JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education
Journal of Educational Measurement
Summer 2015, Vol. 52, No. 2, pp. 200–222

Extended Mixed-Effects Item Response Models With the MH-RM Algorithm

R. Philip Chalmers
York University

A mixed-effects item response theory (IRT) model is presented as a logical extension of the generalized linear mixed-effects modeling approach to formulating explanatory IRT models. Fixed and random coefficients in the extended model are estimated using a Metropolis-Hastings Robbins-Monro (MH-RM) stochastic imputation algorithm to accommodate for increased dimensionality due to modeling multiple design- and trait-based random effects. As a consequence of using this algorithm, more flexible explanatory IRT models, such as the multidimensional four-parameter logistic model, are easily organized and efficiently estimated for unidimensional and multidimensional tests. Rasch versions of the linear latent trait and latent regression model, along with their extensions, are presented and discussed, Monte Carlo simulations are conducted to determine the efficiency of parameter recovery of the MH-RM algorithm, and an empirical example using the extended mixed-effects IRT model is presented.

Item response theory (IRT) is a latent variable framework designed to model how psychological traits interact with individual test items. More specifically, IRT is a probabilistic framework for modeling categorical response data by formalizing the interaction between item and person parameters (Lord, 1980). Person parameters are generally assumed to be latent (i.e., random) and continuously distributed variables that are independent of the items administered, such as mathematics and reading ability, while item properties have traditionally been understood as fixed variables that describe properties inherent only to items. Item-level properties may include how difficult items are, how well they discriminate between individuals with differing abilities, whether other psychological behaviors such as systematic guessing are present, and so on.

Item response models traditionally were introduced as an unconditional representation of the test data, in that there were no exogenous covariates included to model (or "explain") the response probabilities. However, in order to model item-level data conditional on external covariates, and to capitalize on existent software, De Boeck and Wilson (2004) and contributing authors demonstrated that specialized IRT models could be estimated within a generalized linear mixed modeling (GLMM) framework. Specifically, the Rasch model (Rasch, 1960) could be realized when applying a binomial link function, modeling item intercepts as fixed effects, and representing the unidimensional latent trait as a random effect. Because the unconditional Rasch model naturally could be expressed within the GLMM framework, additional fixed and random variables could also be included for predicting the dichotomous response data (De Boeck & Wilson, 2004).

Although the GLMM approach to modeling item response data appears promising, several authors have cautioned the exclusive use of the Rasch family of IRT models for routine item analysis; largely, this is due to the model's inability to adequately represent psychological tests both theoretically and empirically (see, e.g., Divgi, 1986; Goldstein, 1979; Whitely, 1977; Whitely & Dawis, 1974). Rasch models assume that all items measure the latent construct equally well, and this implies that all items have identical pairwise second-order moments[1] (i.e., covariances). The two-parameter logistic model (2PL; Lord & Novick, 1968), on the other hand, was constructed to specifically account for unequal correlations that items often display with the latent trait. Unlike the Rasch model, the 2PL model recognizes that some items are better indicators of latent variables than others, which is a property that is analogous to traditional linear factor analysis. Specifically, the 2PL model contains an additional "slope" or "discrimination" parameter that modulates how a person's "ability" or trait level influences the probability of item endorsement. Further item-design complications arise when accounting for additional systematic item effects, such as guessing or careless behavior (i.e., 3PL and 4PL models), when modeling tests that have a multidimensional structure and contain unequal slope parameters, when modeling general polytomous item response models such as the multidimensional nominal response model (Thissen, Cai, & Bock, 2010), and when mixing dichotomous and polytomous item designs within the same test.

Estimation Approaches for Extended IRT Models

Several authors have presented a mixed array of extensions to the simple Rasch models when including external covariates; for instance, see Embretson (1999), Geerlings and Glas (2001), later chapters in the collected works of De Boeck and Wilson (2004), Cho, De Boeck, Embretson, and Rabe-Hesketh (2014), and several others. Estimation of these models has ranged from implementing simple numerical quadrature-based integration methods to marginalize the required random-effect terms, which can be effective when the number of random effects is small, to implementing more general stochastic estimation techniques such as data-augmented Gibbs sampling. These explanatory IRT models have a wide range of complexity, and are often met with the problem of estimating the desired parameters and their associated statistical variability in a flexible and effective manner. When the number of random-effect terms required grows too large, as is often caused by multidimensionality in the test or by including multiple residual variance terms at different analysis levels, estimating model parameters becomes a computationally burdensome task.

A potential solution to managing more complicated IRT models and design effects with multiple random-effect terms is to employ general purpose Bayesian Markov chain Monte Carlo (MCMC) estimation techniques. The work by Fox and Glas (2001) is one such instance of applying a Gibbs sampler to estimate IRT measurement models with exogenous fixed and random effects simultaneously, and their approach provided a unified framework for modeling a wide range of multilevel IRT models. The multilevel IRT model is a suitable model to implement when there is a hierarchical nesting structure in the ability parameters, such as when students

are nested within school and countries. An unfortunate consequence of the purely Bayesian estimation approach, however, is the increase in estimation times and computational resources required. Hybrid algorithms that contain elements of Bayesian and maximum-likelihood methodology for estimating multilevel IRT models have been proposed to avoid a fully Bayesian approach, thereby decreasing estimation times and facilitate parameter interpretations within a Frequentist framework. This article will explore one of the more recent hybrid estimation algorithms, and demonstrates how the algorithm performs when modeling more general IRT models with exogenous covariates in both the latent trait and intercept parameters.

This article provides an overview of an extended model used to estimate explanatory IRT models. This extended model demonstrates how to include important item-level parameter information not present in Rasch models, such as those found in the 2PL, 3PL, and 4PL models. However, the additional model complexity introduces computational challenges that are difficult to accommodate using standard estimation techniques. Therefore, Cai's (2010a) Metropolis-Hastings Robbins-Monro (MH-RM) hybrid algorithm is implemented to provide reasonable estimates for model parameters, to adequately account for a larger number of random intercept and latent variable effects, and to deal with missing data by using full-information maximum-likelihood estimation. To demonstrate the usefulness of estimating item- and person-level intercept effects with the MH-RM algorithm, various examples are presented, and, where possible, contrasted with existing GLMM software to demonstrate the potential benefits of the MH-RM algorithm compared to the GLMM estimation with the Laplace algorithm.

Generalized Linear Mixed Models

Linear mixed-effects models that do not contain within-cluster covariance structures[2] can be expressed as

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta}, \tag{1}$$

where $\mathbf{y}$ is a continuous dependent variable, $\mathbf{X}$ is an $N \times k$ design matrix for the k fixed-effect coefficients ($\boldsymbol{\beta}$), and $\mathbf{Z}$ is an $N \times p$ design matrix for the p random-effect coefficients ($\boldsymbol{\delta}$) (McCulloch & Searle, 2001). The $\boldsymbol{\delta}$ coefficients are commonly assumed to follow a multivariate normal distribution, $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \Delta)$. From this assumption, it can be seen that the dependent variable will naturally follow the multivariate normal distribution

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta},\ \mathbf{Z}\Delta\mathbf{Z}'\right). \tag{2}$$

When compared to a linear regression containing a simpler covariance structure, $\mathbf{y} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta},\ \sigma^2\mathbf{I}\right)$, Equation 2 illustrates the consequence of including additional random coefficients. The implied covariance matrix formed by $\mathbf{Z}\Delta\mathbf{Z}'$ generates a more complicated interdependence structure to allow covariation between different levels of observation. Because of this added complication, mixed-effects models generally require more intricate estimation algorithms to account for the simultaneous interdependencies.

In cases where **y** contains a finite number of unique realizations, the assumption that **y** is continuous is often untenable. IRT models, for instance, are almost exclusively specialized for stimuli recorded using a categorical rubric (e.g., correct versus incorrect, partial credit scoring, or responding to Likert-type items), and therefore cannot be appropriately modeled by linear multilevel modeling methods. However, the assumption that **y** is continuously distributed can be relaxed by applying special transformations which linearize the model implied expected values. Linear models that require canonical transformations to accommodate for non-normal error structures have traditionally been called generalized linear models (GLM), and their extension, which include a mixture of fixed and random covariate effects, has been called the generalized linear mixed model (GLMM; McCulloch & Searle, 2001).

