Behav Res (2014) 46:823–840
DOI 10.3758/s13428-013-0413-3

A generalized longitudinal mixture IRT model for measuring differential growth in learning environments

Damazo T. Kadengye · Eva Ceulemans · Wim Van den Noortgate

Published online: 21 November 2013
© Psychonomic Society, Inc. 2013

Abstract This article describes a generalized longitudinal mixture item response theory (IRT) model that allows for detecting latent group differences in item response data obtained from electronic learning (e-learning) environments or other learning environments that result in large numbers of items. The described model can be viewed as a combination of a longitudinal Rasch model, a mixture Rasch model, and a random-item IRT model, and it includes some features of the explanatory IRT modeling framework. The model assumes the possible presence of latent classes in item response patterns, due to initial person-level differences before learning takes place, to latent class-specific learning trajectories, or to a combination of both. Moreover, it allows for differential item functioning over the classes. A Bayesian model estimation procedure is described, and the results of a simulation study are presented that indicate that the parameters are recovered well, particularly for conditions with large item sample sizes. The model is also illustrated with an empirical sample data set from a Web-based e-learning environment.

Keywords Item response theory · E-learning · Modeling of growth · Mixture models

D. T. Kadengye · W. Van den Noortgate
Faculty of Psychology and Educational Sciences and ITEC–iMinds, University of Leuven–Kulak, Kortrijk, Belgium

D. T. Kadengye · E. Ceulemans · W. Van den Noortgate
Centre for Methodology of Educational Research, University of Leuven, Leuven, Belgium

D. T. Kadengye (✉)
Faculty of Psychology and Educational Sciences, KU Leuven–Kulak, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium
e-mail: dkadengye@gmail.com

Longitudinal item response theory (IRT) models (Andersen, 1985; Andrade & Tavares, 2005; Embretson, 1991; Fischer, 1989; Paek, Baek, & Wilson, 2012; Roberts & Ma, 2006; von Davier, Xu, & Carstensen, 2009) can be used to measure individual differences in growth (or change) over time. Measurement of growth in student performance between two or more testing occasions is a central topic in educational research and assessment (Fischer, 1995). In educational learning environments, such longitudinal IRT models can be interesting for researchers whose interest is in tracking and monitoring persons' progress over time in response to a treatment such as teaching or training (Miceli, Settanni, & Vidotto, 2008). This is the case in studies in which students are followed from one grade to another over a period of time in such a way that, every academic year, students are subjected to a set of different tests—for instance, at the beginning of the term and at midterm, end of term, and end of year. In this way, educators can evaluate the effectiveness of their employed training procedures, so that they may be in position to make more informed instructional decisions (Safer & Fleischman, 2005).

Modeling growth in individuals

One approach for modeling growth over time is the use of the multidimensional Rasch model of Andersen (1985) or that of Embretson (1991), which can be employed when the same or separate sets of items are repeatedly answered by a group of persons over different measurement occasions in time. For instance, Embretson's multidimensional Rasch model for learning and change assumes the involvement of M abilities in item responses within K measurement occasions, such that (a) on the first measurement occasion ($k = 1$) only an initial ability (θ_{p1}) is involved in the item responses of person p, and (b) on the later $k > 1$ occasions, that ability (θ_{p1}) plus $k - 1$ additional abilities are involved in the performance.

Embretson's model can be expressed as

$$\mathrm{P}\left(Y_{pik}=1\middle|\left(\theta^{*}_{p1},\theta^{*}_{p2},\ldots,\theta^{*}_{pk}\right),\beta_i\right)=\exp\left(\sum_{m=1}^{k}\theta^{*}_{pm}-\beta_i\right)\Big/ 1+\exp\left(\sum_{m=1}^{k}\theta^{*}_{pm}-\beta_i\right), \quad (1)$$

where $\theta^{*}_{p1}=\theta_{p1}$ is the baseline level (Occasion 1) ability for person p ($p=1,2,\ldots,P$), $\theta^{*}_{pk}=\theta_{pk}-\theta_{p(k-1)}$ is the change in ability for person p from occasion $k-1$ to occasion k (for $k=2,\ldots,K$) and $\pi_{pik}=\mathrm{Prob}(Y_{pik}=1|\theta_{pk},\beta_i)$ is the probability that person p with ability $\theta_{pk}=\sum_{m=1}^{k}\theta^{*}_{pm}$ at occasion k will answer item i ($i=1,2,\ldots,I$) with difficulty βi correctly on that particular occasion. The θs are assumed to be multivariate and normally distributed over persons, whereas the item difficulty parameters, βi, are assumed to be the same for all occasions. Note that only $I-1$ item difficulties can be estimated from Eq. 1, unless the intercept is constrained to be zero (otherwise, the model is not identified).

Equation 1, as with other common IRT models, can be regarded as a logistic mixed model (Kamata, 2001). Furthermore, instead of assuming unknown, fixed values of the item difficulties (as in Eq. 1), the considered items can also be regarded as a random sample from a population of items and therefore item difficulties can also be modeled as random effects over items (De Boeck, 2008; De Boeck & Wilson, 2004; Van Den Noortgate, De Boeck, & Meulders, 2003). As such, the item difficulties are assumed to be randomly and normally distributed over items but constant over occasions. In this way, Eq. 1 can be simplified to a generalized linear mixed model (GLMM) with crossed random effects, as

$$\mathrm{logit}\left(\pi_{pik}\right)=b_0+\sum_{k=2}^{K}b_kD_{pik}+\sum_{k=1}^{K}\omega_{pk}D_{pik}+\upsilon_i, \quad (2)$$

where b_0 is the expected initial person ability, b_k is the difference from b_0 at occasion k such that the expected overall ability at occasion k is b_0+b_k, and D_{pik} is a dummy variable equal to 1 if item i was solved by person p at session k, or 0 otherwise. The variable ω_{pk} is the random deviation of the ability of person p from the overall ability at occasion k, that can be assumed multivariate normal $\boldsymbol{N}(\mathbf{0},\boldsymbol{\Sigma}^{*})$ over persons with $\boldsymbol{\Sigma}^{*}$ as the variance–covariance matrix. In this case, the ability estimate of person p during occasion k, θ_{pk} in Eq. 1, is now given by $b_0+b_k+\omega_{pk}$. The υ_is are random item parameters with $\upsilon_i\sim N(0,\sigma^2_{\upsilon})$, where υ_i corresponds to $-\beta_i$ of Eq. 1, and therefore cannot be interpreted as the item difficulty, but rather as the item easiness.

Measuring growth in e-learning environments

A specific application is in electronic learning (e-learning) environments in which persons can freely engage items online as part of formal education to build and improve their ability (for examples, see Desmet, Paulussen, & Wylin, 2006; Klinkenberg, Straatemeier, & Van der Maas, 2011). For each engaged item, an individual gets instant feedback (e.g., about the correctness of the answer) to enhance learning. On the basis of the time points at which the items are engaged, we can define for each person one or more study sessions. For instance, if one day minimally separates answering two items, we could conclude that the person has started a new study session (see, e.g., Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Vlach & Sandhofer, 2012). If we consider responses to items that are engaged in the same study session as being solved at one measurement occasion, Eq. 2 can be employed to track persons' growth from one study session to another. In this case, items need not be the same for all sessions, or for each person, provided that not all items solved at a specific occasion are solved at that occasion only, so that scales are still linked across sessions and across persons. This requirement is even not necessary if item difficulties can already be considered as known—for instance, on the basis of a prior calibration study.

In some e-learning environments, the total number of available items can be very large, but the number of items per study session small. For instance, the item bank of the Math Garden e-learning environment (Klinkenberg et al., 2011) consists of 2,784 items. Yet, each study session corresponds to 15 items only. Such a scenario can lead to many study sessions such that, if Eq. 2 is employed to the data, the number of parameters to estimate can be very large, thereby limiting easy interpretation of the results. Moreover, due to user freedom, the number of items per session over persons is likely to vary greatly, as well as the space between study sessions. This further makes Eq. 2 less suitable for comparisons of individuals from one study session to another. Instead, we can use session as a continuous predictor variable and, for instance, define a person's ability as linearly changing over study sessions. In this way, the GLMM in Eq. 2 can be rewritten as

$$\mathrm{logit}\left(\pi_{pik}\right)=\left(b_0+\omega_{0p}\right)+\left(b_1+\omega_{1p}\right)\mathrm{session}_{pk}+\upsilon_i, \quad (3)$$

where Y_{pik} is equal to 1 if person p answers item i correctly during the kth session, $Y_{pik}\sim\mathrm{binomial}(1,\pi_{pik})$, and $\mathrm{session}_{pk}$ is the kth study session of person p. Therefore, b_0 is the expected initial ability, b_1 is the overall population slope in person abilities for the sessions, ω_{0p} is the random deviation of person p from the overall initial ability b_0, ω_{1p} is the random deviation of person p from the overall slope b_1 over sessions, and υ_i is the random item easiness parameter, as in Eq. 2. The population distribution of the two random person-specific effects (ω_{0p},ω_{1p}) can be assumed to be multivariate

normal $N(\mathbf{0},\boldsymbol{\Sigma})$ over persons, with $\boldsymbol{\Sigma}$ as the variance–covariance matrix. The parameters that are estimated are the mean initial ability and growth slope, b_0 and b_1, as well as the (co)variances of the random effects, σ_υ^2 and Σ. The value of $b_0+\omega_{0p}+(b_1+\omega_{1p})session_{pk}$ corresponds to θ_{pk}, the ability of person p during session k. The linear growth assumption is considered here for simplicity of explanation purposes, but in practice other growth assumptions, such as quadratic or exponential, can be considered.

