could be concluded that the observed average item scores fit the model quite well.

For the scale "Academic Self-Concept," 11 out of 27 LM tests were significant at the 5% level. The sample of students was divided in four subgroups of approximately the same size. Note that here also the average observed scores in the groups increased with the score level of the groups as predicted by the average expected values. The mean absolute difference between the observed and expected values equaled .02, with a maximum absolute difference of .09.

For the third scale, "Attentiveness in the Classroom," 9 out of 20 LM tests were significant at the 5% level. Again, the sample of students was divided in four subgroups of approximately the same size. The largest absolute difference is .11 and the mean average absolute difference is .02. The conclusion was that in spite of the large number of significant LM tests the model is acceptable because the differences in the observed and expected item mean score were small.

(text continued on p. 22)

Table 1
Classical Test Theory Indices for Each Scale at Each Time Point

		Time Point 1			Time Point 2			Time Point 3		
Item	Cat	Mean Score	*SD*	Rit	Mean Score	*SD*	Rit	Mean Score	*SD*	Rit
Summary statistics: Well-Being at School										
1	4	0.68	0.77	.81	0.78	0.81	.83	1.29	0.86	.84
2	4	0.68	0.92	.84	0.87	1.00	.85	1.37	1.09	.87
3	4	0.98	0.91	.79	1.18	0.96	.79	1.70	0.93	.79
4	4	0.85	0.90	.81	0.99	0.96	.83	1.41	0.99	.86
Mean			3.19			3.81			5.76	
SD			2.81			3.09			3.26	
Alpha			.82			.84			.86	
Summary statistics: Academic Self-Concept										
1	4	1.23	0.72	.53	1.74	0.77	.61	1.83	0.74	.61
2	4	1.05	0.91	.62	0.91	0.89	.64	1.20	0.88	.67
3	4	1.32	0.96	.62	0.134	0.95	.67	1.49	0.88	.68
4	4	1.44	0.98	.56	1.42	0.96	.62	1.54	0.90	.56
5	3	0.86	0.65	.66	0.93	0.65	.69	1.12	0.64	.71
6	3	0.83	0.68	.63	0.88	0.67	.68	1.15	0.66	.66
7	4	1.30	0.77	.63	1.28	0.77	.68	1.54	0.72	.66
8	4	1.24	0.77	.63	1.30	0.84	.65	1.58	0.81	.64
9	4	1.87	0.89	.47	1.90	0.86	.45	2.20	0.78	.44
Mean			11.63			1.71			13.65	
SD			4.31			4.61			4.37	
Alpha			.76			.79			.78	
Summary statistics: Attentiveness in the Classroom										
1	4	1.31	0.92	.71	1.43	0.97	.77	1.91	1.00	.79
2	4	1.47	0.90	.73	1.64	0.95	.80	2.03	0.92	.81
3	3	0.90	0.68	.62	1.06	0.71	.72	1.38	0.74	.73
4	4	1.21	0.85	.69	1.35	0.88	.77	1.61	0.89	.77

(continued)

Table 1 (continued)

		Time Point 1			Time Point 2			Time Point 3		
Item	Cat	Mean Score	*SD*	Rit	Mean Score	*SD*	Rit	Mean Score	*SD*	Rit
5	4	1.50	0.98	.71	1.68	1.00	.77	1.92	0.97	.78
6	4	1.44	0.96	.96	1.55	0.96	.74	1.91	0.94	.72
7	4	1.48	0.94	.63	1.64	1.00	.66	1.93	0.97	.68
8	3	1.10	0.77	.58	1.14	0.73	.62	1.46	0.76	.60
9	4	1.06	0.88	.74	1.27	0.94	.81	1.72	0.94	.82
10	4	0.97	0.62	.66	1.06	0.63	.69	1.32	0.68	.69
Mean			12.45			13.81			17.19	
SD			5.77			6.47			6.53	
Alpha			.87			.90			.91	

Note: Cat = number of response categories; Rit = item-total correlations.