Although there are several well established mixed-effects modeling software available for GLMMs, their implementations are not identical. There is a variety of specialized software which can estimate models containing noncontinuous outcomes, such as `PROC GLIMMIX` and `PROC NLMIXED` (SAS Institute Inc., 2010), or open-sourced GLMM software such as the `glmer()` function from the `lme4` package in R (Bates, Maechler, Bolker, & Walker, 2014). SAS has three estimation methods available for categorical and count data: penalized quasi-likelihood (PQL), Laplace approximation, and adaptive quadrature. PQL is a less computationally intensive estimator and gives acceptable results in many cases, but can produce biased estimates as a consequence. When population variance values are large, or events are rare, other estimation approaches are usually preferable. The Laplace estimation option in `PROC GLIMMIX` or the `lme4` package can provide relatively accurate estimates in a variety of circumstances, although these are known to be less accurate than integration-based methods when the number of categories in the dependent variable is low (Joe, 2008). However, Laplace integration often demonstrates reduced estimation times for many classes of GLMM compared to other more intensive integration methods, and therefore may be the most practical approach when estimating complex models. Additionally, SAS's `NLMIXED` procedure uses an adaptive quadrature approach for direct numerical integration that is computationally intensive but produces the most accurate results across a wide variety of applications (Pinheiro & Bates, 2000). However, adaptive quadrature is often limited to models with only two or three random effects because of the exponential increase in computational power required to evaluate multiple adaptive integrals.

Rasch IRT Models as GLMMs

The simplest item response model for dichotomously scored data is the one-parameter logistic (1PL) model. This model is often called the "Rasch" model after Georges Rasch (1960), who emphasized its theoretical importance in measurement theory. The 1PL model consists of a single latent (i.e., random) variable for the ith individual, θ_i, where $\theta \sim \mathcal{N}(0, \sigma^2_\theta)$, and a fixed intercept term for the jth item, d_j. The probability of correctly responding to the jth administered item is expressed as a logistic function

$$P(y_{ij} = 1|\theta_i, d_j) = \frac{\exp(\theta_i + d_j)}{1 + \exp(\theta_i + d_j)}, \tag{3}$$

and the probability of incorrectly responding to the item is the natural complement, $P(y_{ij} = 0|\theta_i, d_j) = 1 - P(y_{ij} = 1|\theta_i, d_j)$. This particular parameterization is the so-called slope-intercept form of IRT models. Therefore, the d parameter in Equation 3 may be interpreted as a quantification of the item "easiness" rather than the usual "difficulty" interpretation.

The Rasch family of IRT models can be specified as a GLMM by recognizing that the J independent d_j parameters can be estimated as $\boldsymbol{\beta}$ coefficients organized by J unique $N \times k$ design matrices, $\mathbf{X}_j$. Additionally, the collection of all $N \times 1$ person-varying θ parameters, $\boldsymbol{\Theta}$, can be defined by a separate item-varying design matrix, $\mathbf{Z}_j$, which controls the random coefficients, $\boldsymbol{\delta}$. In matrix notation we can express this model as

$$\mathbf{P}_j(\mathbf{y} = \mathbf{1}|\boldsymbol{\beta}, \boldsymbol{\delta}) = \frac{\exp(\mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\boldsymbol{\delta})}{1 + \exp(\mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\boldsymbol{\delta})}, \tag{4}$$

where $\mathbf{P}_j$ represents the jth item probability vector for all participants sampled, and $\boldsymbol{\delta}$ represents the collection of all random coefficients.

The fixed coefficients in Equation 4 allow for external predictors to explain variability across persons, items, or both. The variance of the random coefficients are also estimated, and these can be important in determining how much residual variability remains at particular analysis levels following the inclusion of fixed effect variables. As such, one of the goals in multilevel and mixed-effects modeling is often to reduce the residual variability of the random effects so that the deterministic nature of the fixed effect predictors can be interpreted with better confidence. In IRT, the standard Rasch model can be interpreted as an unconditional model because no predictors external to the response data are included. Thus, the variance of the random effect can be readily interpreted as the amount of variability in the participants' ability scores. However, when additional fixed- and random-effect predictors are included at the participant level, the variance of θ becomes conditional on these effects. This is important to recognize because the variability at the person level (i.e., the θ values) can change when additional predictor variables are included in the analysis, and therefore the interpretation of $VAR(\theta)$ will become a conditional or residual variance rather than the unconditional variability of latent ability scores.

Extended Mixed-Effects IRT

The simple 1PL model in Equation 3 is in fact a highly constrained version of the multidimensional four-parameter logistic model,

$$P_j(y_i = 1|\boldsymbol{\theta}_i, \mathbf{a}_j, d_j, g_j, u_j) = g_j + (u_j - g_j)\frac{\exp(\boldsymbol{\theta}_i\mathbf{a}_j + d_j)}{1 + \exp(\boldsymbol{\theta}_i\mathbf{a}_j + d_j)}. \tag{5}$$

Here, $\mathbf{a}_j$ represents an $m \times 1$ vector of slope or discrimination parameters that specify the degree to which an individual's abilities or latent traits ($\boldsymbol{\theta}_i$) influence the probability of endorsing an item, and the latent traits now follow a multivariate normal distribution, $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Two other parameters can be modeled to apply lower (g) and upper (u) asymptotes that are useful for restricting the range of expected probabilities to fall between values other than 0 and 1. These bounds are used to accommodate psychological guessing (g) and careless (u) answering behavior in

testing applications. The model in Equation 5 is not uniquely identified by the data, and therefore requires additional parameter constraints. A common identification constraint is to set the hyper-parameter variance terms for $\boldsymbol{\theta}$ to 1, such that $\boldsymbol{\Sigma}$ is estimated as a correlation matrix.

In order to encompass a wider variety of IRT models that do not belong to the Rasch family, Equation 4 can be further extended to include the missing parameters (g, u, and $\mathbf{a}$) found in the multidimensional 4PL model, and has the form

$$\mathbf{P}_j(\mathbf{y} = \mathbf{1}|\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\delta}, \mathbf{a}_j, g_j, u_j) = g_j + (u_j - g_j)\frac{\exp(\boldsymbol{\Theta}\mathbf{a}_j + \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\boldsymbol{\delta})}{1 + \exp(\boldsymbol{\Theta}\mathbf{a}_j + \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\boldsymbol{\delta})}. \quad (6)$$

For ease of reference, Equation 6 will be referred to as the extended mixed-effects IRT model (EMEIRT). Here, the $\mathbf{Z}$ design matrix no longer contains the ability parameters, as was the case in the previous GLMM representation, and instead only consists of random-effect terms that represent variation in the intercept parameters. The $\boldsymbol{\Theta}$ matrix is the $N \times m$ ability parameters which is now treated distinctly from $\boldsymbol{\delta}$ so that the discrimination parameters can be applied.[3] The $\boldsymbol{\Theta}$ matrix may further be decomposed into fixed and random components to exclusively model person-level ability variation by decomposition $\boldsymbol{\Theta}$ into specific between-subjects matrices, $\boldsymbol{\Theta} = \mathbf{V}\boldsymbol{\Gamma} + \mathbf{W}\boldsymbol{\zeta} + \boldsymbol{\epsilon}$, where $\mathbf{V}$ is an $N \times r$ design matrix for the matrix of fixed-effect parameters, $\boldsymbol{\Gamma}$, $\mathbf{W}$ is an $N \times s$ design matrix for the matrix of random effects, $\boldsymbol{\zeta}$, and $\boldsymbol{\epsilon}$ is an $N \times m$ matrix of residuals. The decomposition of $\boldsymbol{\Theta}$ is in the form of a multivariate linear mixed-effects model (MacCallum, Kim, Malarkey, & Kiecolt-Glaser, 1997), and as such can be used to model separate coefficients for each latent trait while accounting for potential covariation in $\boldsymbol{\Theta}$. The remaining design matrices and coefficients serve the same purpose as in the GLMM approach, where again fixed ($\boldsymbol{\beta}$) and random ($\boldsymbol{\delta}$) coefficients may be included for modeling variation with the intercept parameters at any level of the analysis. However, when estimating Rasch models the $\mathbf{V}$ and $\mathbf{W}$ matrices are generally not required because they can be completely re-expressed using the $\mathbf{X}$ and $\mathbf{Z}$ matrices alone.