In situations in which the interest is in monitoring growth for observed groups of persons rather than in focusing on the whole population of individuals, indicator variables for manifest person groups like gender, course of study, or ethnicity can be included as predictors in Eq. 3 (for examples, see Cho, Athay, & Preacher, 2013; Rost, 1990). This is similar to the explanatory IRT modeling framework of De Boeck and Wilson (2004), in which IRT models are formulated as GLMMs. Inclusion of manifest person groups as predictors offers flexible ways for predicting and understanding differences in ability and preferences (of groups) of persons. For instance, if persons are classified in two or more groups, including in Eq. 3 a set of dummy indicators for group membership and their interactions with the session number would provide parameter estimates for group-specific growth trends. Furthermore, interaction terms of person groups by item indicators can be included in the model to assess the presence of differential item functioning (DIF), as was demonstrated by Van Den Noortgate and De Boeck (2005). A heteroscedastic explanatory IRT model further allows person random effects' variance to be different for the manifest groups' levels.

Including person group indicators is only possible when information on groups has been recorded during data collection. However, information on some sensitive person properties, such as gender or ethnicity, is often not readily available in e-learning environments, due to user independence and freedom, design, or nonresponse. Even when group information is available, Samuelsen (2008) has argued that manifest characteristics can be poor proxies in certain situations for latent groups such as educationally advantaged or disadvantaged students. Rather, it is more common that persons belonging to the same manifest group will belong to different latent classes, because the set of characteristics that is the root of the latent class formation is not exactly the same as that of the manifest group formation (Bilir, 2009). In such situations, understanding the reasons behind differential performance can be difficult (Bilir, 2009; Dai & Mislevy, 2009; Samuelsen, 2008).

Such unobserved heterogeneity can be addressed by mixture IRT models, which infer latent groups from the data (Cho & Cohen, 2010). The mixture IRT models that are employed in most research studies are generic extensions of Rost's (1990) mixture Rasch model, which assumes that a population of persons is composed of a fixed number of exhaustive and mutually exclusive latent classes. In the mixture Rasch model, the logit of the probability of a correct response by person p from class g on item i can be given as

$$\text{logit}\left(\pi_{pig}\right) = b_{0g} + \omega_{pg} + \upsilon_{ig} \text{ and } Y_{pig} \sim \text{binomial}\left(1, \pi_{pig}\right), \quad (4)$$

where $g=1,2,\ldots,G$ is an index for mutually exclusive latent classes, b_{0g} is the expected initial ability for class g, ω_{pg} is the random deviation of person p in class g from b_{0g}—the overall initial ability of class g—and υ_{ig} is the easiness parameter of item i for class g. The structure of the ability in Eq. 4 is such that $\omega_{pg} \sim N(0,\sigma_g^2)$, where σ_g^2 is the class-specific variance of ability. Within each latent class, the model described in Eq. 4 is assumed to hold, but each class may have different item easiness parameters and another distribution of person ability parameters, although the model can be simplified by constraining some parameters to be the same over classes.

Mixture IRT models have been widely used in educational science and psychology. They have, for instance, been applied to investigating individual differences in the selection of response categories in multiple-choice items of a college-level English placement test (Bolt, Cohen, & Wollack, 2001); for detecting latent groups of persons for whom items function differently (Cohen & Bolt, 2005; Mislevy & Verhelst, 1990); for identifying latent subgroups of adolescents on the basis of their propensity to engage in sexually and substance use risky behaviors (Finch & Pierson, 2011); and for describing DIF at both the school and student levels for a standardized mathematics test in a multilevel setting (Cho & Cohen, 2010). In the recent past, mixture IRT models have started to gain considerable attention in longitudinal IRT models for measuring change, specifically those for understanding and explaining differential growth as a result of learning or training in educational environments (cf. Cho, Bottge, Cohen, & Kim, 2011; Cho, Cohen, & Bottge, 2013; Cho, Cohen, Kim, & Bottge, 2010).

This study extends on this growing research domain of longitudinal mixture IRT modeling, but focuses mainly on data sets from Web-based e-learning environments. Specifically, we incorporate the growth trend of Eq. 3 into the mixture IRT model of Eq. 4 in order to estimate differential latent growth that might be present in data sets from e-learning environments. The resulting generalized longitudinal mixture IRT model can be used to detect and compare latent classes in the data as a result of latent differential growth trends or other unobserved characteristics. We expected that this model will provide possibilities for estimates of growth trends specific to unobserved groups in a population of persons, as compared to the longitudinal model in Eq. 3, which provides estimates for the whole population. The remainder of the article will be organized as follows: First, the generalized longitudinal mixture IRT model is formulated, followed by a brief discussion on the estimation of model parameters. Next, a simulation study design used to evaluate the model is described. The obtained results are then discussed by focusing on the ability

of the presented model to recover the simulated parameters and the correct classification of participants under different conditions. The proposed model is then applied to an empirical example, and finally, a brief discussion is given.

The generalized longitudinal mixture IRT model

The proposed generalized longitudinal mixture IRT model allows for a mixture of latent classes of persons that differ from each other in two different aspects. First, the proposed model assumes that response patterns may be heterogeneous over classes due to initial person-level differences (Mislevy & Verhelst, 1990; Rost, 1990). This can, for instance, be the case in situations in which the random person ability values are heterogeneous, with persons belonging to one class assumed to have, on average, higher abilities than their counterparts in other classes before learning takes place. Furthermore, the proposed model assumes that growth trends over time can also be heterogeneous, with persons in one class tending to have, on average, higher growth trends than those in other classes. This can happen, for instance, when one group of persons is more motivated to learn or solve items than the other, or when one group has better learning strategies than the other. The model, however, does not exclude the possibility that there may be no initial differences between latent person classes, but only class-specific growth trends, and vice versa. This is because the utility of mixture IRT modeling approach lies in the fact that latent classes, although not immediately observable, are defined by certain shared response patterns that can be used to help explain item-level performance about how the members of one latent class differ from those of another (Cho & Cohen, 2010).

In this regard, the probability of getting a correct score in the generalized mixture longitudinal IRT model can be obtained from Eqs. 3 and 4 as follows:

$$\text{logit}\left(\pi_{pigk}\right) = b_{0g} + \left(b_{1g} + \omega_{1pg}\right)session_{pgk} + \omega_{0pg} + \upsilon_{ig}, \quad (5)$$

where b_{0g} and b_{1g} are defined as before, but each is now specific to latent class g; ω_{0pg} and ω_{1pg} are the respective class-specific random deviations for initial ability and the sessions' slopes, assumed to be distributed with $\omega_{0pg} \sim N(0,\sigma^2_{\omega 0g})$ and $\omega_{1pg} \sim N(0,\sigma^2_{\omega 1g})$; and υ_{ig} is the random effect for item i within class g, with class-specific distribution assumptions such that $\upsilon_{ig} \sim N(0,\sigma^2_{\upsilon g})$. In this model, each person is parameterized by a class-specific parameter, such that the ability parameter for person p in any given study session is given by $\theta_{pgk}=b_{0g}+\omega_{0pg}+(b_{1g}+\omega_{1pg})\text{session}_{pgk}$. For each latent class, the item easiness values υ_{ig} can be obtained and investigated further. When some of the item easiness values are different across the latent classes, this is considered as DIF.

From Eqs. 4 and 5, if, for instance, one has two classes and a dummy indicator for each class is added, one cannot estimate a common intercept anymore; rather, class-specific intercepts are obtained. Similarly, for Eq. 5, if interaction terms between session and the two dummies are included, one can not include the main effect of session anymore to allow for model identification. The model in Eq. 5 can be looked at as belonging to the family of the GLMMs described by De Boeck and Wilson (2004), or to the generalized linear latent and mixed models (GLLAMM; cf. Rabe-Hesketh, Skrondal, & Pickles, 2004a), since both an item response model and a mixture model are expressed simultaneously in it. The specificity of this model lies in the incorporation of the learning trends and the fact that these trends can be specific to unobserved discrete classes. Once obtained, relations with external variables can be identified in order to understand what the different latent classes are.

Parameter estimation

The parameters of the generalized longitudinal mixture IRT model can be estimated using Bayesian inference via the Markov-chain Monte Carlo (MCMC) estimation algorithm in WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). MCMC estimation algorithms are receiving increasing attention in mixture IRT models (see, e.g., Bilir, 2009; Bolt et al., 2001; Cho & Cohen, 2010; Cohen & Bolt, 2005; Dai & Mislevy, 2009; Frederickx, Tuerlinckx, De Boeck, & Magis, 2010). In MCMC estimation, a Markov chain is simulated in which values representing parameters of the model are repeatedly sampled from their full conditional posterior distributions over a large number of iterations. To derive the posterior distributions for each parameter in a mixture IRT model, it is first necessary to specify their prior distributions (Cohen & Bolt, 2005). The posterior distribution of the parameters of interest $\boldsymbol{\Omega}$, given the observed data $\boldsymbol{Y}$, is then obtained using Bayes's theorem:

$$P(\boldsymbol{\Omega}|\boldsymbol{Y}) = \frac{P(\boldsymbol{\Omega})P(\boldsymbol{Y}|\boldsymbol{\Omega})}{P(\boldsymbol{Y})} \quad (6)$$

where $P(\boldsymbol{\Omega})$ is the prior density for $\boldsymbol{\Omega}$, $P(\boldsymbol{Y}|\boldsymbol{\Omega})$ is the conditional probability of $\boldsymbol{Y}$ given $\boldsymbol{\Omega}$, and $P(\boldsymbol{Y})$ is the marginal density of the observed data $\boldsymbol{Y}$. The posterior of the model parameters are computed from Eq. 6, whose integral is typically evaluated using Monte Carlo integration techniques. For a detailed discussion of the MCMC estimation algorithm for mixture IRT models, interested readers are referred to Bolt et al. (2001).