Table 2
Results of the Lagrange Multiplier (LM) Test to Evaluate Fit of the Item Response Functions for the Scale "Well-Being at School"

					Group 1		Group 2		Group 3	
Time Point	Item	LM	Prob	Abs Dif	Obs	Exp	Obs	Exp	Obs	Exp
1	1	6.67	.04	.03	0.32	0.32	0.90	0.89	1.45	1.48
	2	16.17	.00	.04	0.22	0.21	0.87	0.86	1.81	1.85
	3	32.83	.00	.09	0.63	0.61	1.31	1.31	1.94	2.03
	4	14.06	.00	.07	0.46	0.45	1.16	1.14	1.85	1.91
2	1	1.46	.48	.02	0.31	0.31	0.89	0.87	1.58	1.59
	2	15.87	.00	.02	0.30	0.28	0.91	0.93	1.95	1.97
	3	29.11	.00	.08	0.74	0.72	1.38	1.37	2.07	2.15

	4	12.22	.00	.03	0.47	0.46	1.15	1.15	1.98	2.01
3	1	12.86	.00	.02	0.70	0.70	1.26	1.28	1.70	1.72
	2	17.09	.00	.01	0.60	0.59	1.33	1.34	2.08	2.07
	3	7.88	.05	.04	1.22	1.22	1.84	1.80	2.18	2.19
	4	1.06	.79	.02	0.72	0.72	1.41	1.41	2.00	2.02

Note: Prob = probability; Abs Dif = absolute difference; Obs = observed average item scores; Exp = expected average item scores.

Table 3
Results of the Lagrange Multiplier (LM) Test to Evaluate Fit of the Item Response Functions for the Scale "Academic Self-Concept"

					Group 1		Group 2		Group 3		Group 4	
Time Point	Item	LM	Prob	Abs Dif	Obs	Exp	Obs	Exp	Obs	Exp	Obs	Exp
1	1	7.95	.05	.03	1.37	1.37	1.68	1.71	1.90	1.87	2.10	2.07
	2	11.87	.01	.07	0.45	0.50	0.85	0.87	1.23	1.16	1.61	1.64
	3	9.51	.02	.07	0.70	0.77	1.21	1.19	1.53	1.49	1.91	1.92
	4	0.64	.89	.01	0.98	0.97	1.36	1.36	1.60	1.62	1.99	1.98
	5	10.62	.01	.03	0.32	0.33	0.73	0.71	0.99	0.98	1.31	1.34
	6	16.90	.00	.05	0.34	0.33	0.67	0.67	0.98	0.94	1.26	1.31
	7	15.17	.00	.07	0.77	0.77	1.19	1.19	1.48	1.45	1.78	1.85
	8	2.37	.50	.02	0.73	0.72	1.12	1.10	1.35	1.36	1.76	1.77
	9	2.41	.49	.04	1.56	1.54	1.81	1.85	2.04	2.01	2.21	2.23
2	1	6.50	.09	.05	1.30	1.28	1.69	1.71	1.90	1.93	2.29	2.24
	2	5.32	.15	.05	0.37	0.37	0.67	0.69	1.05	1.00	1.53	1.57
	3	2.34	.50	.02	0.71	0.72	1.16	1.18	1.56	1.54	2.07	2.06
	4	0.50	.92	.02	0.86	0.87	1.28	1.26	1.58	1.57	2.03	2.05
	5	12.80	.01	.04	0.39	0.38	0.76	0.78	1.08	1.05	1.41	1.45
	6	5.11	.16	.02	0.35	0.36	0.74	0.72	0.98	0.98	1.36	1.38

(continued)

Table 3 (continued)

					Group 1		Group 2		Group 3		Group 4	
Time Point	Item	LM	Prob	Abs Dif	Obs	Exp	Obs	Exp	Obs	Exp	Obs	Exp
	7	8.51	.04	.04	0.68	0.71	1.21	1.17	1.47	1.46	1.83	1.86
	8	1.14	.77	.02	0.77	0.77	1.15	1.14	1.42	1.42	1.90	1.92
	9	10.12	.02	.05	1.62	1.60	1.90	1.87	2.02	2.03	2.18	2.23
3	1	2.72	.61	.03	1.37	1.39	1.78	1.75	1.92	1.92	2.11	2.11
	2	10.23	.04	.05	0.60	0.60	0.90	0.94	1.23	1.18	1.52	1.49
	3	6.18	.19	.04	0.85	0.89	1.33	1.29	1.54	1.53	1.84	1.83
	4	1.66	.80	.03	1.10	1.09	1.35	1.38	1.59	1.56	1.77	1.80
	5	7.81	.10	.02	0.61	0.61	0.95	0.93	1.11	1.10	1.32	1.32
	6	9.87	.04	.03	0.71	0.71	1.01	0.98	1.14	1.13	1.33	1.34
	7	1.11	.89	.02	1.03	1.03	1.38	1.39	1.61	1.59	1.81	1.82
	8	9.10	.06	.04	1.04	1.08	1.40	1.38	1.62	1.58	1.83	1.83
	9	10.95	.03	.09	1.90	1.93	2.25	2.16	2.27	2.27	2.41	2.39

Note: Prob = probability; Abs Dif = absolute difference; Obs = observed average item scores; Exp = expected average item scores.