The design matrices $\mathbf{X}_j$ and $\mathbf{Z}_j$ can be structured for multiple purposes. However, they are primarily organized to control the effects caused by the item-level intercept parameters. If covariate designs that affect the person-level are also included in these design matrices then the covariates will directly influence the response probabilities for the selected participants. It is interesting to note that the application of person-level covariates through $\mathbf{X}_j$ and $\mathbf{Z}_j$ are largely independent of the discrimination pattern induced by $\mathbf{a}_j$, and this can have subtle but important consequences in analyses when compared to techniques such as latent regression modeling to explain between-person variability (e.g., Adams, Wilson, & Wang, 1997). If, for instance, an item has no correlation with the latent variables in $\boldsymbol{\Theta}$, such that $\mathbf{a}_j = \mathbf{0}$, then the model reduces to a standard GLMM without any person-level residual terms; however, the person-level covariates (e.g., IQ scores obtained from an external test) still can affect the probability of item endorsement. This property can be contrasted with the use of the strictly person-level design matrices, $\mathbf{V}$ and $\mathbf{W}$, that serve to explain variability in the $\boldsymbol{\Theta}$ scores directly, and therefore will only indirectly affect the response probabilities through the weighting coefficients in $\mathbf{a}_j$. Furthermore, special organization of $\mathbf{V}$ and $\mathbf{W}$ will result in conditional IRT models for $\boldsymbol{\Theta}$ that have seen widespread

use, such as the latent regression model when only **V** is included along with a single random coefficient for each participant, or the multilevel IRT model (e.g., Fox & Glas, 2001) if random-effect terms are structured within the **W** matrix.

In addition to the modeling exogenous covariates, identification constraints should be considered because they can affect the model interpretation in different ways. As was the case when interpreting the variance terms in Equation 4, including external covariates at the person level will affect the estimate of the $VAR(\boldsymbol{\Theta})$ terms. However, because the model variances for $\boldsymbol{\Theta}$ are often constrained to 1 for identification, the associated $\hat{\mathbf{a}}$ estimates will be affected by the person-level intercepts instead. Therefore, when interpreting the $\hat{\mathbf{a}}$ estimates, slopes should only be regarded as genuine discrimination parameters only when no person-level covariates have been modeled, and as coefficients that modulate the residual variances of $\boldsymbol{\Theta}$ otherwise. If alternative identification constraints are imposed, such as fixing various **a** parameters to 1 and freely estimating the variance terms for $\boldsymbol{\Theta}$, then the remaining parameters will be estimated in reference to these constraints and may help to alleviate the interpretation of the slope parameters and, in some cases, improve stability in the estimated parameters. Note that this approach is often adopted in the structural equation modeling literature when modeling regression effects with latent variables (Bollen, 1989).

Although Equation 6 may appear to be a relatively simple extension of Equation 4, there are some caveats that make its estimation difficult. To begin, the model does not have a canonical link similar to Equation 4, so capitalizing on standard GLMM estimation software does not appear to be a feasible solution. We can see that the model does not contain strictly additive components either because the inner product $\boldsymbol{\Theta}\mathbf{a}_j$ is a combination of the fixed item slopes with a random person-level ability (or residual) term. Therefore, a method to estimate nonlinear mixed-effects models with specialized constraints for the expected values is required. Additional specification difficulties arise in the application of Equation 6 when considering non-nested or crossed random effects, such as in situations where multilevel grouping structures are accommodated for $\boldsymbol{\Theta}$, or when including random item-level terms commonly implemented in experimental test designs (De Boeck & Wilson, 2004). As well, high dimensionality due to estimating multiple random effects, in addition to the already potentially problematic $\boldsymbol{\Theta}$ effects in multidimensional IRT tests, is often numerically problematic in even the simplest cases. With the growing popularity of multidimensional confirmatory IRT models (e.g., Cai, 2010b), it is not uncommon to model three or more latent dimensions for $\boldsymbol{\Theta}$ in addition to other random grouping structures.

MH-RM Algorithm for Extended Mixed-Effects IRT Models

In order to avoid the computational complexity of evaluating multiple integrals in confirmatory item response models, Cai (2010a) proposed a hybrid algorithm which benefits from both Bayesian and frequentist approaches to parameter estimation which he called the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm. The MH-RM is an application inspired from Fisher's (1925) observation that the gradient of the observed-data log-likelihood is equal to the *expectation* of the gradient for the complete-data log-likelihood. The algorithm draws from this relationship by imputing plausible values for the random effects with an MH sampler

to form a pseudo-complete data set, and updates the fixed-effect estimates using the more analytically tractable complete-data gradient vector and Hessian matrix. However, because the stochastic draws in the MH samples are noisy and do not deterministically converge to the maximum-likelihood location, an RM filter is applied to dampen the imputation effects and to slowly converge to the maximum-likelihood solution (Cai, 2010a).

In addition to being useful for confirmatory item factor analysis, the MH-RM can be modified to include additional fixed- and random-effect components at different levels of the analysis; hence, it can be used to estimate Equation 6 directly. Because Equation 6 contains a mix of fixed and random coefficients, the strategy for utilizing the MH-RM is to impute plausible random coefficients by drawing from the complete-data likelihood

$$L(\boldsymbol{\Psi}) = \left[\prod_{i=1}^{N} \prod_{j=1}^{J} P(y_{ij} = 1|\boldsymbol{\theta}_i, \boldsymbol{\beta}, \delta, \boldsymbol{\gamma}, \boldsymbol{\zeta}, \boldsymbol{\Xi}, \boldsymbol{\psi}_j)^{c_{ij}} \cdot \right.$$
$$\left. \times [1 - P(y_{ij} = 1|\boldsymbol{\theta}_i, \boldsymbol{\beta}, \delta, \boldsymbol{\gamma}, \boldsymbol{\zeta}, \boldsymbol{\Xi}, \boldsymbol{\psi}_j)]^{1-c_{ij}} \right] \times \left[\prod_{k=1}^{K} \prod_{n=1}^{n|k} \phi(\boldsymbol{\delta}_{nk}|\boldsymbol{\mu}_{\delta_k}, \boldsymbol{\Sigma}_{\delta_k}) \right]$$
$$\times \left[\prod_{h=1}^{H} \prod_{p=1}^{p|h} \phi(\boldsymbol{\zeta}_{ph}|\boldsymbol{\mu}_{\zeta_h}, \boldsymbol{\Sigma}_{\zeta_h}) \right] \times \left[\prod_{i=1}^{N} \phi(\boldsymbol{\Theta}_i|\mathbf{0}, \boldsymbol{\Sigma}_{(\boldsymbol{\Theta}|\boldsymbol{\gamma},\boldsymbol{\zeta})}) \right], \tag{7}$$

where $c_{ij} = 1$ if the item is endorsed by subject i and 0 otherwise, $\boldsymbol{\psi}_j$ is the collection of parameters for the jth item in the 4PL model, $\boldsymbol{\Psi}$ is the collection of all estimated fixed effect and hyper-distribution parameters, and ϕ is multivariate density function for the $K + H + m$ clusters of random effects and residual terms given their respective mean ($\boldsymbol{\mu}$) and covariance matrices ($\boldsymbol{\Sigma}$).

Prior to conditioning the MH draws from the complete-data likelihood on the fixed-effect parameters, each random coefficient must be sampled conditional on all other imputed random coefficients. This can be conceptualized as drawing from a generalized linear model with an "offterm," representing the composite of random coefficients for each individual nested within each item (McCullach & Nelder, 1989). Once the random effects are all obtained to form a pseudo-complete dataset, fixed-effect estimates can be updated using methods typically found in optimizing generalized linear models. After a sufficient number of initial burn-in iterations, the RM filter can then be applied to help cancel the noise of the MH sampling of the random coefficients.