Typically, vague or noninformative priors are used for parameters about which not much is known beforehand (Frederickx et al., 2010), but this may not always work out well in practice. In fact, Bilir (2009) and Cho and Cohen (2010) noted that noninformative priors tend to fail in providing enough bounds for the parameter estimates of a mixture IRT model. Various prior choices for the parameters,

especially those for the standard deviation components in random-effects models, have been proposed in the Bayesian literature, with no standard solution (Frederickx et al., 2010). In general, however, one considers all parameters of the model to be random variables, and mildly informative priors that provide more stable fitting procedures for the model parameters and hyperparameters can be applied (Cho & Cohen, 2010; Cohen & Bolt, 2005). For the proposed generalized longitudinal mixture IRT model, we adopted priors similar to those used in the study of Bolt et al. (2001), such that

$$\begin{aligned} b_{0g} &\sim \text{Normal}\ (0,1)\ , \\ b_{1g} &\sim \text{Normal}\ (0,1)\ , \\ \tau_{\omega_{0g}} &\sim \text{Gamma}\ (2,4)\ , \\ \sigma_{\omega_{1g}} &\sim \text{Normal}(0,1)I(0), \\ \tau_{vg} &\sim \text{Gamma}\ (2,4)\ , \end{aligned} \tag{7}$$

where $\tau_{\omega_{0g}}$ is the class-specific precision (the reciprocal of the variance) of the initial person random effects, such that $\sigma^2_{\omega_{0g}} = 1/\tau_{\omega_{0g}}$, $\sigma_{\omega_{1g}}$ is the prior of the standard deviation of random deviations from the class-specific sessions' slopes—with $I(0,)$ indicating that observations of $\sigma_{\omega_{1g}}$ are sampled above 0—and τ_{vg} is the class-specific precision of the item random effects such that, $\sigma^2_{vg}=1/\tau_{vg}$.

Within mixture models, for each person the posterior probabilities for belonging to the latent classes are estimated, and these can be used to decide which latent class a person is most likely to belong to. These probabilities are constrained such that $\sum_{g=1}^{G}\varphi_{pg}=1$ and $0\leq\varphi_{pg}\leq 1$, with φ_{pg} being equal to the estimated probability that person p belongs to class g. On the basis of these probabilities, we are also able to estimate the mixing proportions, which are the relative sizes of the latent classes—that is, the probabilities of specific classes occurring in the overall population ($\varphi_g=\sum_p\varphi_{pg}$). The posterior probabilities of a person's class membership is defined according to the frequency with which the person is sampled into each latent class over the course of the MCMC chain. This defines the posterior probability of a person's membership in a specific latent class in such a way that a person's estimated class membership is determined by the proportion of times an individual is sampled into each class over the course of the Markov chain (Cohen & Bolt, 2005). In the presence of person manifest groups, a multinomial logistic regression can be used as a prior to predict these probabilities (Bilir, 2009; Cho & Cohen, 2010; Dai & Mislevy, 2009). In the studies of Cohen and Bolt (2005) and Frederickx et al. (2010), φ_g was predicted by using a Dirichlet distribution as a prior. In a similar way, for the proposed generalized longitudinal mixture IRT model, the person-specific probabilities of class membership φ_{pg} can be predicted from $\varphi_{pg}\sim$ Dirichlet(1,1), when $G = 2$.

The number of latent classes G in a mixture IRT model is typically unknown. This makes the parameter space of infinite dimension. Therefore, a range of mixture models, each with a different number of latent classes, can be fitted, and then a selection is made out of the different models. As in the study of Cho and Cohen (2010), the Akaike information criterion (AIC) and Bayesian information criterion (BIC) can be used for selection of the model with a number of latent classes that fits best to the data. The model with the lowest AIC and BIC values is preferred. For more on (and for other) model selection procedures in mixture dichotomous IRT models, we refer the reader to Li, Cohen, Kim, and Cho (2009). Other important issues to watch out for in MCMC sampling are detection of convergence of MCMC runs and the problem of label switching—which, for instance, occurs when latent classes change meaning over the estimation chain. The consequence of label switching and/or lack of convergence is a distortion of the estimated values for the parameters of interest. In WinBUGS, several options—including time series, kernel density, and autocorrelation function plots—provide ways for monitoring the behavior of MCMC runs for estimated values after a few initial iterations, considered to be unstable, are discarded.

Basic simulation study

A basic simulation study was carried out in order to assess the performance of the proposed methodology, assuming that items are engaged by a population of persons composed of $G = 2$ discrete latent classes and varying different factors. The model was then assessed in a population with $G = 3$ for comparability purposes (but varying fewer factors, due to the feasibility of the study). The general model specification in WinBUGS is provided in the Appendix.

Data generation

Data were generated using the statistical computing software R (R Development Core Team, 2012) according to Eq. 5. A binary score, Y_{pigk}, for item i by person p during study session k within latent group g was generated from $Y_{pigk}\sim\text{Bernoulli}(1,\pi_{pigk})$. The latent class indicators, together with item difficulty, person ability, and random slope values, were then deleted from the data before exporting them to WinBUGS, through which the analysis was done. Typically, estimation of mixture IRT models is very computer intensive, and the estimation time is normally a linear function of sample size, with larger sample sizes taking longer to estimate (Bilir, 2009; Frederickx et al., 2010).

Six factors were considered in the simulation study: (1) the percentage of DIF items, (2) item sample size, (3) person sample size, (4) mixing proportions, (5) the difference between classes in the initial ability distribution, and (6) the difference between classes in the distribution of the growth slopes' deviations. In order to make the simulation study feasible, other factors were not varied, resulting in a small

number of conditions. Furthermore, only a small number of replicated data sets (five) per condition were studied, which is not uncommon in the mixture IRT literature (see, e.g., Cho & Cohen, 2010). Specific levels of these factors are described below.

Percentage of DIF items

In situations in which the same set of dichotomously scored items is administered to more than one class of persons, items many not be invariant across the classes (Muthén & Lehman, 1985)—for instance, due to DIF. Most DIF simulation studies vary the percentages of DIF items. For instance, Cho and Cohen (2010) considered 10 % and 30 % of the total items to exhibit DIF. Samuelsen (2008), on the other hand, considered 10 %, 30 %, and 50 %. Given the findings of Cho and Cohen, the presence of more DIF items improves the accuracy of latent class and parameter recovery. But DeMars and Lau (2011) cautioned that "If a large percentage of the items show DIF, then the test would seem to be measuring disparate constructs in the two groups, making it nonsensical to speak of matching examinees on ability or trait" (p. 601). Given these arguments, 30 % of the items were induced to exhibit DIF in the first simulation study, with $G = 2$ classes. For purposes of comparison, additional simulations with 15 % of the items exhibiting DIF were carried out only for those factors for which the model performed well with 30 % DIF items. In this regard, persons in the first latent class were assigned items with easiness values simulated from a standard normal distribution, $\upsilon \sim N(0,1)$, whereas persons in the second latent class were assigned the same item set, but with the last 30 % (or 15 % for the additional simulations) of the total item set containing DIF. To induce DIF, we added some random noise generated from $N(0,1)$ on the easiness values of these items. The resulting values were then divided by $\sqrt{2}$, such that the final set of I items had zero mean and unity variance.

Number of items (I)

Most simulations in mixture IRT studies use relatively small item samples, such as 20 items (Bilir, 2009; Samuelsen, 2008), 30 items (Dai & Mislevy, 2009; Li et al., 2009), and 40 items (Cho & Cohen, 2010). In this study, we considered three larger, and therefore more realistic, item sample sizes of $I = 60$, 120, and 240. In addition, we allowed the items to be randomly distributed in five study sessions for each person with equal probabilities. This was partly intended to mimic situations in e-learning environments, where users have freedom to independently engage any of the presented items in study sessions of choice. By considering the three item samples, we investigated the performance of the proposed model when the numbers of items in each session k are few (I_k=12), moderate (I_k=24), and large (I_k=48).

Number of persons (P)

Mixture IRT models tend to fail with small sample sizes. The results from the study of Frederickx et al. (2010), who compared samples of 500 and 1,000 persons in a mixture IRT model with random items, suggested that the average misclassification rate was least in the larger person samples. In this study, we considered two moderate sample sizes of P=350 and P=700 persons, mainly due to estimation time limitations.

Mixing proportions (φ_g)

In the study of Frederickx et al. (2010), each class was simulated to consist of 50 % of persons. In this study, two levels of mixing proportions were studied, φ_1=φ_2=.5 and φ_1=.3 versus φ_2=.7, in order to investigate performance of the model given balanced and unbalanced designs. These proportions are comparable to those of other studies in which different proportions were compared (see, e.g., Dai & Mislevy, 2009).

Distribution of initial ability (ω_{0pg})

The initial ability for each latent class was simulated from a normal distribution with different means and unity variance. Our major interest in this simulation study was to evaluate the performance of the proposed model given different standardized mean sizes between the latent classes. Initial investigations with conditions of smaller standardized mean differences between the two latent classes showed that the model failed to efficiently separate the latent classes. In this regard, we considered two cases, namely

1. Medium difference: $\omega_{0p1} \sim N(-.375,1)$ and $\omega_{0p2} \sim N(.375,1)$, and
2. Large difference: $\omega_{0p1} \sim N(-.75,1)$ and $\omega_{0p2} \sim N(.75,1)$.

The standardized mean difference, computed as $(\mu_2-\mu_1)/\sigma$, was therefore equal to 0.75 or 1.5, for the medium or large size, respectively.

Distribution of growth slopes (ω_{1pg})

As in the case of the initial ability parameters, slopes were simulated from a normal distribution, and two cases of medium and large standardized mean differences between latent classes were considered, namely

1. Medium difference: $\omega_{1p1} \sim N(-.0375,.01)$ and $\omega_{1p2} \sim N(.0375,.01)$,and
2. Large difference: $\omega_{1p1} \sim N(-.075,.01)$ and $\omega_{1p2} \sim N(.075,.01)$.

We thus considered four distributional design combinations for the two groups, namely (1) MM, with medium

standardized mean differences between the initial ability and slope distributions, respectively; (2) ML, with medium and large standardized mean differences between the initial ability and slope distributions, respectively; (3) LM, with large and medium standardized mean differences between the initial ability and slope distributions, respectively; and (4) LL, with large standardized mean differences between the initial ability and slope distributions, respectively.

Analysis procedure

By varying the five design factors, 48 conditions were created in total: 3 (item sample sizes) × 2 (person sample sizes) × 2 (mixing proportions) × 2 (initial ability distributions) × 2 (slope distributions). For each considered condition, five data sets were simulated in the R software, resulting in a total of 240 data sets. Each of the simulated data sets, consisting only of the session indicators and the binary scores, was then exported to WinBUGS for the analysis (see the macro shown in the Appendix).