Table 4
Results of the Lagrange Multiplier (LM) Test to Evaluate Fit of the Item Response Functions for the Scale "Attentiveness in the Classroom"

					Group 1		Group 2		Group 3		Group 4	
Time Point	Item	LM	Prob	Abs Dif	Obs	Exp	Obs	Exp	Obs	Exp	Obs	Exp
1	1	4.08	.25	.03	0.68	0.66	1.17	1.17	1.50	1.53	2.14	2.13
	2	6.80	.08	.04	0.82	0.80	1.31	1.36	1.74	1.72	2.35	2.35
	3	11.16	.01	.06	0.43	0.45	0.86	0.80	1.01	1.01	1.38	1.42
	4	7.08	.07	.03	0.61	0.61	1.10	1.07	1.39	1.40	1.96	1.99

	5	7.36	.06	.05	0.82	0.81	1.44	1.39	1.76	1.79	2.41	2.42
	6	12.06	.01	.05	0.80	0.79	1.38	1.33	1.67	1.69	2.24	2.28
	7	17.19	.00	.07	0.99	0.92	1.36	1.42	1.72	1.73	2.22	2.22
	8	41.19	.00	.11	0.61	0.65	1.12	1.00	1.24	1.23	1.56	1.67
	9	4.66	.20	.04	0.36	0.37	0.91	0.87	1.22	1.24	1.96	1.96
	10	12.86	.00	.05	0.50	0.53	0.96	0.91	1.11	1.10	1.46	1.49
2	1	3.66	.45	.03	0.61	0.59	1.10	1.12	1.50	1.50	1.87	1.90
	2	5.00	.29	.04	0.80	0.78	1.30	1.34	1.75	1.73	2.14	2.16
	3	8.12	.09	.03	0.48	0.48	0.91	0.88	1.06	1.09	1.29	1.31
	4	2.06	.73	.02	0.59	0.60	1.08	1.07	1.37	1.37	1.74	1.73
	5	2.54	.64	.03	0.81	0.79	1.37	1.39	1.83	1.81	2.20	2.23
	6	9.68	.05	.07	0.74	0.77	1.33	1.30	1.73	1.67	1.95	2.02
	7	2.21	.70	.03	0.93	0.91	1.42	1.42	1.72	1.75	2.06	2.05
	8	31.44	.00	.09	0.59	0.64	1.02	0.98	1.27	1.18	1.38	1.38
	9	2.94	.57	.02	0.39	0.41	0.95	0.93	1.31	1.29	1.66	1.68
	10	11.04	.03	.03	0.52	0.55	0.95	0.93	1.08	1.09	1.31	1.28
3	1	11.30	.02	.04	1.01	0.99	1.49	1.49	1.84	1.88	2.33	2.37
	2	2.23	.69	.03	1.13	1.14	1.68	1.65	2.03	2.04	2.46	2.48
	3	2.91	.57	.02	0.78	0.77	1.08	1.09	1.27	1.29	1.60	1.60
	4	6.10	.19	.02	0.84	0.83	1.21	1.21	1.52	1.50	1.91	1.93
	5	10.94	.03	.05	0.97	1.00	1.58	1.53	1.94	1.94	2.40	2.39
	6	0.61	.96	.02	1.13	1.12	1.57	1.57	1.90	1.90	2.25	2.27
	7	3.91	.42	.03	1.20	1.17	1.63	1.64	1.93	1.95	2.29	2.29
	8	0.24	.99	.01	0.98	0.97	1.22	1.21	1.38	1.39	1.64	1.63
	9	2.35	.67	.02	0.82	0.81	1.24	1.26	1.62	1.62	2.13	2.12
	10	7.87	.10	.05	0.80	0.81	1.06	1.07	1.21	1.23	1.52	1.47

Note: Prob = probability; Abs Dif = absolute difference; Obs = observed average item scores; Exp = expected average item scores.