To effectively make use of the RM filter, a burn-in and aggregation stage should be performed so that the starting values for the MH-RM algorithm are not too discrepant from the ML estimates prior to applying the RM filter. However, when random coefficient terms other than the unconditional $\boldsymbol{\Theta}$ effects are included, an additional burn-in stage is recommended so that the $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ terms will be closer to their ML estimates before drawing the additional $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\zeta}}$ coefficients. It is also important to monitor the proportion of accepted draws in the MH samplers for each set of random effects so that the parameter space will be adequately covered

across successive iterations. Finally, after stable ML estimates have been obtained, the observed information matrix may be computed by using information generated throughout the RM estimation stage (Cai, 2010a) or by Monte Carlo integration when treating the ML estimates as fixed values (Monroe & Cai, 2014). For more details regarding the MH-RM algorithm and its implementation, refer to Cai (2010a, b), Chalmers (2012), Chalmers and Flora (2014), and the references therein.

Applications of Mixed-Effects IRT Models

The linear latent trait model (LLTM; Fischer, 1983) can be understood as a design-constrained version of the simple Rasch model, and is useful when modeling items with similar intercept parameters often due to known or suspected test-design effects. The model has the form

$$\mathbf{P}_j(\mathbf{y} = \mathbf{1}|\boldsymbol{\Theta}, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\Theta} + \mathbf{X}_j\boldsymbol{\beta})}{1 + \exp(\boldsymbol{\Theta} + \mathbf{X}_j\boldsymbol{\beta})}, \tag{8}$$

where $\mathbf{X}_j$ is an $N \times k$ design matrix for the item intercepts $\boldsymbol{\beta}$. The number of columns in $\mathbf{X}_j$ is generally less than the number of items, and therefore the number of unique intercepts to model is less than the number of items. For example, a test may be designed such that similarly easy items are administered in the first half, while similarly difficult items are administered in the second half. Therefore, it may be beneficial to constrain the first and second halves of the intercepts in the test to be equal, effectively reducing the number intercept parameters to only two rather than estimating a unique intercept for each item. This type of reduction offers a much more parsimonious model, and can provide smaller standard errors due to increasing the overall degrees of freedom.

Equation 6 can be reduced to the traditional LLTM model by setting all the $\mathbf{a}_j$ slope parameters to 1 and dropping the $\mathbf{Z}\boldsymbol{\delta}$ term by fixing $\mathbf{Z}$ to be a matrix of $\mathbf{0}$'s. Following that, the $\mathbf{X}_j$ design matrices can be constructed so that relevant item intercepts are applied only to particular items. In the six-item example presented above, we could generate the LLTM item design by setting

$$\mathbf{X}_1 = \mathbf{X}_2 = \mathbf{X}_3 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{X}_4 = \mathbf{X}_5 = \mathbf{X}_6 = \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix},$$

and estimate the elements $\boldsymbol{\beta}' = [\beta_{first}\ \beta_{last}]$. Additional item-level predictor variables may be included if other types of item design effects are present, or if there are additional group or group-by-item interaction terms. However, when using the EMEIRT model additional item-level random effects can readily be included by specifying appropriate $\mathbf{Z}$ matrices, and the effect of $\boldsymbol{\Theta}$ need not be considered constant for all items, which is achieved by estimating the fixed item-slope parameters. Hence, the 2PL model, as well as the more general 4PL model, can be modeled in the way that the LLTM has traditionally been applied to Rasch item types.

Alternatively, if only group-level predictors are included (i.e., no item-level designs, including the usual item intercepts), then the corresponding fixed- and random-effect design matrices will all be the same across items, such that $\mathbf{X} = \mathbf{X}_j$ and $\mathbf{Z} = \mathbf{Z}_j$ for all j items. For the $N = 5$, $J = 6$ example, if each individual has a known IQ score of 101, 95, 98, 110, and 85, respectively, and the first two individuals are male while the last three are female, a possible design matrix for determining whether these effects influence the probability of positive endorsement could be

$$\mathbf{X} = \mathbf{X}_j = \begin{bmatrix} 101 & 1 & 0 \\ 95 & 1 & 0 \\ 98 & 0 & 1 \\ 110 & 0 & 1 \\ 85 & 0 & 1 \end{bmatrix}$$

for the corresponding coefficients $\boldsymbol{\beta} = [\beta_{IQ}\ \beta_{Male}\ \beta_{Female}]$. Further manipulating the EMEIRT model can allow the slope parameters ($\mathbf{a}$) to be estimated, but, if desired, these can be treated distinctly from the person-level fixed effects. Hence, the approach is not only restricted to Rasch-type models, nor is it limited to modeling person-level covariates indirectly through the $\boldsymbol{\Theta}$ terms as in latent regression analysis. Treating the $\boldsymbol{\beta}$ coefficients independently of $\boldsymbol{\Theta}$ also has interesting benefits when the test structure is multidimensional, in that the covariates can be interpreted in probabilistic terms rather than in reference to which latent trait indirectly affects the item probability; this is demonstrated further in the empirical example in the empirical analysis section. Finally, additional item and group-by-item interaction designs may be included, and further random effects at the item- or person-level can be added to induce residual variability effects.

Simulated and Empirical Examples

To demonstrate how the MH-RM algorithm can be used to estimate extended mixed-effects IRT models, we will consider how the variations of the latent regression (Adams et al., 1997) and linear latent trait models (LLTM; Fischer, 1983), as well as their extensions, can be estimated with this algorithm. These two models will be explored using simulated data to understand their relationship with the traditional Rasch and EMEIRT parameterizations, as well as to observe how effective the MH-RM algorithm is at recovering known population parameters. Data generation and estimation of the IRT and EMEIRT models were performed in R (R Core Team, 2014) using the `mirt` package (Chalmers, 2012). Throughout the simulations and examples, the MH-RM algorithm was terminated after three consecutive iterations fell below .001; 200 and 50 iterations were used for the burn-in and MH-RM starting value estimation stages, respectively, and 5000 Monte Carlo draws were used to approximate the observed-data log-likelihood.

Simulated Example

To demonstrate how the general EMEIRT model outlined above can be manipulated, a data set consisting of 15 dichotomous items was generated with a sample

Table 1
Item-Level Parameter Estimates for Model 5

	Simulated				Estimated				
Item	a	d_{group_1}	d_{group_2}	g	$\hat{a}$	SE_a	$\hat{g}_{logit}$	$SE_{\hat{g}_{logit}}$	$\hat{g}$
1	0.6	−1.5	1.5	0.2	0.586	.097	−1.346	.079	.207
2	1.1	−1.5	1.5	.2	1.241	.108	−1.448	.078	.190
3	1.3	−1.5	1.5	.2	1.315	.113	−1.342	.077	.207
4	0.8	−1.5	1.5	.2	1.009	.105	−1.351	.078	.206
5	0.7	−1.5	1.5	.2	1.011	.109	−1.348	.078	.206
6	1.3	0	0	.2	1.113	.113	−1.430	.084	.193
7	1.8	0	0	.2	1.726	.169	−1.391	.081	.199
8	0.7	0	0	.2	0.711	.088	−1.356	.087	.205
9	2.3	0	0	.2	2.172	.263	−1.390	.079	.199
10	1.6	0	0	.2	1.412	.150	−1.328	.082	.209
11	1.0	1.5	−1.5	.2	1.077	.104	−1.351	.077	.206
12	0.9	1.5	−1.5	.2	0.796	.100	−1.358	.078	.205
13	1.0	1.5	−1.5	.2	1.075	.111	−1.376	.077	.202
14	0.9	1.5	−1.5	.2	0.995	.103	−1.384	.078	.200
15	1.0	1.5	−1.5	.2	0.963	.108	−1.390	.078	.199

size of $N = 2,000$, where the θ parameters were drawn from a standard normal distribution. Item parameters used to simulate the data are presented in Table 1, and are organized into two groups. The design was chosen to emulate a scenario where the slope parameters are equal across groups; however, both groups show a unique item-design effect such that there is a group-by-item interaction effect present. A fixed lower-bound parameter of $g = .2$ was also included to indicate that these items were all scored from 5-option multiple choice test items. Finally, an arbitrary independent predictor variable, signifying an irrelevant and uncorrected standardized score on an external test, was drawn from a standard normal distribution which was completely independent of the characteristics of the test.[4]

To estimate the effect of the slope parameters, a unidimensional Rasch and 3PL model was estimated from Equation 5 using a multiple-group estimation method where there were no cross-group equality constraints imposed. To help ensure proper convergence and model stability, an informative normal prior distribution was added to the g parameters with a mean of -1.386 and a $SD = .15$.[5] Following convergence, the model information statistics were compared to determine which model provided a better fit to the data. The Rasch model produced AIC and BIC values of 34,785.28 and 34,964.51, respectively, whereas the 3PL model had AIC and BIC values of 34,629.82 and 35,133.90. The AIC statistics suggest that the 3PL model should be preferred over the Rasch model for these data, and this observation was expected given the population parameters used to simulate the data. The multiple-group 3PL model is referred to as "Model 1," and is treated as an over-parameterized model to be further simplified.