For all conditions in this study, a burn-in of 3,000 iterations was used, followed by 7,000 post-burn-in iterations that were summarized to obtain the parameter estimates of interest. Note, in this respect, that the 10,000 iterations for one data set of 350 persons take approximately four and a half hours to run when using one chain on a computer with an Intel Core i5 CPU, 2.67 GHz, and 4 GB RAM. We were interested in the recovery of population parameters for the class-specific means, variances, and mixing proportions used in the data simulation. This was assessed by comparing the summaries of posterior estimates with the population values used to generate the data. That is, for the fixed effects' parameters, the averages of the 7,000 posterior estimates summarized over the five replicated data sets were compared to the values used in data generation. The respective standard deviations of these posterior estimates were also obtained in order to evaluate how reliable the averages were. For the variance components, the median values of the posterior estimates were considered, because variance was bounded above zero, and therefore was positively skewed. Also reported are the misclassification rates, averaged over the five replications for each condition.

Results

Model evaluation

Monitoring convergence Convergence of the MCMC chains was monitored by examining the time series, kernel density, and autocorrelation plots. In the early steps of MCMC sampling within WinBUGS, the values of the parameters sampled are dependent on the values of the parameters from the previous steps. This dependence is expected to reduce exponentially as the sampling continues and to reach a stationary state that signifies convergence (Bilir, 2009; Lunn et al., 2000). Autocorrelation plots, for instance, provide evidence of how much the consecutive samples are correlated, and these are expected to approach the value of zero for all chains as a sign of achieving convergence. This is portrayed by the sample plots for the different parameters of interest on one simulated data sample in Fig. 1. Convergence can also be monitored from the kernel density plots, which are the posterior density plots for the parameters of interest, and a smooth shape for these plots is an indicator of convergence (Bilir, 2009; Lunn et al., 2000). In the case of nonconvergence, the kernel density plots of some of the parameters are expected to be multimodal and lumpy distributions. As such, the density plots in Fig. 2 provide evidence that all parameters have converged for this specific data set. A few exceptions were noted in the intercepts (b_{01} and b_{02}), most especially in those conditions with smaller item and person samples, as well as in conditions in which the standardized mean differences between the two latent classes were moderate.

Detection of label switching Two types of label switching were investigated in this study. The first one (denoted Type I) occurs within a single MCMC chain, when latent classes change meaning over the estimation chain and can be detected by the presence of distinct jumps in the time-series plots or multimodal kernel density plots for the MCMC iterations (Stephens, 2000). Figure 2 shows a sample of kernel density plots that do not give evidence for Type I label switching, and this was the case for all of the data sets in the considered conditions. The second type of label switching (denoted Type II) that we looked out for occurs when the latent classes switch among the replicated data sets within the same condition. For example, in the five replicated data sets of a certain condition, persons initially generated within class $g=1$ may be consistently assigned to class $g=2$ during the MCMC runs for the first three of the five data sets, whereas persons in the other two data sets may be correctly recovered within class $g=1$, which is the class that they were initially generated in. This can be detected by comparing the latent class-specific parameter estimates with those used in the data generation, and labels to be applied to each of the latent classes can be determined (Cho & Cohen, 2010). Type II label switching is common in simulation studies and was noticed in this study. Furthermore, this type of label switching was commonly observed among smaller person and item sample sizes, and also in the unbalanced conditions. In this study, latent class-specific parameter estimates for all of the data sets in all conditions were compared to those used in data generation, and those parameters whose labels had switched were determined and corrected.

Misclassification proportions Table 1 gives an overview of the average misclassification proportions for the different

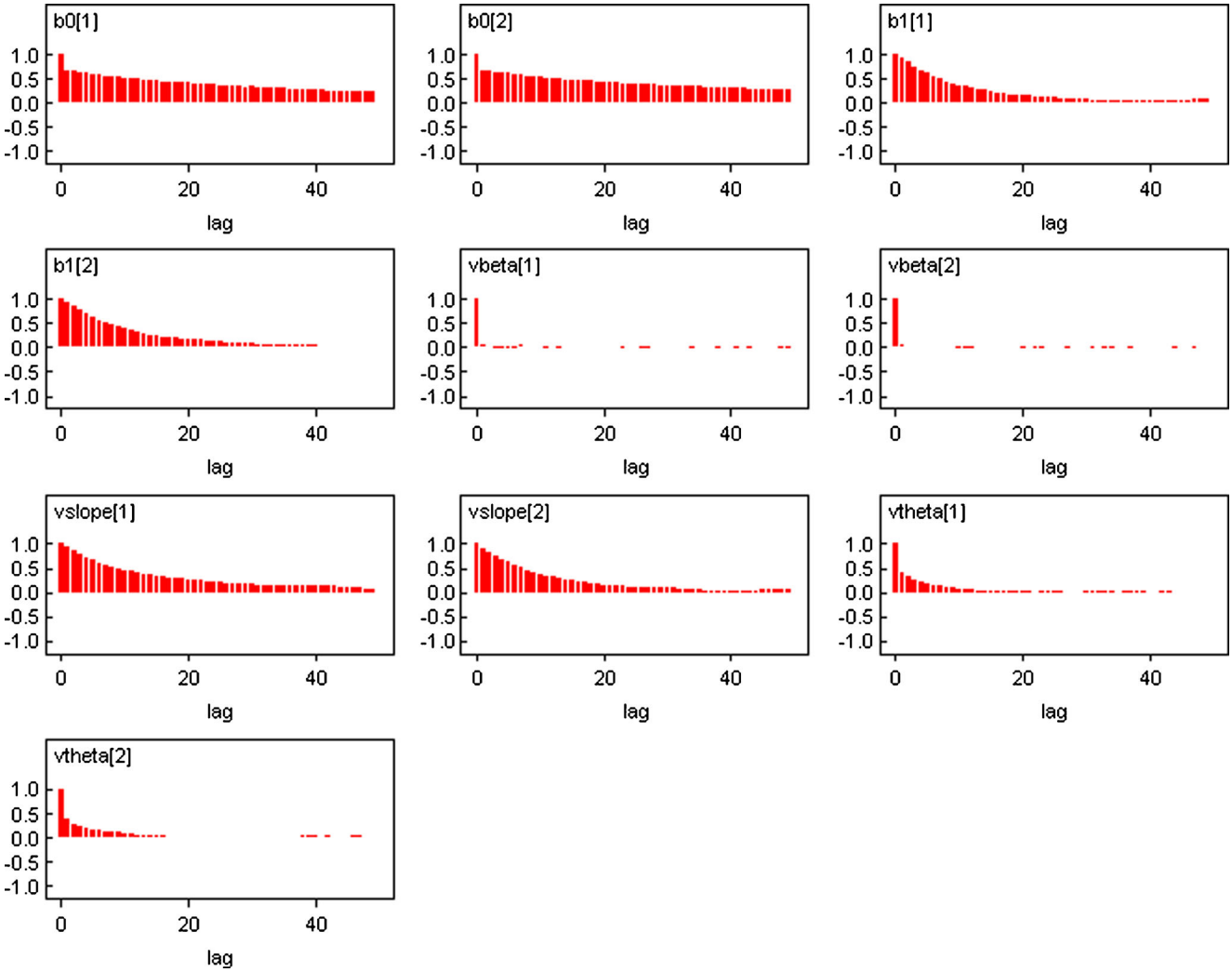

Fig. 1 Autocorrelation plots for the balanced design of one data set in the LL condition with 700 persons and 240 items

design combinations considered in this study. These are the proportions of individuals who were wrongly classified in a latent class that they did not belong to a priori, averaged over the five replicated data sets per condition. One can observe that all conditions with smaller item samples resulted in larger misclassification rates. In both the balanced and unbalanced designs, the misclassification rates reduced with an increase in the number of items, whether with small or large person samples. As compared to the balanced data samples, it can be noted that unbalanced samples tended to result in larger misclassification proportions. A similar observation is also made for conditions comprising latent classes that are closer to each other—that is, when the standardized mean differences between the latent classes are moderate. With larger item samples set at $I = 240$, however, misclassification rates tend to be comparable for the different designs—more specifically, for the large person samples set at $P = 700$. Therefore, the proposed generalized mixture longitudinal IRT model seems to perform quite well in allocating persons to their rightful latent classes only for the large item sample sizes, for balanced designs, and/or for samples with large standardized mean differences between the latent classes.

Parameter recovery

To determine the performance of the proposed generalized mixture longitudinal IRT model, recovery of the generating parameters was investigated. In this respect, recovery was assessed by comparing the mixing proportions, fixed effects, and population parameters of the random effects with the generating values.