Model Fit Over Time Points

Next, MML estimates were computed with the responses of all time points jointly for the three scales separately. The LM test was computed for the GPCM with the multivariate normal distribution. To investigate the effect of modeling the dependency, the LM test was also computed for the estimates of the GPCM with a standard normal distribution. The results of these LM tests are presented in Tables 5, 6, and 7 for the scales "Well-Being at School," "Academic Self-Concept," and "Attentiveness in the Classroom," respectively.

First, consider the results of the scale "Well-Being at School" in Table 5. The degrees of freedom for the LM test is the number of answer categories minus one, which equals four for all items of the scale "Well-Being at School." For the unidimensional model, all LM tests were highly significant.

For the multidimensional model, only one LM test was not significant at a 5% significance level. However, taking the dependency of the responses into account did result in expected item mean scores that were much closer to the observed item mean scores; the mean absolute difference decreased from .25 for the unidimensional model to .02 for the multidimensional model, which is quite small.

The results for the scale "Academic Self-Concept" are shown in Table 6. The degrees of freedom of each LM test are presented in the column labeled "*df*." Note that for the unidimensional model, 1 out of 27 LM tests was not significant at the 5% level, which was also the case for the multidimensional model. However, the mean absolute difference decreased from .09 (unidimensional) to .05 (multidimensional). Again, the latter difference is quite small.

For the scale "Attentiveness in the Classroom" (Table 7), the number of nonsignificant LM tests at the 5% level increased from one to four while the dependency was taken into account. However, a lot of LM tests remained significant. On the other hand, the mean absolute difference decreased as well from .18 to .04. So also here the multidimensional model produced quite acceptable results.

To visualize the analyses using linear regression, Figure 1 shows for the scale "Attentiveness in the Classroom" the scatter plots with linear regression line for the discrimination (first panel) and location parameters (second panel) of Time Points 1 and 2 with their confidence intervals. These plots show that most of the confidence intervals did include the regression line. Table 8 shows the results of all combinations for all scales. The columns labeled "Obs" give the number of observed intervals, and the columns labeled "Incl" give the number of intervals that included the regression line. Mostly, half of the intervals did include the regression line. For Time Points 1 and 3, the minimum number of intervals that include the regression line was found. This might be a consequence of the large time difference.

In spite of the large number of significant LM tests and the large number of confidence intervals that did not include that the regression line, we still conclude that the multidimensional model is acceptable. In case of the LM tests, the differences between the observed and expected item means were quite small when the multidimensional GPCM was used to estimate the expected item means. The large number of

(text continued on p. 30)

Table 5
Results of the Lagrange Multiplier (LM) Test to Evaluate Constancy of the Location Parameters Over Time for the Scale "Well-Being at School" (degrees of freedom equal to 4)

		Unidimensional				Multidimensional			
Time Point	Item	LM	Prob	Obs	Exp	LM	Prob	Obs	Exp
1	1	834.6	.00	0.68	0.92	50.9	.00	0.68	0.70
	2	733.8	.00	0.69	0.97	44.8	.00	0.69	0.71
	3	577.5	.00	0.99	1.29	138.9	.00	0.99	1.04
	4	513.5	.00	0.86	1.09	56.2	.00	0.86	0.84
2	1	225.9	.00	0.79	0.92	126.6	.00	0.79	0.84
	2	110.2	.00	0.86	0.97	20.1	.00	0.86	0.87
	3	91.5	.00	1.19	1.29	15.3	.00	1.18	1.19
	4	83.9	.00	0.99	1.09	5.8	.21	0.99	0.99
3	1	1602.5	.00	1.29	0.92	337.5	.00	1.29	1.18
	2	1195.3	.00	1.37	0.97	181.6	.00	1.37	1.31
	3	1605.1	.00	1.70	1.29	534.2	.00	1.70	1.58
	4	1007.4	.00	1.41	1.09	169.7	.00	1.41	1.40

Note: Prob = probability; Obs = observed average item scores; Exp = expected average item scores.