Following the computation of the unconditional multiple-group IRT models, five additional models were estimated. These models were:

- Model 2 – A single group 3PL model. This model was hypothesized to have the worst fit to the data because it ignores the "group" effect.
- Model 3 – A single group 3PL model that included an item-design to reduce the degrees of freedom. The item-design constrained the intercepts for items 1–5, 6–10, and 11–15 to be the same. This model was hypothesized to have a better fit than Model 2 in terms of the information criteria, though ignoring the Group $\times$ Item-design effect will still negatively affect the accuracy of predicting the data.
- Model 4 – A 3PL model with an item-design by group interaction effect. This allows the item-design coefficients to vary across group, and therefore six intercepts are used to model the design effect. This model should provide the best fit to the data because it was the population model used to generate the data.
- Model 5 – This model is the same as Model 3, with the inclusion of a continuous predictor variable that was simulated to have no relationship with the model. It was believed that the inclusion of this parameter would not significantly improve the log-likelihood, the information statistics would increase, and the parameter estimate for the continuous variable effect would be nonsignificant.
- Model 6 – This model is the same as Model 4, except the item types were estimated as Rasch models instead of 3PL models. It was estimated to explore the potential consequences of ignoring important item-level properties (in this case, slopes and lower-bound parameters).

The estimated item-level coefficients for Model 5, as well as the population generating parameters, are displayed in Table 1. In Table 1 the logit of the lower-bound (g) parameters are reported with their standard errors instead of the raw values. Logits are reported for g because internally the `mirt` package reparameterizes the g parameters to obtain more optimal behavior with the Newton-Raphson optimizer used in the MH-RM algorithm. Additionally, the logit transformation allows the standard errors for g to be nonsymmetric, which is appropriate because the parameter is bounded between $0 \leq g \leq 1$.

The model with the highest log-likelihood was Model 1, which also contained the lowest degrees of freedom because it contained the most freely estimated parameters. However, the information statistics for Model 1 were not as low as Model 4 and 5, or even the BIC in Model 6, indicating that the model was not as parsimonious as it could be. Comparing Models 1 and 4 with a likelihood ratio test revealed that the models were not significantly different from each other, $\chi^2(54) = 27.012$, $p = .999$, and therefore from both an information and likelihood-ratio perspective Model 4 should be preferred over Model 1. Overall, Model 4 contained the lowest AIC and BIC values, and although Model 5 improved the log-likelihood slightly it did not result in statistically improved model fit, $\chi^2(1) = .831$, $p = .362$.

Table 2 displays the estimated design parameters for Models 5 and 6. Both models contained the same intercept design as the population, with the additional inclusion of an arbitrary continuous predictor variable. As was described above, Model 6 was selected to have a 1PL design for each item rather than the population-generated 3PL model, and the consequences of this misspecification are clear from the intercept

Table 2
Design-Level Parameter Estimates for Model 4 and Model 5

	Parameters	$\hat{\beta}$	SE	z	p
Model 5	Item-design low	−1.620	.094	−17.175	.0000
	Item-design mid	−0.074	.076	−0.974	.1651
	Item-design high	1.484	.061	24.492	.0000
	Group two	3.175	.116	27.423	.0000
	Continuous predictor	0.006	.030	0.194	.4230
	Item-design mid × Group two	−3.024	.125	−24.245	.0000
	Item-design high × Group two	−6.317	.144	−43.757	.0000
Model 6	Item-design low	−0.591	.039	−15.170	.0000
	Item-design mid	0.415	.038	10.821	.0000
	Item-design high	1.707	.045	37.592	.0000
	Group two	2.319	.060	38.822	.0000
	Continuous predictor	0.006	.023	0.260	.3974
	Item-design mid × Group two	−2.232	.067	−33.564	.0000
	Item-design high × Group two	−4.673	.073	−64.291	.0000

parameters. In addition to providing much smaller standard errors than Model 5, Model 6 also demonstrated severe bias in the parameter estimates for the intercept item design. Reconstructing the item intercepts from the design parameters in Model 5, we can see that, for the first group, the intercepts were −1.620, −0.074, 1.484, while for the second group the intercepts were 1.555, 0.076, and −1.657, respectively. These intercept parameters were reasonably close to the population parameters in Table 1. However, Model 6 produced the values −0.591, 0.415, and 1.707 for group one, and 1.727, 0.501, and −0.647 for group two, respectively, which were quite discrepant from the simulated population values. Model 6 demonstrates the consequences of making inferences about population parameters when inappropriate IRT models are selected. Furthermore, biased parameters have negative consequences for secondary procedures that make use of the obtained item-parameter estimates (such as computing $\hat{\theta}$ estimates).

Alternatively, the simulated data could have been analyzed by treating the item-intercept parameters as random effects with the goal of discovering sufficient item-level covariates that explain their variability (De Boeck, 2008). Using this approach, a baseline model containing the grouping and continuous predictor variables were fit to the data using 3PL IRT models for each item, while simultaneously including a random item-by-group intercept effect. This resulted in an unconditional variance of $\hat{\sigma}^2_{items \times group} = 4.291$, suggesting that there was substantial variability in the 30 random-intercept coefficients. However, after including the fixed Item-design × Group interaction effect, the residual variability dropped to $\hat{\sigma}^2_{items \times group} = .0003$. This suggested that nearly all the variation in the Item × Group intercepts were removed by these simple but highly significant coefficients, $\chi^2(4) = 6,179.871,\ p < .0001$.

Simulation 1

In addition to comparing nested models using a single data set, it is important to evaluate how effective the MH-RM algorithm is at recovering population parameters in general. To accomplish this, 1,000 additional data sets using the same specifications listed in the previous section were generated and estimated using the configuration in Model 4. Bias and root mean square deviation (RMSD) statistics were computed to gauge the parameter recovery accuracy and efficacy, and can be found in the appendix. In general, the algorithm appeared to recover the population parameters with very little bias, and with a reasonable amount of precision across both the item and design parameters. The largest RMSD for the item parameters was in the item with the largest slope ($a = 2.3$, RMSD $= .252$), while the largest RMSD for the design parameters was for the last five item's Group $\times$ Item-design intercept ($\beta = -6.0$, RMSD $= .117$). These observations are to be expected because more extreme coefficients tend to have larger standard errors. The models converged relatively quickly, requiring an average of 18.41 seconds ($SD = 1.63$) to reach a stable maximum-likelihood location. Overall, the algorithm demonstrated acceptable estimation accuracy, and agrees with previous simulation studies that have implemented the algorithm (see, e.g., Cai, 2010b; Chalmers & Flora, 2014).