Mixing proportions In Table 2, the average proportions of Class 1 membership by sample sizes for the different conditions are given. Proportions for Class 2 membership can be obtained by subtracting the Class 1 values from unity. Note that for the balanced conditions, the values recovered for group membership in the person-level mixtures were

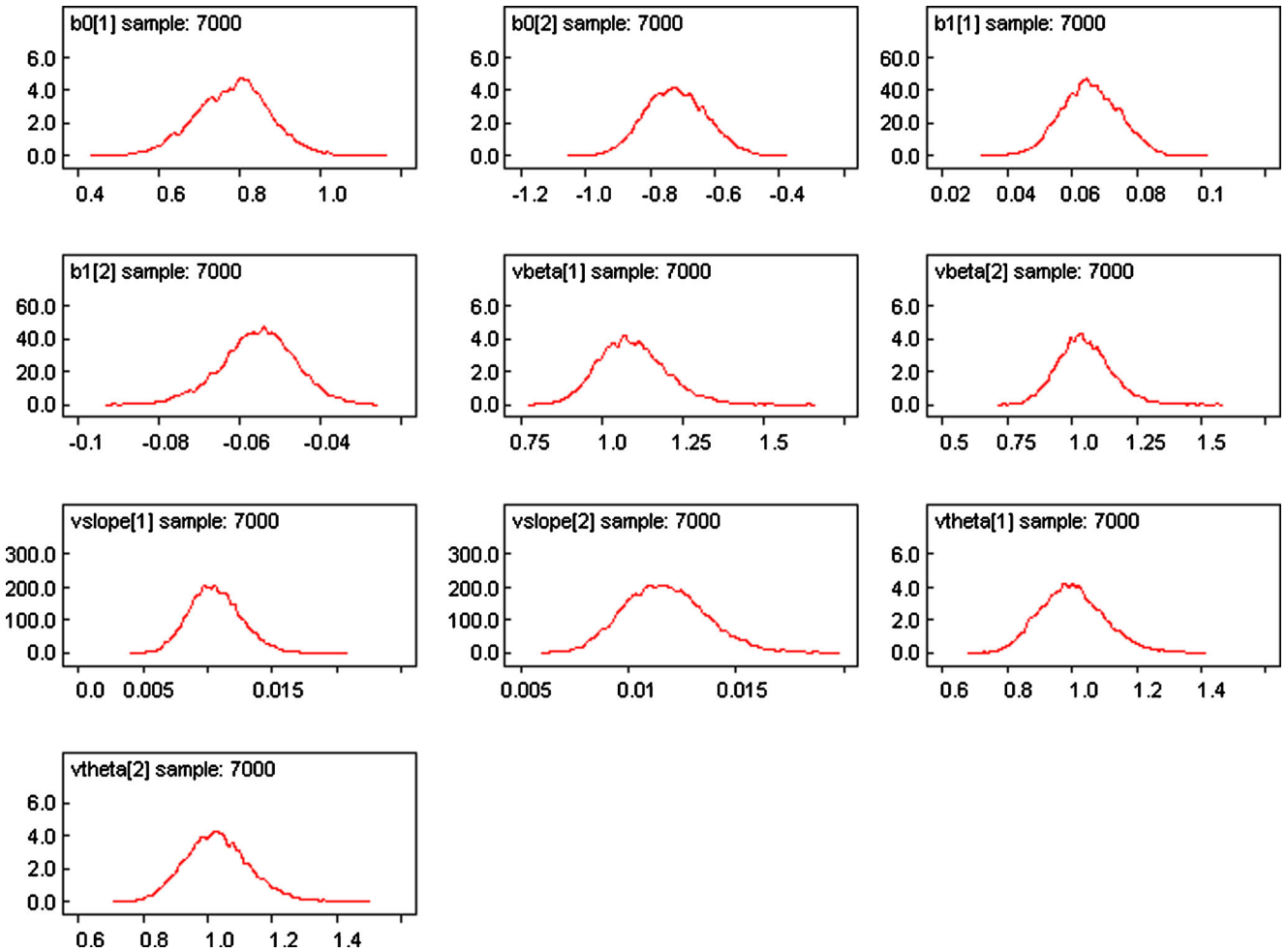

Fig. 2 Kernel density plots for the balanced design of one data set in the LL condition with 700 persons and 240 items

comparable to the generating values, for all item and person sample sizes (Table 2). Due to the discussed misclassification rates in Table 1, however, it is important to note that these groups can contain a sizeable number of the wrong people. In the unbalanced conditions, however, the mixing proportions were greatly biased, with the recovered proportions tending toward those from the balanced designs—more specifically, in the conditions with 60 and 120 items. This was noticed in all conditions, with both small and large person sample sizes, though we noticed a slight improvement for conditions having large standardized mean differences between the latent classes. For conditions with 240 items, the mixing proportions were much less biased.

Table 1 Average misclassification proportions, by sample designs and sizes

Design	$P = 350$			$P = 700$		
	$I = 60$	$I = 120$	$I = 240$	$I = 60$	$I = 120$	$I = 240$
Balanced samples						
MM	.33	.22	.14	.29	.17	.10
ML	.25	.18	.08	.23	.17	.09
LM	.25	.15	.08	.17	.11	.09
LL	.19	.13	.09	.16	.11	.06
Unbalanced samples						
MM	.38	.26	.16	.27	.27	.11
ML	.40	.27	.11	.29	.18	.12
LM	.34	.21	.12	.25	.21	.12
LL	.25	.17	.09	.19	.16	.10

Fixed-effect estimates In Tables 3 and 4, the recovery of the fixed effects' parameters computed from the estimated posterior means of the parameters of interest across all data sets can be assessed by comparing the mean estimates with the population values used to generate the data. Note that most of the parameters for balanced sample designs shown in Table 3 are recovered rather well, because the estimated means are close to the population values for almost all of the considered conditions. For the unbalanced designs in Table 4, on the other hand, most conditions have greatly flawed estimates—notably, for latent Class 1 with 30 % of the persons—which could be expected on the basis of the misclassification

Table 2 Average membership proportions for the first class, by sample designs and sizes

Design	P = 350			P = 700		
	I = 60	I = 120	I = 240	I = 60	I = 120	I = 240
Balanced samples (φ_1=.5)						
MM	.50	.49	.51	.48	.50	.51
ML	.50	.48	.49	.49	.51	.49
LM	.51	.48	.51	.50	.50	.50
LL	.50	.49	.48	.51	.50	.49
Unbalanced samples (φ_1=.3)						
MM	.51	.46	.41	.47	.47	.37
ML	.52	.46	.34	.47	.43	.37
LM	.49	45	.36	.46	.45	.39
LL	.47	.43	.35	.44	.43	.37

proportions discussed above. This observation is especially prevalent in all of the smaller sample size conditions, with the larger ones showing a slight improvement. All of the conditions in Tables 3 and 4 with 240 items, however, seem to result in estimates that are generally comparable to those used in data generation. This implies that the proposed model performs better in balanced than in unbalanced designs. In the case of the latter designs, the sample sizes (more especially, the number of items), as well as the standardized mean differences between the latent classes, need to be large. Note also that the mean standard deviations of the estimates are generally large in both the balanced and unbalanced conditions with 60 items, but more in the MM conditions. This variability in estimates is partly due to the small number of items used to estimate session-specific effects.

Random-effect variance estimates In this section, we assess the recovery of the random effects' variance estimates for the initial abilities, the session slopes, and the item difficulties. Samples with 120 and 240 items seemed to result in slightly better estimates that were much closer to the population values than were those based on 60 items. Furthermore, for conditions with 60 items, smaller person sample size conditions seemed to result in variance estimates for initial abilities ($\sigma^2_{0\omega 1}$ and $\sigma^2_{0\omega 2}$) with positive biases larger than those from the large person sample sizes. In general, however, variance estimates were recovered relatively well for all of the conditions, such that the median variance estimates were more or less equal to the population values, and the conclusions drawn from all item sample sizes were not very different. In Table 5, we present the random-effect variance estimates only for samples with 240 items. From this table, it can be observed that both the small and large person sample size conditions have estimates that are not very different from the population values. In the unbalanced conditions, the latent classes with the smaller proportion of persons (30 %) seem to have larger biases for the variance estimates for both initial abilities ($\sigma^2_{0\omega 1}$) and slopes ($\sigma^2_{1\omega 1}$). This is perhaps because, as compared to the large sample sizes, small samples are likely to result in large standard errors of the estimates. For the item random effects, no major differences in the variance estimates seem to be present for all of the conditions. It should be noted, however, that these observations are based on the median values of only five replicated data sets, which is rather a small number. From the present observations, however, the proposed model seems to perform well in recovering the variances of the considered random effects.

Additional simulations

In this section, and given the results of the basic simulation study so far, two extensions of the simulation design were studied. The performance of the proposed mixture IRT model was investigated, first, for a case in which only 15 % of the items exhibited DIF, and second, for G = 3 latent classes. These two conditions were studied for the balanced study design, I = 240 items, and the case of large standardized mean differences between the initial ability and slope distributions. We did this because the model seemed to perform well for these factor levels in the previous simulation study, but more so because, due to computation requirements, it was not feasible to combine all of the previously considered factor levels for the additional simulations. Specific factors for these simulations are given hereunder.

15 % of DIF items

The considered factors were P=700 persons, I=240 items, G=2, and the LL design, with both large initial differences in ability for $\omega_{0p1}\sim N(-.75,1)$ and $\omega_{0p2}\sim N(.75,1)$, and large differences in growth slopes for $\omega_{1p1}\sim N(-.075,.01)$ and $\omega_{1p2}\sim N(.075,.01)$. The procedures for data generation and analysis were similar to those followed before.

Three ($\boldsymbol{G}$=3) latent classes

Here, the P=700 persons were distributed as P_1=P_2=233 and P_3=234, with each class engaging all of the I=240 items. Furthermore, the distributions of initial abilities for the classes were taken to be $\omega_{0p1}\sim N(-.75,1)$, $\omega_{0p2}\sim N(.00,1)$, and $\omega_{0p3}\sim N(.75,1)$, whereas those for growth slopes were taken to be $\omega_{1p1}\sim N(-.075,.01)$, $\omega_{1p2}\sim N(.00,.01)$, and $\omega_{1p3}\sim N(.075,.01)$, ensuring that moderate to large standardized differences between the latent classes would be retained. These values imply that the first class of persons started low and on average declined over time, that the second group was neither declining nor improving, and that the ability of the third latent class, which started highest, was on average

Table 3 Mean estimates and their mean standard deviations for the fixed effects' parameters, by designs and sizes for balanced samples