Table 6
Results of the Lagrange Multiplier (LM) Test to Evaluate Constancy of the Location Parameters Over Time for the Scale "Academic Self-Concept"

			Unidimensional				Multidimensional			
Time Point	Item	*df*	LM	Prob	Obs	Exp	LM	Prob	Obs	Exp
1	1	4	30.2	.00	1.73	1.77	68.9	.00	1.73	1.71
	2	4	8.4	.06	1.05	1.06	516.3	.00	1.05	0.98
	3	4	47.8	.00	1.33	1.39	58.1	.00	1.33	1.31
	4	4	49.1	.00	1.46	1.47	287.0	.00	1.46	1.40
	5	3	167.9	.00	0.87	0.97	667.2	.00	0.87	0.91
	6	3	166.2	.00	0.84	0.96	931.7	.00	0.84	0.90
	7	4	62.8	.00	1.30	1.38	17.8	.00	1.30	1.31
	8	4	152.4	.00	1.23	1.37	553.9	.00	1.23	1.29
	9	4	57.1	.00	1.87	1.99	165.0	.00	1.87	1.94
2	1	4	12.6	.01	1.75	1.77	54.7	.00	1.75	1.72
	2	4	123.9	.00	0.93	1.06	633.3	.00	0.93	1.01
	3	4	12.8	.01	1.35	1.39	32.3	.00	1.35	1.33
	4	4	23.8	.00	1.42	1.47	8.2	.08	1.42	1.42
	5	3	19.9	.00	0.94	0.97	24.7	.00	0.94	0.93
	6	3	70.1	.00	0.89	0.96	211.5	.00	0.89	0.92
	7	4	66.9	.00	1.29	1.38	145.8	.00	1.29	1.33
	8	4	47.5	.00	1.30	1.37	37.3	.00	1.30	1.32
	9	4	49.1	.00	1.90	1.99	73.3	.00	1.90	1.95

3	1	4	37.5	.00	1.83	1.77	10.8	.03	1.83	1.86
	2	4	177.2	.00	1.20	1.06	57.9	.00	1.20	1.19
	3	4	161.4	.00	1.49	1.39	94.5	.00	1.49	1.52
	4	4	74.5	.00	1.53	1.47	75.0	.00	1.53	1.58
	5	3	395.4	.00	1.12	0.97	85.9	.00	1.12	1.07
	6	3	699.5	.00	1.16	0.96	278.9	.00	1.16	1.07
	7	4	357.4	.00	1.54	1.38	88.5	.00	1.54	1.49
	8	4	397.8	.00	1.56	1.37	107.4	.00	1.56	1.48
	9	4	338.9	.00	2.20	1.99	176.3	.00	2.20	2.06

Note: Prob = probability; Obs = observed average item scores; Exp = expected average item scores.

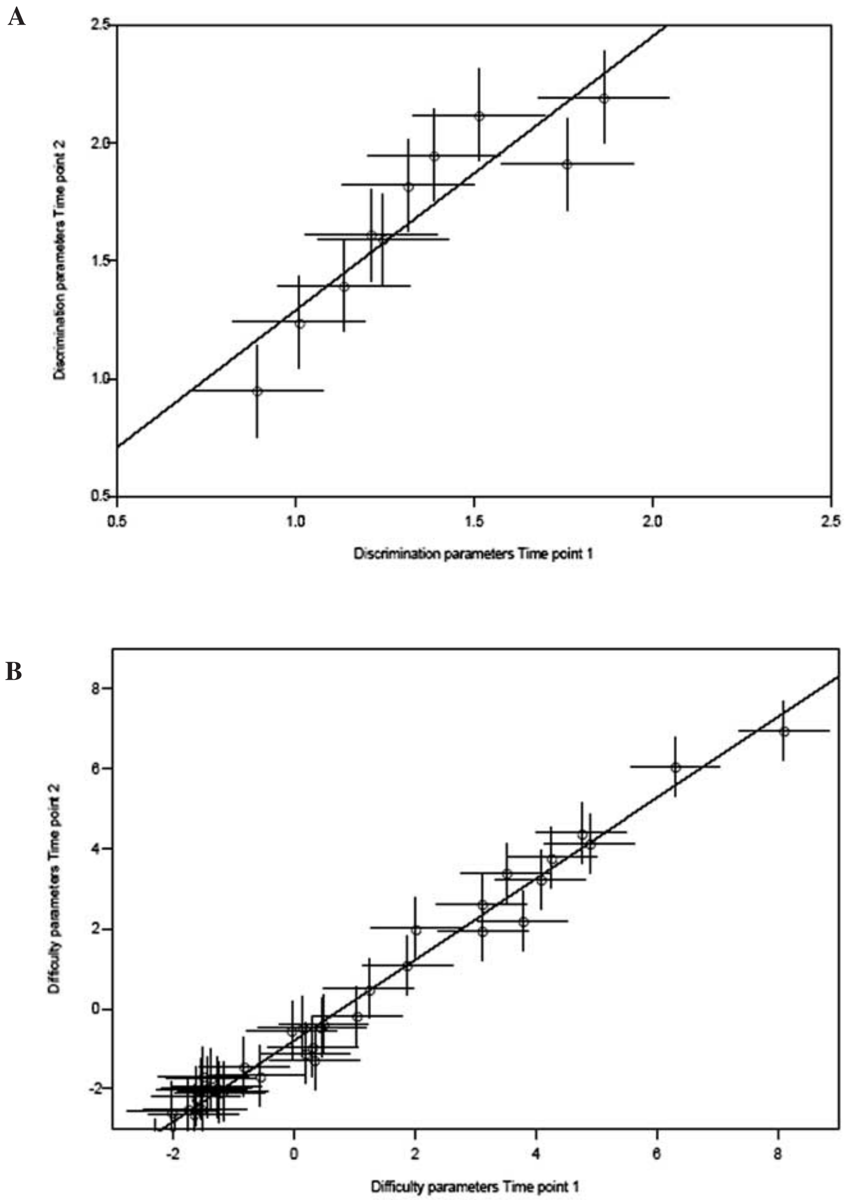
Table 7
Results of the Lagrange Multiplier (LM) Test to Evaluate Constancy of the Location Parameters Over Time for the Scale “Attentiveness in the Classroom”