Simulation 2

To further evaluate how the MH-RM algorithm can recover parameters when person-level effects and test multidimensionality are present, a second simulation study was conducted. This simulation consisted of a three-factor test with 30 dichotomous items, where each factor uniquely loaded on 10 items to form a simple-structure factor pattern. Two sets of models were estimated: the conditional 2PL model and the conditional Rasch model. For the conditional 2PL model, the slope parameters were drawn from a log-normal distribution, $a \sim \ln \mathcal{N}(.2, .3)$, with intercept parameters drawn from $d \sim \mathcal{N}(0, 1)$; for the conditional Rasch model all slopes were set equal to 1. The $\boldsymbol{\Theta}$ parameters for were drawn from a multivariate normal distribution with a mean vector of $\mathbf{0}$ and correlation matrix $\boldsymbol{\Sigma}$, where all factors inter-correlated $r = .40$. Finally, two person-level effects were constructed: a group membership variable where each subject was assigned to one of three balanced groups, and a continuous covariate drawn from a standard normal distribution. The $\boldsymbol{\beta}$ coefficients for the four conditional intercepts were $\boldsymbol{\beta}' = [\beta_{G1}\ \beta_{G2}\ \beta_{G3}\ \beta_C] = [.0,\ 1.0,\ 2.0,\ .5]$.

To recover the slope parameters in the conditional 2PL model, one slope on each of the latent factors was fixed to 1 to properly control the metric, and in turn the latent variance terms were freely estimated instead. The results from the second simulation design after performing 1,000 replications for $N = 600$ are displayed in Table 3. Overall, the MH-RM algorithm recovered all parameters with little bias and with reasonable efficiency. The Rasch model was consistently recovered with greater precision compared to the 2PL model; this was to be expected because the model is more parsimonious, and in general is easier to estimate because it only contains intercept terms. What can be concluded from this simulation is that overall the MH-RM algorithm is effective at recovering person-level covariate parameters, and remains

Table 3
Simulation 2 Results for Conditional Rasch and 2PL Models

		Parameters						
Model	Statistic	β_{G2}	β_{G3}	β_C	d	$\mathbf{a}$	$VAR(\boldsymbol{\Theta})$	$COV(\boldsymbol{\Theta})$
Rasch	Bias	−.0171	−.0061	−.0005	.0065	—	−.0184	.0000
	RMSD	.0924	.0946	.0376	.1299	—	.1028	.0599
2PL	Bias	−.0029	.0087	.0015	.0048	.0671	−.0508	−.0248
	RMSD	.0983	.1042	.0403	.1548	.2453	.2021	.0744

Table 4
Simulation 3 Results and Estimation Time (With Standard Deviations) for Unconditional Rasch Model With Cross-Random Effects Using the Laplace (`lme4`*) and MH-RM (*`mirt`*) Algorithm*

	Bias		RMSD		Average Estimation
Estimator	$VAR(\boldsymbol{\theta})$	$VAR(\mathbf{d})$	$VAR(\boldsymbol{\theta})$	$VAR(\mathbf{d})$	Time (in seconds)
Laplace	.007	.033	.078	.203	27.90 (3.04)
MH-RM	.003	.033	.078	.205	47.29 (4.02)

effective in tests with multidimensional and correlated latent traits. What is more, the estimation times were also quite low despite the multidimensionality, where the conditional 2PL model converged to a maximum in 18.52 second on average ($SD = 2.12$), while the conditional Rasch model converged after 8.92 seconds ($SD = .46$).

Simulation 3

Lastly, two simulations were performed to determine the degree to which the corresponding hyper-parameters for estimated random effects could be recovered in crossed and multilevel analysis designs, respectively. For the first simulation, a Rasch model with no covariates was constructed to form a crossed random-effect design, where $N = 1{,}000$ individual variation parameters ($\boldsymbol{\Theta}$) were drawn from a normal distribution with a variance of 1.5, and $J = 50$ item-level intercept parameters were drawn from a normal distribution with a variance of one. These crossed-random effects were estimated using the MH-RM algorithm, and the simulated data were also estimated using the Laplace algorithm available in the `lme4` package (Bates et al., 2014). A total of 1,000 data sets were estimated using these two methods, and results are presented in Table 4.

As can be seen in Table 4, the Laplace and MH-RM estimation algorithms reach comparable bias and efficiency results for the random ability and intercept effects. RMSD values for the variance of the intercept parameters were much higher than the ability parameters because the number of exchangeable clusters was relatively low, especially compared to the number of clusters in the random ability effect. As well, the MH-RM required slightly more estimation time to reach a stable solution

Table 5
Simulation 3 Results and Estimation Time (With Standard Deviations) for the Unconditional Three-Dimensional Rasch Model With a Hierarchical Grouping Effect

	Bias (RMSD)				Average Estimation
Estimator	$VAR(\boldsymbol{\Theta})$	$COV(\boldsymbol{\Theta})$	$VAR(\boldsymbol{\delta})$	$\mathbf{d}$	Time (in seconds)
Laplace	−.043 (.056)	−.004 (.038)	−.022 (.111)	.006 (.098)	3,537.82 (608.39)
MH-RM	−.013 (.060)	−.001 (.040)	−.042 (.115)	.031 (.106)	69.84 (6.723)

than the GLMM with the Laplace algorithm, but not by a large factor. Overall, the random-effect hyper-parameters appeared to be recovered with equivalent precision regardless of the estimation methodology. Of note, both algorithms were negatively affected by smaller cluster sizes, in that smaller cluster sizes resulted in more bias and larger RMSD values.

In the second simulation study, a multilevel Rasch model was generated with one grouping random-effect term acting on the person level. The simulated tests contained $N = 2,000$ subjects and $J = 30$ items, where the underlying factor structure was composed from a three-dimensional Rasch model with a simple structure pattern. For the random group effects, 100 coefficients were drawn from a normal distribution, $\boldsymbol{\delta} \sim N(0, .75)$, and organized such that each cluster contained 20 person-level observations each. In turn, the latent traits were organized to uniquely load on 10 items, and possessed an inter-factor covariance structure of

$$COV(\boldsymbol{\Theta}) = \begin{bmatrix} 1 & & \\ .5 & 1 & \\ .5 & .5 & 1 \end{bmatrix}.$$

Individual $\boldsymbol{\Theta}_i$ elements were generated from a multivariate normal distribution conditional on the mean structure implied by the hierarchical grouping effect. Finally, the intercept parameters were drawn from a normal distribution, $\mathbf{d} \sim N(0, .5)$, and were estimated as fixed-effect parameters. This design contained a total of four random effects (three from the test structure and one from the grouping effect) and 37 fixed-effect parameters (7 variances and covariances, and 30 item intercepts).

Monte Carlo simulations for the second design were estimated using 1,000 replications and again fit using the MH-RM algorithm and the `glmer()` function in `lme4` with the Laplace algorithm. Results of this simulation were organized and displayed in Table 5. As was the case in the previous simulation, the MH-RM and Laplace algorithm appear to recover the variance and covariance parameters well, where the RMSD values are again inversely proportional to the size of the clusters. As well, the fixed item-level intercepts were recovered adequately using both approaches, demonstrating comparable bias and RMSD values. However, where these approaches starkly differed was in the overall estimation time required for the models to converge. The Laplace algorithm required approximately 59 minutes on average to converge to a maximum-likelihood solution, whereas the MH-RM required only 70 seconds on average. This suggests that the MH-RM algorithm can potentially decrease estimation times by at least 50 orders of magnitude in some designs, and

for more realistic sample and item-bank sizes the MH-RM may well be a viable general purpose estimator, even for IRT models that generally fit within the GLMM framework.

Empirical Example

Next, the EMEIRT model was applied to an empirical data set consisting of 84 mathematics items which were administered to 5,456 high-school students sampled from 274 schools in the United States. The data were obtained from the 2003 administration of the Programme for International Student Assessment (PISA) study. The PISA is an ongoing survey for assessing whether 15-year-old students have acquired essential knowledge and skill sets that are important in future career and academic opportunities. The PISA largely focuses on mathematical reasoning skills while also measuring reading, science, and problem-solving abilities, though to a lesser degree. The survey is administrated every three years in a wide variety of countries, and has been an informative tool for modifying international educational systems.