Design	Pop. value	Mean estimates						Mean standard deviations					
		$P = 350$			$P = 700$			$P = 350$			$P = 700$		
		$I = 60$	$I = 120$	$I = 240$	$I = 60$	$I = 120$	$I = 240$	$I = 60$	$I = 120$	$I = 240$	$I = 60$	$I = 120$	$I = 240$
MM													
b_{01}	–.375	–.204	–.325	–.193	–.177	–.292	–.355	.204	.146	.114	.169	.128	.093
b_{02}	.375	.195	.269	.374	.200	.311	.348	.206	.145	.113	.162	.115	.091
b_{11}	–.0375	–.019	–.031	–.038	–.030	–.037	–.035	.027	.018	.013	.019	.012	.009
b_{12}	.0375	.020	.035	.027	.029	.039	.034	.026	.018	.013	.018	.012	.009
ML													
b_{01}	–.375	–.264	–.359	–.389	–.229	–.318	–.416	.194	.144	.111	.172	.126	.095
b_{02}	.375	.295	.279	.359	.337	.271	.359	.200	.140	.113	.174	.127	.084
b_{11}	–.075	–.066	–.068	–.070	–.025	–.076	–.070	.026	.017	.013	.024	.013	.009
b_{12}	.075	.065	.058	.076	.055	.070	.065	.027	.017	.013	.023	.013	.009
LM													
b_{01}	–.750	–.439	–.676	–.706	–.666	–.710	–.692	.206	.155	.113	.176	.125	.092
b_{02}	.750	.457	.571	.769	.614	.721	.776	.209	.149	.112	.162	.127	.089
b_{11}	–.0375	–.030	–.043	–.035	–.033	–.033	–.037	.025	.017	.013	.019	.012	.009
b_{12}	.0375	.040	.023	.035	.037	.037	.035	.025	.017	.013	.018	.011	.009
LL													
b_{01}	–.750	–.511	–.669	–.659	–.682	–.770	–.802	.221	.159	.124	.155	.120	.092
b_{02}	.750	.563	.607	.683	.649	.705	.688	.219	.148	.117	.155	.115	.089
b_{11}	–.075	–.060	–.069	–.064	–.066	–.071	–.070	.029	.019	.014	.017	.012	.009
b_{12}	.075	.057	.066	.068	.076	.069	.074	.029	.018	.014	.017	.012	.009

improving over time. DIF was induced in the last 15 % of the items only for classes $g = 2$ and $g = 3$, following procedures similar to those used before.

For $G = 2$ latent classes and 15 % of the items in the second class having DIF, the average misclassification proportion was 11 %, which is about twice as large as that observed within similar conditions but at 30 % DIF. This implies that the presence of fewer DIF items (15 % as opposed to 30 %) reduces the accuracy of latent class separation resulting in increased misclassification rates. For parameter recovery, the average mixing proportion was .50, whereas the mean estimates for the fixed effects' parameters and the median estimates for the random effects' parameters are shown in Table 6. These estimates are comparable to the population values and are generally not different from those obtained under similar conditions but with 30 % DIF—that is, for the LL and $I = 240$ conditions in Tables 3 and 5.

For $G = 3$ latent classes, the recovered average mixing proportions of .32, .34, and .34 for Classes 1, 2, and 3, respectively, were generally balanced and comparable to the population proportions. The average misclassification rate of persons between classes $g = 1$ and $g = 2$ was 24 %, that between classes $g = 1$ and $g = 3$ was 11 %, and that between classes $g = 2$ and $g = 3$ was 13 %. As such, we found a noticeable general increase in misclassification rates for this condition, perhaps due to the low percentage of DIF items set at 15 %. Furthermore, the large misclassification rates, especially between classes $g = 1$ and $g = 2$ and between classes $g = 2$ and $g = 3$, may be attributed to the reduced standardized mean difference between initial abilities and slope distributions between each pair of the two sets of classes. Nonetheless, the recovered mean estimates for the fixed effects' parameters and the median estimates for the random effects' parameters shown in Table 6 are generally comparable to the population values. As such, the proposed model seems to perform fairly well in recovering the means of the fixed effects and the variances of the considered random effects, even for $G = 3$ latent classes, and on the basis of the results of the first simulation study, these estimates would be expected to improve if more than 15 % of the items contained DIF.

Example

An empirical sample data set from the Maths Garden e-learning environment (Klinkenberg et al., 2011) was analyzed using the longitudinal IRT model in Eq. 3, which assumes no latent classes, and the results were compared to those obtained from the longitudinal mixture IRT model in Eq. 5. The Maths Garden is a Web-based computer adaptive practice and monitoring system used in The Netherlands to provide more time

Table 4 Mean estimates and their mean standard deviations for the fixed effects' parameters, by designs and sizes for unbalanced samples

Design	Pop. value	Mean estimates						Mean standard deviations					
		$P = 350$			$P = 700$			$P = 350$			$P = 700$		
		$I = 60$	$I = 120$	$I = 240$	$I = 60$	$I = 120$	$I = 240$	$I = 60$	$I = 120$	$I = 240$	$I = 60$	$I = 120$	$I = 240$
MM													
b_{01}	–.375	.093	–.115	–.162	–.077	–.077	–.211	.211	.158	.135	.166	.129	.099
b_{02}	.375	.300	.409	.341	.424	.423	.424	.215	.140	.112	.161	.127	.088
b_{11}	–.0375	.007	–.007	–.020	–.020	.003	–.020	.028	.020	.014	.019	.013	.010
b_{12}	.0375	.029	.037	.043	.041	.025	.043	.029	.018	.012	.018	.013	.008
ML													
b_{01}	–.375	.040	–.014	–.264	–.098	–.274	–.185	.204	.155	.133	.175	.119	.106
b_{02}	.375	.289	.361	.401	.292	.400	.342	.207	.148	.104	.161	.114	.086
b_{11}	–.075	–.001	–.021	–.067	–.026	–.037	–.051	.034	.020	.017	.020	.013	.011
b_{12}	.075	.047	.069	.072	.081	.087	.073	.031	.017	.008	.018	.011	.009
LM													
b_{01}	–.750	.138	–.310	–.389	–.168	–.202	–.425	.226	.164	.153	.182	.141	.116
b_{02}	.750	.511	.761	.803	.735	.785	.827	.222	.152	.105	.177	.134	.091
b_{11}	–.0375	–.003	–.011	–.027	–.007	–.017	–.024	.029	.019	.016	.019	.013	.011
b_{12}	.0375	.049	.049	.042	.030	.044	.041	.029	.017	.009	.019	.012	.008
LL													
b_{01}	–.750	–.240	–.265	–.562	–.248	–.298	–.506	.272	.169	.153	.169	.137	.117
b_{02}	.750	.704	.741	.761	.783	.829	.729	.277	.136	.104	.164	.124	.085
b_{11}	–.075	–.017	–.047	–.056	–.037	–.041	–.048	.032	.020	.016	.019	.014	.012
b_{12}	.075	.070	.092	.075	.089	.077	.082	.030	.016	.011	.017	.012	.008

Table 5 Median variance estimates for the random effects' parameters, by designs, for samples with 240 items

Design	Sample size	Persons				Items	
		$\sigma^2_{0\omega 1}$	$\sigma^2_{1\omega 2}$	$\sigma^2_{1\omega 1}$	$\sigma^2_{1\omega 2}$	σ^2_{v1}	σ^2_{v2}
Pop. value		1.00	1.00	0.01	0.01	1.00	1.00
Balanced samples							
MM	350	1.097	1.091	0.011	0.010	1.052	1.069
ML	350	1.110	1.061	0.011	0.012	1.026	1.134
LM	350	1.016	1.028	0.013	0.010	1.058	1.091
LL	350	1.129	1.009	0.011	0.011	0.962	1.032
MM	700	0.996	1.053	0.011	0.009	1.014	1.019
ML	700	1.023	0.980	0.010	0.011	1.054	1.094
LM	700	1.064	1.049	0.011	0.010	1.076	1.011
LL	700	1.035	1.030	0.010	0.011	1.002	0.991
Unbalanced samples							
MM	350	1.284	1.135	0.013	0.008	0.955	1.007
ML	350	1.076	1.004	0.010	0.009	1.081	1.128
LM	350	1.299	1.142	0.014	0.010	0.939	0.958
LL	350	1.399	1.053	0.008	0.011	0.981	1.024
MM	700	1.086	1.015	0.011	0.009	0.955	0.989
ML	700	1.213	1.021	0.014	0.010	0.990	1.005
LM	700	1.499	1.013	0.011	0.010	0.934	0.948
LL	700	1.197	0.988	0.014	0.010	0.908	0.981

Table 6 Mean estimates of the fixed effects' parameters, with their mean standard deviations, and median estimates of the random effects' parameters for samples with 15 % differential item functioning

Parameter	Popn. value	Estimate	*SD*
$G = 2$			
Fixed effects			
b_{01}	–0.750	–0.679	0.095
b_{02}	0.750	0.727	0.097
b_{11}	–0.075	–0.069	0.009
b_{12}	0.075	0.067	0.010
Random effects			
Persons			
$\sigma^2_{0\omega 1}$	1.000	1.094	
$\sigma^2_{0\omega 2}$	1.000	1.099	
$\sigma^2_{1\omega 1}$	0.010	0.010	
$\sigma^2_{1\omega 2}$	0.010	0.012	
Items			
σ^2_{v1}	1.000	1.085	
σ^2_{v2}	1.000	1.056	
$G = 3$			
Fixed effects			
b_{01}	–0.750	–0.467	0.148
b_{02}	0.000	–0.073	0.136
b_{03}	0.750	0.633	0.126
b_{11}	–0.075	–0.053	0.016
b_{12}	0.000	–0.017	0.013
b_{13}	0.075	0.066	0.013
Random effects			
Persons			
$\sigma^2_{0\omega 1}$	1.000	1.360	
$\sigma^2_{0\omega 2}$	1.000	1.166	
$\sigma^2_{0\omega 3}$	1.000	1.109	
$\sigma^2_{1\omega 1}$	0.010	0.011	
$\sigma^2_{1\omega 1}$	0.010	0.013	
$\sigma^2_{1\omega 3}$	0.010	0.011	
Items			
σ^2_{v1}	1.000	1.075	
σ^2_{v2}	1.000	1.063	
σ^2_{v3}	1.000	1.080	

to practice and maintain basic arithmetic skills in order to improve the ability of individual students. Persons can regularly engage specific exercise items that are adjusted to their individual ability levels using an IRT model, and it includes immediate feedback to persons on the exercises that they engage in, thereby ensuring that learning takes place. According to Klinkenberg et al., the Maths Garden was specially developed to improve mathematics education in primary schools by combining arithmetic practice and measurement in a playful manner using computerized educational games. The data set that was used in Klinkenberg et al.'s article consisted of data collected in the 2008/2009 academic year from pupils from across The Netherlands, within the age range of 4–12 years, enrolled in kindergarten and grades 1–6 of primary school. In the present article, however, the analyzed data set, for the 2010/2011 academic year, consisted of the item responses of participants 4–18 years of age, with additional, older participants enrolled in the first four years of secondary school education. For a detailed design and implementation methodology of the Maths Garden e-learning environment, we refer the interested reader to Klinkenberg et al.