			Unidimensional				Multidimensional			
Time Point	Item	*df*	LM	Prob	Obs	Exp	LM	Prob	Obs	Exp
1	1	4	335.4	.00	1.31	1.55	39.9	.00	1.31	1.36
	2	4	399.5	.00	1.47	1.71	66.6	.00	1.47	1.52
	3	3	401.1	.00	0.90	1.12	191.7	.00	0.90	0.98
	4	4	205.0	.00	1.21	1.39	27.0	.00	1.21	1.21
	5	4	204.7	.00	1.51	1.70	24.7	.00	1.51	1.50
	6	4	176.3	.00	1.45	1.63	25.2	.00	1.45	1.45
	7	4	182.0	.00	1.50	1.69	5.0	.29	1.50	1.52
	8	3	122.0	.00	1.10	1.24	37.5	.00	1.10	1.11
	9	4	658.7	.00	1.06	1.35	202.8	.00	1.06	1.15
	10	3	255.7	.00	0.98	1.12	32.9	.00	0.98	1.00
2	1	4	80.1	.00	1.43	1.55	63.4	.00	1.43	1.50
	2	4	44.5	.00	1.64	1.71	17.9	.00	1.64	1.67
	3	3	35.5	.00	1.06	1.12	14.6	.00	1.06	1.08
	4	4	11.6	.02	1.35	1.39	2.1	.71	1.35	1.34
	5	4	8.4	.08	1.67	1.70	8.1	.08	1.67	1.65
	6	4	33.9	.00	1.56	1.63	12.1	.02	1.56	1.59
	7	4	16.9	.00	1.64	1.69	5.2	.26	1.64	1.65
	8	3	56.6	.00	1.15	1.24	38.0	.00	1.15	1.21
	9	4	54.8	.00	1.27	1.35	20.1	.00	1.27	1.30
	10	3	40.3	.00	1.07	1.12	15.4	.00	1.07	1.09
3	1	4	813.7	.00	1.91	1.55	104.4	.00	1.91	1.79

2	4	1011.4	.00	2.03	1.71	102.5	.00	2.03	1.95
3	3	855.3	.00	1.39	1.12	148.3	.00	1.39	1.28
4	4	429.1	.00	1.60	1.39	50.3	.00	1.60	1.60
5	4	389.7	.00	1.92	1.70	43.0	.00	1.92	1.95
6	4	479.4	.00	1.90	1.63	37.6	.00	1.90	1.86
7	4	307.9	.00	1.93	1.69	10.5	.03	1.93	1.90
8	3	502.9	.00	1.46	1.24	101.8	.00	1.46	1.39
9	4	1303.0	.00	1.71	1.35	218.4	.00	1.71	1.59
10	3	478.0	.00	1.32	1.12	42.4	.00	1.32	1.27

Note: Prob = probability; Obs = observed average item scores; Exp = expected average item scores.