The majority of the 84 items in the mathematics section of the PISA were scored in a correct-incorrect format. However, a small selection of items included a partial credit scoring format, and this information was included in the subsequent IRT analyses to maximize the information provided by the test. Additionally, there were large amounts of missing data due to the administration of partial forms (13 total), which were created so that students were not required to answer all 84 mathematics items. Of the potential $5{,}456 \times 84$ item responses, only 31% of the entire data matrix was observed due to the administration of partial forms.

Given the complexity of the item content and design, the dimensionality of the test was investigated using exploratory full-information maximum-likelihood (FIML) item factor analysis with the MH-RM algorithm (Cai, 2010a). Dichotomous items were fit with the multidimensional 2PL model, and the partial credit items were fit with the multidimensional graded (i.e., ordinal) response model. Initially, a unidimensional IRT model was fit to the data, which revealed that approximately one-third of the test displayed negative discrimination coefficients along this unidimensional construct. Further item factor analysis models were fit to the data, and the BIC selection criteria indicated that at least four factors were required to sufficiently explain the response patterns in the data. Items were removed from further analysis if, following a biquartimin rotation (Jennrich & Bentler, 2011), they displayed negative loadings on the general dimension or had no standardized loadings greater than |.15| on any dimension. The biquartimin rotation was selected to capture a general math factor with correlated subcomponents, and is often appropriate for ability tests which aim to measure a common trait. This reduced the test size to only 58 items, where 53 items loaded on the primary factor, and 36, 21, and 14 items loaded on the three specific factors with various cross-loadings. The observed factor pattern for these items was also used to construct a confirmatory item factor analysis pattern of loadings, which was to be used in subsequent analyses.

Next, an EMEIRT model was fit using five explanatory fixed-effects predictor variables which were believed to influence the observed response probabilities. The fixed effects were gender (male or female), birth year (born in 1988 or 1987), country of birth (United States or foreign), English spoken at home, and proportion

Table 6
Linear Mixed-Effects Model With EAP Predicted $\hat{\theta}$ Scores, and Unidimensional and Multidimensional EMEIRT Models for PISA Data

	Linear Mixed Model with EAP Estimates			Unidimensional EMEIRT			Multidimensional EMEIRT		
	$\hat{\beta}$	SE	t	$\hat{\beta}$	SE	z	$\hat{\beta}$	SE	z
Fixed effects:									
Male	−.081	.025	−3.193	−.064	.031	−1.936	−.073	.038	−1.921
Born in 1988	.018	.044	0.410	.096	.045	1.948	.105	.053	1.972
Born in USA	.066	.057	1.157	.017	.067	0.615	.058	.070	0.819
Speaks English at home	−.228	.049	−4.594	−.192	.070	−2.966	−.236	.047	−5.021
Ratio of teachers to students	.184	.230	0.797	.189	.230	0.757	.093	.276	0.336
	$\hat{\sigma}^2$	SE	t	$\hat{\sigma}^2$	SE	z	$\hat{\sigma}^2$	SE	z
Random effects:									
School ID	.043	–	–	.051	–	–	.014	–	–

of teachers to students in each classroom (a level 2 predictor). School ID was modeled as a level 2 random intercept effect to allow for school differences to be captured. For comparative reasons, a unidimensional EMEIRT and a linear mixed-effects model using EAP estimates from the unidimensional IRT model were estimated. The linear mixed-effects model was estimated to show the effect ignoring both factor structure and measurement precision of $\hat{\boldsymbol{\Theta}}$, while the unidimensional EMEIRT model reflected only ignoring the factor structure. The linear mixed-effects model was organized to explain variation in the unidimensional $\hat{\boldsymbol{\Theta}}$ scores, which in turn indirectly explains different response probabilities at the item level. The EMEIRT models, on the other hand, attempted to model the external covariates independent of the factor structure and therefore influence the response pattern probabilities directly after accounting for the factor structure.

As is seen in Table 6, the linear mixed-effects model suggested that there may be meaningful variation between schools with respect to the unidimensional ability scores, and therefore some schools may have slightly higher $\hat{\boldsymbol{\theta}}$ values compared to others. What is concluded from the multidimensional EMEIRT models, however, is that schools did not vary much in their item response probability patterns, and therefore schools generally are not different in their response patterns. These results are not necessarily contradictory, but instead reflect different aspects of the data set. In the linear mixed-effects model, the $\hat{\boldsymbol{\beta}}$ coefficients represent the relative differences in $\hat{\boldsymbol{\Theta}}$ scores due to regressing the scores on the covariates, while in the EMEIRT models the $\hat{\boldsymbol{\beta}}$s reflect changes in relative probability scores conditional on the covariates at the item level directly after controlling for the test's factor structure.

What is clear from this empirical example is that the choice of whether to model differences in $\hat{\boldsymbol{\Theta}}$, or whether to model the item probabilities directly, is important to consider when studying exogenous covariates such as gender or country of origin. Furthermore, if there is reason to believe that exogenous covariates will

have different effects for each item then additional intercept parameters could be modeled within each item by further manipulating the $\mathbf{X}_j$ matrices and adding additional $\boldsymbol{\beta}$ terms to the EMEIRT model. Achieving a similar item-level effect using latent regression methods is more difficult because the focus is on the $\boldsymbol{\Theta}$ scores rather than on the item properties. Hence, regardless of whether the test structure is unidimensional or multidimensional, constructing regression coefficients based on the $\mathbf{X}_j$ design matrices in the EMEIRT model provides a clear interpretation of what the regression coefficients directly represent at the item level.

Discussion

The GLMM can be effective at modeling Rasch-based IRT models, and in doing so the flexibility of the mixed-effects framework and existent software can be adapted for individual test items. The GLMM can include important covariates at the item-response level that are believed to be important predictors of the propensity for item endorsement, and is also capable of modeling item-level designs directly. However, many IRT models in common usage cannot be estimated in this framework, and instead alternative approaches are required so that a wider range of IRT models can be utilized. More importantly, it may be too much to demand from GLMM or general purpose estimation software (such as SAS NLMIXED) to estimate IRT models in item analysis work. For instance, estimation with `lme4` package can take an extended amount of time even for simpler Rasch models (see De Boeck et al., 2011, as well as the simulations above), and although SAS NLMIXED can be manipulated to estimate EMEIRT models the estimation time required to reach convergence for unidimensional tests (and even more so for multidimensional tests) may take hours or even days (e.g., see Sheu, Chen, Su, & Wang, 2005).

The parameters in the EMEIRT model outlined in this article, on the other hand, can be effectively and efficiently estimated by combining the strengths of Bayesian and ML estimation techniques through the MH-RM algorithm. This approach may be ideal for a variety of reasons, namely: to estimate fixed and random effects for a wider variety of dichotomous and polytomous IRT models, to model item- and person-level fixed and random covariates, and to reasonably accommodate higher dimensionality from estimating a larger number of random effects. These estimation properties provide accurate estimates of IRT and exogenous covariate parameters, offer a level of flexibility similar to a fully Bayesian MCMC approach, potentially decrease the estimation times by a meaningful factor, and allow for more automated software implementations to be developed, such as the routines currently available in the `mirt` package (Chalmers, 2012).

In the "Empirical Example" subsection above, test multidimensionality was appropriately accommodated in spite of the presence of a large amount of missing data, and a mixture of dichotomous and polytomous IRT models was fit simultaneously to the data. The EMEIRT model was able to provide a highly general representation of the test's underlying structure and design components, and the direct effect of external covariates were modeled to determine how the external covariants influenced the probability of endorsing specific item categories. As demonstrated in the simulated data example, model complexity can be reduced by

comparing nested models through interpreting frequentist statistics such as AIC, BIC, or likelihood-ratio tests, and, compared to general purpose GLMM software, estimation times for fitting IRT models with multidimensional structures may be dramatically decreased. Not explored in this article is the potential use of the observed information matrix for further linear hypothesis testing using the Wald (1943) method. The Wald approach may be useful for testing additional linear parameter hypotheses within larger and more complicated models because it does not require the estimation of a nested model; however, the effectiveness of this method requires further investigation within the EMEIRT model, and largely depends upon the quality of the estimated observed information matrix.