The four domains of addition, subtraction, multiplication, and division are administered within the Maths Garden environment, but only the first domain—addition—was analyzed for this article. The data set consisted of response scores to 610 items by 3,803 persons, but the analysis in this article focuses on the 1,335 individuals who engaged in at least 50 exercise items and on the 290 items that were answered by at least 150 distinct persons. The highest number of engaged items by an individual was 211, with a median and mean of 69 and 76 exercise items, respectively. For each item engaged by an individual, a time-stamp is recorded automatically by the environment. We used this variable—the time at which distinct items were answered by a person—to compute study sessions for the individuals, and one calendar day was used as the minimum spacing time between two consecutive study sessions for an individual (see, e.g., Cepeda et al., 2008; Vlach & Sandhofer, 2012). For the analyzed data sample, only 15.8 % of the considered persons had up to ten study sessions (highest number), whereas 54.5 % had up to six study sessions (average number). The response variable was the first attempt score on an item by a person, with 1 = pass and 0 = fail.

For this example, Eq. 5 with $G = 2$ latent classes was compared to Eq. 3 (with $G = 1$), simply for illustration purposes. In normal mixture model practice, however, it is important to compare the results of several values of G and then to select the best-fitting model—for instance, by using information criteria. Moreover, we observed that the MCMC chains for all of the model parameters with $G = 2$ were not very stable until up to the 3,000th iteration. Perhaps this was due to the sparse nature of this data set (i.e., not all persons answered all of the 290 items), further limiting us to investigate more latent classes. As such, a burn-in of 5,000 iterations was used, followed by 10,000 post-burn-in iterations that were summarized in order to obtain the parameter estimates of interest. The averages of these 10,000 posterior estimates with their respective standard deviations for the fixed effects' parameters, as well as the median values of the posterior estimates for the variance components, are given in Table 7. From this table, we can observe that the mixture IRT Model 5 is more parsimonious when compared to the general Model 3, because it shows smaller AIC and BIC values. From Model 3, we get the average initial ability level and the ability growth for the whole population of persons. For instance, we

observed that persons had an average probability of success of .73 on an average item before use of the e-learning environment. This is the antilogit of the intercept value—that is, $\text{logit}^{-1}(0.982) = .73$. Model 3 further shows that, on average, all persons' ability to succeed on an average item increased over study sessions, due to the positive slope estimates. From Model 5, on the other hand, we observed that persons in latent Class 1 had, on average, a lower probability of success of .68, as compared to the value of .76 for the persons in the second latent class. For the sessions' slopes, it appears that on average the probabilities of success over time for persons in Class 1 were about twice as high as those of their counterparts in Class 2. A closer look at the average logit estimates for the slopes of Class 2 indicates that persons in this latent class were not learning over time. This is because the standard deviation of the slope average for Class 2 is about twice as big as its estimate, an indication that this estimate is less reliable. We furthermore observed more variability in the logits of success for persons in latent Class 1, for both the intercepts and slopes, than was the case for persons in latent Class 2. Additionally, the items engaged in by persons in Class 1 seem to have been more similar (i.e., to have had less variability) than those engaged in by persons in Class 2. As such, e-learning managers can utilize extra information from Model 5 to make necessary educational interventions for persons in specific latent classes, which is not the case with the overall population-averaged estimates given by Model 3.

Table 7 Mean estimates of the fixed effects' parameters, with their mean standard deviations, and median estimates of the random effects' parameters of the item response theory models for the Maths Garden environment

Parameter	Model 3		Model 5	
	Estimate	*SD*	Estimate	*SD*
Fixed effects				
Intercept	0.982	0.040		
Class 1			0.746	0.072
Class 2			1.157	0.063
Session slopes	0.015	0.005		
Class 1			0.024	0.021
Class 2			0.012	0.021
Random effects				
Persons				
Intercepts	0.395			
Class 1			0.410	
Class 2			0.335	
Session slopes	0.299			
Class 1			0.630	
Class 2			0.064	
Items	0.315			
Class 1			0.419	
Class 2			0.289	
Fit statistics				
AIC	108,500		108,100	
BIC	108,500		108,200	

Discussion, limitations, and directions for further research

In this article, a generalized mixture longitudinal item response theory (IRT) model was described for modeling latent group differences in growth for item response data from learning environments. The model described in this study utilizes features of the longitudinal or multidimensional Rasch model (Andersen, 1985; Embretson, 1991), the random-item-response model (De Boeck, 2008; Van Den Noortgate et al., 2003), and the mixture Rasch model (Rost, 1990). The proposed model provides possibilities of estimating latent group-specific growth trends, relative to longitudinal IRT growth models that consist of no person groups. Moreover, defining the difficulty of the items as random (besides the common random person abilities) is an efficient and realistic way to model the variation in item difficulty (Frederickx et al., 2010), which is quite appealing for learning environments that tend to include a large item bank—for instance, Web-based electronic learning environments.

A unique feature of the presented model is that it has mixtures of latent classes of persons that differ from each other in two different aspects, which leads to differences in the difficulty levels of the items. First, it assumes that response patterns may be heterogeneous over classes, due to initial person-level differences (before learning takes place), and second, that growth trends may be heterogeneous over time, with persons in one class tending to have, on average, higher growth trends than those in other classes. The model, however, does not exclude the possibility that there may be no initial differences between latent person classes, but only class-specific growth trend differences, and vice versa. As such, the presented model provides an opportunity to determine whether latent classes exist that differ in their item response patterns as a result of existing initial latent differences within the population, or in case these differences are a result of learning within web-based e-learning environments or a combination of both situations.

A Bayesian solution was described for estimation of the model parameters, as implemented in WinBUGS—a freely available software program—and a simulation study was presented to investigate the performance of the proposed model under some practical conditions. In this respect, the generating parameters were generally recovered well for most of the considered conditions—more specifically, for large item sample sizes, balanced designs, and conditions in which standardized mean differences between latent classes were considered large. In these cases, the proposed generalized mixture longitudinal IRT model seemed to give clear-cut classifications of persons into their respective predetermined latent classes, with minimal

misclassification proportions. Although the results from the utilized simulation study are specific to the considered conditions with 350 and 700 persons, comparison of the results from the two cases (as well as other preliminary simulations not reported here) provide evidence that the larger the number of items, the more likely that the model will provide lower misclassification rates and more reliable parameter estimates, whether with two or three latent classes. Furthermore, the obtained simulation study results appear to support previous suggestions regarding potential problems of mixture IRT models with small sample sizes of less than 380 persons (cf. Bilir, 2009; Cho, Cohen, & Kim, 2006). In this study, for most conditions with samples of 350 persons, higher misclassification rates and poorly recovered model parameters were noted, especially within the unbalanced designs. The proposed longitudinal mixture IRT model was further illustrated with an empirical example data set from a Web-based e-learning environment—the Maths Garden (Klinkenberg et al., 2011). The results that we obtained showed that the proposed mixture IRT model is a parsimonious approach and provides additional latent information when compared to the results obtained from a longitudinal IRT model that assumes no latent classes.

Several limitations of the proposed model have been noted. For instance, the model seems to be unreliable within conditions of unbalanced sample sizes or when the standardized mean differences between latent classes are not large, and more so for most conditions with item samples of 60 and 120. In these cases, the model did not even recover the generating parameters very well for the conditions of large person sample sizes, with large standardized mean differences of 1.5 between latent classes for initial abilities as well as slopes (though slightly improved). An additional drawback that is common with mixture random-item IRT models such as the one considered in the present study is that they cannot be run directly in all common statistical software packages, but require flexible software such as WinBUGS. Frederickx et al. (2010) noted that such a limitation can hinder the applicability of mixture IRT models, because one needs to be familiar with Bayesian statistics, including MCMC computation techniques, in order to utilize the model. Moreover, the inference processes of mixture IRT models using MCMC typically require substantial computing time in order to obtain usable results (Cho & Cohen, 2010; von Davier & Yamamoto, 2007). Further research could be aimed at developing or testing alternative methods for fitting the proposed generalized mixture longitudinal IRT model in an efficient and quick way—such as by using the gllamm package in the STATA statistical software (Rabe-Hesketh, Skrondal, & Pickles, 2004b). Furthermore, the considered assumption of a linear change of the latent trait over time may lead to model misspecification and/or to overextraction of classes for the actual data analyses. The proposed model, however, can easily be extended, and model fit indices can be used to choose the best-fitting model, after accounting for the complexity of the models. It is expected that more complex models (such as quadratic models) may require even more items and/or persons to yield unbiased and precise estimates, but an evaluation of these models was outside the scope of the present article. Additionally, the provided simulation study ignored the problem of the accuracy of estimations of the number of latent classes, because we did not fit models with different numbers of classes, followed by a model selection procedure to pick the most appropriate number. A further evaluation of the proposed modeling approach with respect to the quality of parameter estimates, on the one hand, and correct model selection, on the other, will be important.

Other possible extensions of the proposed model exist. In this study, for instance, two and three latent person classes were studied, but more can be introduced in the model by specifying a different value for G in the WinBUGS macro (see the Appendix). It is also possible to study the presence and magnitude of differential functioning items (DIF). This may be achieved, for instance, by extracting class-specific posterior means of the item difficulty estimates and studying them further. It would also be interesting to extend the proposed model by including manifest person (and perhaps item) properties as predictors, as in the explanatory IRT modeling framework (De Boeck & Wilson, 2004). For instance, the utilized Maths Garden data set comprised some manifest person properties, including gender, school, grade, and age, but these were not modeled because manifest explanatory modeling was not the focus of the present article. Bilir (2009), however, noted that the inclusion of both manifest and latent groups/properties can provide a good comparison on the basis of which to explain or understand the underlying causes of DIF—whether it is due to the presence of manifest groups, latent classes, or a combination of both. In the spirit of this article, inclusion of such properties might provide useful information, in so far as the classification of growth trends is concerned, whether the differences in growth trends are due to manifest and/or to latent groups. Furthermore, the proposed model is limited in that it is based on the Rasch model, and discrimination parameters were not incorporated. However, actual correlations among the items could be made to differ by making the Rasch model not fit the actual data well. Future research that takes these issues into account will greatly improve the proposed model.