Figure 1
Parameter Estimates at Different Time Points



Note: For the scale "Attentiveness in the Classroom,' Panel (a) gives a scatter plot of the discrimination parameters of Time Points 1 and 2 with the estimated regression line. The horizontal and vertical lines represent the corresponding confidence intervals. Panel (b) shows a scatter plot of the location parameters of Time Points 1 and 2.

Table 8
Evaluation of the Stability of the Item Parameters Using Linear Regression

	Location Parameters						Discrimination Parameters					
	Well-Being		Self-Concept		Attentiveness		Well-Being		Self-Concept		Attentiveness	
Time Point	Obs	Incl	Obs	Incl	Obs	Incl	Obs	Incl	Obs	Incl	Obs	Incl
1	16	10	34	17	37	20	4	2	9	5	10	6
2	16	11	34	19	37	19	4	2	9	5	10	5
1	16	3	34	7	37	6	4	2	9	3	10	3
3	16	4	34	8	37	10	4	2	9	4	10	3
2	16	3	34	10	37	8	4	3	9	6	10	5
3	16	6	34	12	37	12	4	3	9	6	10	5

Note: For each scale, for all pairs of time points the number of observed 95% confidence intervals are presented and the number of intervals that included the regression line. Obs = observed intervals; Incl = intervals that included the regression line.

result of the large sample size ($N \sim 1{,}942$). As a consequence of a large sample size, the standard errors of the item parameters were small, which resulted in small confidence intervals.

Evaluation of Trends

In the previous section, analyses were presented that support the validity of the constructs "Well-Being at School," "Academic Self-Concept," and "Attentiveness in the Classroom." In this section, the trends over time will be investigated. Table 9 shows for each scale the mean on each time point and the covariance matrix over time points. The left panels give the MML estimates per scale. The means and covariance matrix were estimated concurrently with the item parameters. The right panels pertain to a concurrent MML estimate of all item parameters, the means of all scales at all time points and the complete covariance matrix. The three covariance matrices displayed are only a part of the complete covariance matrix, because the latter matrix also contains the covariances between the scales at the different time points.

The standard errors of the estimates are given within parentheses. The values marked by a superscripted *a* are the correlations between the latent dimensions. For all scales, the mean increased over time. Note that to identify the multidimensional GPCM, the means of the latent abilities at the last time point were set equal to zero. From comparing the estimated means with their standard errors, it can be concluded that the means at the different time points differed significantly. Therefore, the null hypothesis of no trend was rejected. Examining the correlation of the latent variables over time for each scale, it can be seen that the correlation structures showed the same pattern. The highest correlation was between the latent variables of Time Points 1 and 2. As expected, because of the largest time difference the lowest correlation was between the latent variables of Time Points 1 and 3. Finally, it can be seen that the concurrent estimates and estimates obtained for separately for all three scales are very close.

Conclusion and Discussion

The present article shows the application of multidimensional IRT models for analyzing repeated measures. IRT models provide direct estimates of the relations between the latent variable on several time points and draw on the (often strong) relationship between the latent variables to produce more accurate parameter estimates. Furthermore, incomplete designs and missing responses are no longer obstacles, and all the available information is used to estimate more accurate item and population parameters.

The between-item multidimensional GPCM was proposed to model repeated measures. However, the approach is not limited to the GPCM, and many other unidimensional IRT models can be generalized to between-item multidimensional IRT models. The item and population parameters were estimated with an adapted MML estimation procedure, imposing linear restrictions on the item parameters to obtain constant item

Table 9

Evaluation of Trends: The Mean and Covariance and Correlation Matrix Estimated for Each Scale and Estimated Concurrently for All Scales (the Left Panel Gives the Separate Estimates, and the Right Panel Gives the Concurrent Estimates)