Although the outlined EMEIRT model appears to be very powerful and flexible, much more work needs to be investigated before the complete utility of this model and estimation approach can be concluded. For instance, throughout the simulations and examples the $\mathbf{V}$ and $\mathbf{W}$ matrices, which could be used to model multivariate latent regression and multilevel ability effects, were not explored. The fixed and random effects in the latent regression component ($\boldsymbol{\Gamma}$ and $\boldsymbol{\zeta}$) generate further complications in the model in that they are weighted by the slope parameters $\mathbf{a}_j$, therefore drawing random effects in $\boldsymbol{\zeta}$ and $\boldsymbol{\epsilon}$ must account for this information, and maximizing the $\boldsymbol{\Gamma}$ terms will require the additional weighting affect borne from simultaneously estimated components in $\mathbf{a}_j$. Furthermore, additional IRT parameters may be decomposed into fixed and random components, such as deconstructing the slope coefficients to explain differential variability across populations (e.g., see Cho et al., 2014), or potentially decomposing other model effects, such as the lower-bound parameters, to explain variation in other aspects of the model.

This paper outlined several applications of the EMEIRT model that can be realized by manipulating item-intercept designs, and demonstrated how the MH-RM algorithm appears to be a suitable estimation tool for managing a large number of fixed and random effects. More in-depth simulation work using the EMEIRT model with the MH-RM algorithm should be investigated to better understand how known population parameters and item properties will influence the precision and efficiency of the parameter estimates. Nevertheless, the currently available explanatory mixed-effects IRT models explored in this article are powerful tools for psychometricians to explore if the ultimate goal is to capture and explain information in psychological and educational tests, and the MH-RM algorithm with the EMEIRT model appears to be one such tool that can be used to help understand and statistically model these effects.

Notes

[1]A Rasch model parameterization in linear factor analysis is analogous to extracting a single factor, constraining all the factor loadings to be equal to a constant, and freely estimating the latent factor variance parameter.

[2]To simplify the presentation, we will only explore mixed-effects models that do not contain within-cluster correlation structures that are common in longitudinal models. We do this so that all random effect terms can be expressed by a single design matrix throughout.

[3]A similar line of reasoning exists for the graded response model (Samejima, 1969), as well as several other well-known IRT models, where the expected prob-

ability functions are modified to include information regarding the fixed and random effects while simultaneously preserving the traditional or slope-intercept parameterizations.

[4]Numerical complications can arise when applying the MH-RM with continuous covariates when the means of continuous predictor variables are too far away from zero (i.e., not mean-centered) or have high variability. The exponentiated terms $\mathbf{c}_j = (\mathbf{V}\boldsymbol{\gamma} + \mathbf{W}\boldsymbol{\zeta} + \boldsymbol{\epsilon})\mathbf{a}_j + \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\boldsymbol{\delta}$ in Equation 6 will only remain accurate when $\frac{\exp(35)}{1+\exp(35)} \geq P(1|z) \geq \frac{\exp(-35)}{1+\exp(-35)}$ due to limitations in numerical floating point storage. Therefore, to keep the magnitude of all elements in $\mathbf{c}_j$ less than 35, the predictor variables should be mean-centered and potentially rescaled (e.g., standardized) to lie between -10 to 10 or less to help avoid numerical underflow.

[5]A normal prior is appropriate for the g parameters because internally the parameters are transformed into $g' = \log(\frac{g}{1-g})$ for better numerical stability during estimation.

Appendix

Bias and Root Mean Squared Deviation (RMSD) for Simulated Data Example with 1,000 Replications

	Population		Bias		RMSD	
Item Number	a	g	a	g	a	g
1	0.6	.2	−.0019	.0005	.0951	.0076
2	1.1	.2	.0080	.0005	.1085	.0073
3	1.3	.2	.0151	.0004	.1126	.0074
4	0.8	.2	.0014	.0000	.0986	.0073
5	0.7	.2	−.0029	.0004	.0915	.0071
6	1.3	.2	.0089	.0004	.1247	.0072
7	1.8	.2	.0193	.0003	.1823	.0073
8	0.7	.2	−.0054	.0005	.0846	.0063
9	2.3	.2	.0321	.0006	.2521	.0074
10	1.6	.2	.0146	.0004	.1513	.0068
11	1.0	.2	.0040	.0005	.0993	.0076
12	0.9	.2	.0027	.0000	.1003	.0076
13	1.0	.2	.0048	.0004	.1013	.0076
14	0.9	.2	.0034	.0005	.0981	.0074
15	1.0	.2	.0076	−.0001	.1007	.0073

Design-Parameters	Population	Bias	RMSD
Item-design low	−1.5	−.0039	.0667
Item-design mid	0.0	.0037	.0629
Item-design high	1.5	.0048	.0550
Group two	3.0	.0045	.0897
Continuous predictor	0.0	.0004	.0300
Item-design mid × group two	−3.0	−.0080	.1095
Item-design high × group two	−6.0	−.0158	.1170

Acknowledgment

Special thanks to Jolynn Pek, David Flora, Victoria Ng, and four anonymous reviewers for providing insightful comments that improved the quality of this manuscript.

References

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

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4* [computer software manual]. Retrieved from http://CRAN.R-project.org/package=lme4 R package version 1.1-7

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, *75*, 33–57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*, 307–335.

Chalmers, R. P. (2012). MIRT: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. http://www.jstatsoft.org/v48/i06

Chalmers, R. P., & Flora, D. B. (2014). Maximum-likelihood estimation of noncompensatory IRT models with the MH-RM algorithm. *Applied Psychological Measurement*, *38*, 339–358.

Cho, S.-J., De Boeck, P., Embretson, S., & Rabe-Hesketh, S. (2014). Additive multilevel item structure models with random residuals: Item modeling for explanation and item generation. *Psychometrika*, *79*, 84–104.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*(12), 1–28.

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

Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement*, *23*, 283–298.

Embretson, S. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*, 407–433.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika*, *48*, 3–26.

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, *22*, 700–725.

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

Geerlings, H., & Glas, C. A. W. (2001). Modeling rule-based item generation. *Psychometrika*, *76*, 337–359.

Goldstein, H. (1979). Consequences of using the Rasch model for educational assessment. *British Educational Research Journal*, *5*, 211–220.

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika*, *76*, 537–549.

Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. *Computational Statistics and Data Analysis*, *52*, 5066–5074.

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

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

MacCallum, R. C., Kim, C., Malarkey, W. B., & Kiecolt-Glaser, J. K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research*, *32*, 215–253.

McCullach, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London, UK: Chapman & Hall.

McCulloch, C., & Searle, S. R. (2001). *Generalized, linear and mixed models*. New York, NY: John Wiley & Sons.

Monroe, S., & Cai, L. (2014). Estimation of a Ramsay-curve item response theory model by the Metropolis-Hastings Robbins-Monro algorithm. *Educational and Psychological Measurement*, *74*, 343–369.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York, NY: Springer.

R Core Team. (2014). *R: A language and environment for statistical computing* [computer software manual]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrics 17. *Psychometrika Monographs*, *34*(4).

SAS Institute Inc. (2010). *The SAS system for Windows.* Release 9.2 [computer software manual]. Cary, NC: SAS Institute. Retrieved from http://www.sas.com/

Sheu, C.-F., Chen, C.-T., Su, Y.-H., & Wang, W.-C. (2005). Using SAS PROC NLMIXED to fit item response theory models. *Behavior Research Methods*, *37*(2), 202–218.

Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Development and applications* (pp. 43–75). New York, NY: Taylor & Francis.

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

Whitely, S. E. (1977). Models, meaning and misunderstandings: Some issues in applying Rasch's theory. *Journal of Educational Measurement*, *14*, 227–235.

Whitely, S. E., & Dawis, R. V. (1974). The nature of the objectivity with the Rasch model. *Journal of Educational Measurement*, *11*, 163–178.

Author

R. PHILIP CHALMERS is a doctoral student, York University, 4700 Keele Street, Toronto, ON, M3J 1P3, Canada; rphilip.chalmers@gmail.com. His current research interests include modeling latent variables with factor analysis, structural equation, and item response theory techniques.