Author Note Kind acknowledgments to Han L. J. van der Maas, University of Amsterdam, and to Oefenweb.nl for providing the data set from the Maths Garden learning environment. We also thank two anonymous reviewers and Editor in Chief Gregory Francis for their insightful comments and suggestions.

Appendix: WinBUGS macro

```
############################################################
### Generalized Longitudinal Mixture IRT Model: Equation 4 ##
############################################################
# J: the number of students
# I: the number of items
# beta: item difficulty
# theta: initial ability
# etaw: slopes
# taup: Precision of initial ability estimates (deviations from b0)
# taui: Precision of item difficulty estimates
# taus: Precision of slope estimates (deviations from b1)
# class: categorical latent class indicator
# G: number of latent classes
# b0: Class-specific mean intercept
# b1: Class-specific mean slope of sessions
# pi: probability of class membership at person level
# ses: Study sessions indicators

model
{
#Read Response Data
for (j in 1:J) {
 for (i in 1:I) {
 r[j,i]<-resp[j,i]
 }}
# IRT model
for (j in 1:J) {
 for (i in 1:I) {
 prop[j,i]<- exp(b1[class[j]]*ses[j,i]+ taus[class[j]]*etaw[j]*ses[j,i]+
theta[j]+ beta[i,class[j]])
 p[j,i]<- prop[j,i]/(1 + prop[j,i])
 r[j,i]~dbern(p[j,i])
  }
#Persons
 theta[j] ~ dnorm(b0[class[j]],taup[class[j]])
 class[j] ~ dcat(pi[j,1:G])
 etaw[j] ~ dnorm(0,1)
       }
#Items
for (i in 1:I){
  for (g in 1:G){
  beta[i,g]~dnorm(0,taui[g])
  }}
#Priors
#Probabilities of group membership
for (j in 1:J){
pi[j,1:G]~ddirch(alpha[])
       }
for (g in 1:G) {
#Class mean
 b0[g]~ dnorm(0,1)
#Slopes
 b1[g]~ dnorm(0,1)
#Ability
       taup[g]~dgamma(2,4)
       vtheta[g] <- 1/taup[g]
#Variance in slopes
       taus[g]~dnorm(0,1)I(0,)
       vslope[g] <- pow(taus[g],2)
#Difficulty in items
       taui[g]~dgamma(2,4)
       vbeta[g] <- 1/taui[g]
                    }
       }
#Data
# load DATA here (in separate file)

list(J=350, I=60, G=2, alpha=c(1,1),
ses=structure(.Data=c(
1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3,3,4,4
,4,4,4,4,4,4,4,4,4,4,5,5,5,5,5,5,5,5,5,5,
……
1,3,2,4,1,1,2,4,5,4,5,3,2,2,5,4,1,2,3,1,4,5,4,2,2,3,5,5,2),.Dim=c(350,60)),

resp=structure(.Data=c(
0,1,0,0,1,1,1,1,0,1,0,0,0,0,1,0,1,0,1,1,0,0,1,1,0,1,1,0,0,0,0,0,1,0,1,1,0,
……
1,1,1,1,1,1,1,1,0,0,1,1,1,0,1,1,1,0,0,1,0,0,1,0,1,1),.Dim=c(350,60)))

# Initials
list(b0=c(0.0,0.0),b1=c(0.0,0.0),taup=c(1,1),taus=c(1,1),taui=c(1,1))
```

References

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50,* 3–16. doi:10.1007/BF02294143

Andrade, D. F., & Tavares, H. R. (2005). Item response theory for longitudinal data: Population parameter estimation. *Journal of Multivariate Analysis, 95,* 1–22. doi:10.1016/j.jmva.2004.07.005

Bilir, M. K. (2009, July 29). *Mixture item response theory-Mimic Model: Simultaneous estimation of differential item functioning for manifest groups and latent classes*. Florida State University. Retrieved from http://diginole.lib.fsu.edu/etd/3761

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics, 26,* 381–409. doi:10.3102/10769986026004381

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning. A temporal ridgeline of optimal retention. *Psychological Science, 19,* 1095–1102. doi:10.1111/j.1467-9280.2008.02209.x

Cho, S.-J., Athay, M., & Preacher, K. J. (2013a). Measuring change for a multidimensional test using a generalized explanatory longitudinal item response model. *British Journal of Mathematical and Statistical Psychology, 66,* 353–381. doi:10.1111/j.2044-8317.2012.02058.x

Cho, S.-J., Bottge, B., Cohen, A. S., & Kim, S.-H. (2011). Detecting cognitive change in the math skills of low-achieving adolescents. *Journal of Special Education, 45,* 67–76.

Cho, S.-J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics, 35,* 336–370. doi:10.3102/1076998609353111

Cho, S.-J., Cohen, A. S., & Bottge, B. (2013b). Detecting intervention effects using a multilevel latent transition analysis with a mixture IRT model. *Psychometrika, 78,* 576–600.

Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2006, June). *An investigation of priors on the probabilities of mixtures in the mixture Rasch model*. Paper presented at the Annual Meeting of the Psychometric Society, Montreal, Canada.

Cho, S.-J., Cohen, A. S., Kim, S.-H., & Bottge, B. (2010). Latent transition analysis with a mixture IRT measurement model. *Applied Psychological Measurement, 34,* 583–604.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42,* 133–148. doi:10.1111/j.1745-3984.2005.00007

Dai, Y., & Mislevy, R. (2009). *A mixture Rasch model with a covariate. A simulation study via Bayesian Markoc Chian Monte Carlo estimation*. University of Maryland. Retrieved from http://drum.lib.umd.edu/handle/1903/9926

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73,* 533–559. doi:10.1007/s11336-008-9092-x

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

DeMars, C. E., & Lau, A. (2011). Differential item functioning detection with latent classes. How accurately can we detect who is responding differentially? *Educational and Psychological Measurement, 71,* 597–616. doi:10.1177/0013164411404221

Desmet, P., Paulussen, H., & Wylin, B. (2006). FRANEL: A public online language learning environment, based on broadcast material. In E. Pearson & P. Bohman (Eds.), *Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications* (pp. 2307–2308). Chesapeake VA: AACE. Retrieved from www.editlib.org/p/23329

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56,* 495–515. doi:10.1007/BF02294487

Finch, W. H., & Pierson, E. E. (2011). A mixture IRT analysis of risky youth behavior. *Frontiers in Quantitative Psychology and Measurement, 2,* 1–10. doi:10.3389/fpsyg.2011.00098

Fischer, G. H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika, 54,* 599–624. doi:10.1007/BF02296399

Fischer, G. H. (1995). Some neglected problems in IRT. *Psychometrika, 60,* 459–487. doi:10.1007/BF02294324

Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: A Random Item Mixture Model to detect differential item functioning. *Journal of Educational Measurement, 47,* 432–457. doi:10.1111/j.1745-3984.2010.00122.x

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38,* 79–93. doi:10.1111/j.1745-3984.2001.tb01117.x

Klinkenberg, S., Straatemeier, M., & Van der Maas, H. L. J. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education, 57,* 1813–1824. doi:10.1016/j.compedu.2011.02.003

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement, 33,* 353–373.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10,* 325–337. doi:10.1023/A:1008929526011

Miceli, R., Settanni, M., & Vidotto, G. (2008). Measuring change in training programs: An empirical illustration. *Psychology Science Quarterly, 50,* 433–447.

Mislevy, R., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55,* 195–215. doi:10.1007/BF02295283

Muthén, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational and Behavioral Statistics, 10,* 133–142. doi:10.3102/10769986010002133

Paek, I., Baek, S.-G., & Wilson, M. (2012). An IRT modeling of change over time for repeated measures item response data using a random weights linear logistic test model approach. *Asia Pacific Education Review, 13,* 487–494. doi:10.1007/s12564-012-9210-4

R Development Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004a). Generalized multilevel structural equation modelling. *Psychometrika, 69,* 167–190.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004b). GLLAMM manual (U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 160). Retrieved from http://biostats.bepress.com/ucbbiostat/paper160

Roberts, J. S., & Ma, Q. (2006). IRT models for the assessment of change across repeated measurements. In R. W. Lissitz (Ed.), *Longitudinal and value added modeling of student performance* (pp. 100–127). Maple Grove, MN: JAM Press.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14,* 271–282. doi:10.1177/014662169001400305

Safer, N., & Fleischman, S. (2005). How student progress monitoring improves instruction. *Educational Leadership, 62,* 81–83.

Samuelsen, K. M. (2008). Examining differential item functioning from a latent mixture perspective. In G. R. Hancock & K. M. Samuelsen

(Eds.), *Advances in latent variable mixture models* (pp. 177–197). Charlotte, NC: Information Age.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B, 62,* 795–809. doi:10.1111/1467-9868.00265

Van Den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics, 30,* 443–464. doi:10.3102/10769986030004443

Van Den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics, 28,* 369–386. doi:10.3102/10769986028004369

Vlach, H. A., & Sandhofer, C. M. (2012). Distributing learning over time: The spacing effect in children's acquisition and generalization of science concepts. *Child Development, 83,* 1137–1144.

von Davier, M., Xu, X., & Carstensen, C. H. (2009). *Using the general diagnostic model to measure learning and change in a longitudinal large-scale assessment* (Technical Report No. ETS RR-09-28). Educational Testing Services. Retrieved from www.ets.org/Media/Research/pdf/RR-09-28.pdf

von Davier, M., & Yamamoto, K. (2007). Mixture-distribution and HYBRID Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (Statistics for Social and Behavioral Sciences, pp. 99–115). New York, NY: Springer.