				Covariances								Covariances					
				Time Point								Time Point					
Scale	Time Point	Mean		1		2		3		Mean		1		2		3	
Academic	1	–0.240	(.031)	1.590	(.022)					–0.223	(.031)	1.418	(.023)				
Self-Concept	2	–0.207	(.031)	1.189	(.022)	1.771	(.024)			–0.198	(.031)	1.076	(.022)	1.424	(.022)		
				.709[a]								.757[a]					
	3	0.000		0.787	(.021)	0.930	(.022)	1.501	(.020)	0.000		0.759	(.022)	0.905	(.022)	1.498	(.021)
				.509[a]		.570[a]						.521[a]		.620[a]			
Attentiveness	1	–0.323	(.033)	1.268	(.023)					–0.312	(.033)	1.257	(.022)				
in the Classroom	2	–0.128	(.032)	0.876	(.022)	1.424	(.022)			–0.111	(.033)	0.736	(.021)	1.491	(.022)		
				.652[a]								.538[a]					
	3	0.000	0.559	(.020)	0.805	(.020)	1.419	(.025)	0.000	0.588	(.019)	0.864	(.021)	1.473	(.023)		
				.417[a]		.566[a]						.432[a]		.583[a]			
Well-Being	1	–0.780	(.038)	1.957	(.025)					–0.699	(.033)	1.957	(.022)				
at School	2	–0.491	(.033)	1.536	(.024)	2.191	(.412)			–0.477	(.033)	1.496	(.021)	1.916	(.022)		
				.742[a]								.773[a]					
	3	0.000		0.588	(.020)	0.964	(.451)	1.973	(.023)	0.000		0.578	(.025)	1.008	(.021)	1.969	(.023)
				.299[a]	.464[a]							.294[a]		.519[a]			

Note: Standard errors appear in parentheses.

a. These values are the corresponding correlations.

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

To investigate the validity of the results, model fit within and between time points was examined using an LM test to evaluate the fit of the item response functions. The LM statistics were significant in more than half of the cases. The explanation is that the power of the tests increases dramatically with the number of observations. Therefore, with large sample sizes, the sizes of the residuals are much more informative because they give an indication of the seriousness of the model violation. In the present example, the model violations were judged as acceptable. The reason for implying the LM statistics at all is that they give a motivation for considering specific residuals such as Equations 11 and 12, because the alternative hypotheses on which the tests are based indicate which model violations are exactly targeted. Further, the LM statistics contain a proper estimate of the covariance matrix with which the residuals should be weighted (see Glas & Suarez-Falcon, 2003).

References

Adams, J. A., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50*, 3-16.

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

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: 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.

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

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

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*, 495-515.

Fischer, G. H. (2003). The precision of gain scores under an item response theory perspective: A comparison of asymptotic and exact conditional inference about change. *Applied Psychological Measurement, 27*, 3-26.

Fischer, G. H., & Parzer, P. (1991). An extension of the rating-scale model with an application to the measurement of change. *Psychometrika, 56*, 637-651.

Fischer, G. H., & Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika, 59*, 177-192.

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

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

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

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

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

Holman, R., & Glas, C. A. W. (2005). Modelling non-ignorable missing data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology, 58*, 1-17.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating*. New York: Springer.

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

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

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

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

Muraki, E., & Bock, R. D. (2002). *PARSCALE: Parameter scaling of rating data*. Chicago: Scientific Software, Inc.

Muthén, L. K., & Muthén, B. O. (2003). *Mplus version 2.14*. Los Angeles, CA: Muthén & Muthén.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50-64.

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). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*, 25-36.

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

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

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.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408.

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

Thissen, D., Chen, W.-H., & Bock, R. D. (2002). *Multilog*. Lincolnwood, IL: Scientific Software International.

Trivellato, U. (1999). Issues in the design and analysis of panel studies: A cursory review. *Quality & Quantity, 33*, 339-352.

Van Damme, J., De Fraine, B., Van Landeghem, G., Opdenakker, M.-C., & Onghena, P. (2002). A new study on education effectiveness in secondary schools in Flanders: An introduction. *School Effectiveness and School Improvement, 13*, 383-397.

Van Landeghem, G., & Van Damme, J. (2002, May 30). *De evolutie van individueel welbevinden, academisch zelfconcept en prestaties doorheen het middelbaar onderwijs: Verschillen tussen scholen* [Changes in well-being, academic self-concept and achievement of secondary school students: Differences between schools]. Symposium presentation at the Onderwijs Research Dagen, Antwerp, Belgium.

Verhelst, N. D., & Glas, C. A. W. (1995). Dynamic generalizations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 181-201). New York: Springer-Verlag.

Wang, W. C., & Wu, C. Y. (2004). Gain score in item response theory as an effect size measure. *Educational and Psychological Measurement, 64*, 758-780.

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

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

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2002). *Bilog-MG*. Lincolnwood, IL: Scientific Software International.