Journal of Educational and Behavioral Statistics
August 2012, Vol. 37, No. 4, pp. 518–542
DOI: 10.3102/1076998611417628

Profile-Likelihood Approach for Estimating Generalized Linear Mixed Models With Factor Structures

Minjeong Jeon
University of California, Berkeley

Sophia Rabe-Hesketh
University of California, Berkeley and
Institute of Education, University of London

In this article, the authors suggest a profile-likelihood approach for estimating complex models by maximum likelihood (ML) using standard software and minimal programming. The method works whenever setting some of the parameters of the model to known constants turns the model into a standard model. An important class of models that can be estimated this way is generalized linear mixed models with factor structures. Such models are useful in educational research, for example, for estimation of value-added teacher or school effects with persistence parameters and for analysis of large-scale assessment data using multilevel item response models with discrimination parameters. The authors describe the profile-likelihood approach, implement it in the R software, and apply the method to longitudinal data and binary item response data. Simulation studies and comparison with `gllamm` *show that the profile-likelihood method performs well in both types of applications. The authors also briefly discuss other types of models that can be estimated using the profile-likelihood idea.*

Keywords: *generalized linear mixed models; crossed random effects; factor structures; persistence parameters; ML estimation; item response theory (IRT); multilevel IRT;* `gllamm`*; R software*

1. Introduction

In this article, we suggest a profile-likelihood approach for estimating complex models by maximum likelihood (ML) using standard software and minimal programming. The method works whenever setting some of the parameters of the model to known constants turns the model into a standard model. An important class of models that can be estimated this way are generalized linear mixed models with factor structures.

Generalized linear mixed models, also known as multilevel or hierarchical generalized linear models (Goldstein, 2003; Raudenbush & Bryk, 2002), are

popular models for longitudinal data as well as cross-sectional data with units nested in clusters, the canonical example being students nested in schools. Models with crossed random effects (e.g., Goldstein, 1987; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Raudenbush, 1993) to handle data with two or more nonnested classifications such as schools and neighborhoods are also becoming increasingly popular. These models can be estimated in a wide range of general purpose and specialized software packages. In this article, we consider an extension of these models to include factor structures and propose a method for estimating the extended models by ML.

To define this extension, we start by considering a standard two-level generalized linear mixed model with M random effects. The linear predictor for unit i in cluster j can be written as follows:

$$g(\mu_{ij}) = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\boldsymbol{\delta}_j = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \sum_{m=1}^{M} \delta_{mj} z_{mij}, \tag{1}$$

where μ_{ij} is the conditional expectation of the response variable, $g(\cdot)$ is the link function, and $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij} = (z_{1ij}, z_{2ij}, \ldots, z_{Mij})'$ are covariate vectors with fixed coefficient $\boldsymbol{\beta}$ and random coefficients $\boldsymbol{\delta}_j = (\delta_{1ij}, \delta_{2ij}, \ldots, \delta_{Mij})'$, respectively. A simple application of this model would be a linear growth curve model for student achievement where the link function is the identity link and the covariates in both the fixed and the random parts of the model are a constant and time, $\mathbf{x}_{ij} = \mathbf{z}_{ij} = (1, t_{ij})'$. Then, the intercept β_0 (coefficient of the constant) and the coefficient β_1 of time represent mean initial achievement (when time is 0) and mean linear growth, respectively. Subject j's growth trajectory can differ from the mean growth trajectory by having its own intercept $\beta_0 + \delta_{0j}$ and coefficient $\beta_1 + \delta_{1j}$ of time. Variability in growth trajectories is captured by the covariance matrix of $(\delta_{0j}, \delta_{1j})$.

On the right-hand side of Equation (1), we see that δ_{mj} is the random coefficient of z_{mij} (or a random intercept if $z_{mij} = 1$). This part of the model is extended to include factor structures as follows:

$$g(\mu_{ij}) = \mathbf{x}_{ij}'\boldsymbol{\beta} + \sum_{m=1}^{M} \delta_{mj} \mathbf{w}_{mij}'\boldsymbol{\lambda}_m. \tag{2}$$

Here, $\mathbf{w}_{mij}$ is a vector of covariates and known constants associated with the random effect or latent variable δ_{mj} and $\boldsymbol{\lambda}_m$ is a vector of coefficients that we will refer to as factor loadings. As we will see in the applications, models with further levels of nested and/or crossed random effects (e.g., students nested within middle school and high school) with factor structures can also be estimated. When the factor loadings are set to known constants, we can define covariates $z_{mij} = \mathbf{w}_{mij}'\boldsymbol{\lambda}_m$ and the model becomes a standard generalized linear mixed model.

The reason for the name "factor loading" is that a confirmatory factor model can be specified as follows: The indicators or responses y_{ij} to different items i ($i = 1, \ldots, I$) that measure the common factors for persons j ($j = 1, \ldots, J$) are stacked into a single (unidimensional) response variable $\mathbf{y}$. The resulting data are treated as two-level data with item–person combinations ij at Level 1 nested in persons j at Level 2. The random effects or common factors δ_{mj} vary at Level 2, taking the same value for all responses by the same person. For the mth common factor δ_{mj}, the elements of $\mathbf{w}_{mij}$ are dummy variables for the items that load on that factor. For instance, if Items 1 through 3 load on the first factor δ_{1j}, then $\mathbf{w}_{1ij}$ is a three-dimensional vector, taking the values $(1, 0, 0)'$ when $i = 1$, $(0, 1, 0)'$ when $i = 2$, and $(0, 0, 1)'$ when $i = 3$. Letting $\boldsymbol{\lambda}_1 = (\lambda_{11}, \lambda_{12}, \lambda_{13})'$, we see that δ_{1j} is multiplied by $\mathbf{w}_{11j}'\boldsymbol{\lambda}_1 = \lambda_{11}$ for Item 1, $\mathbf{w}_{12j}'\boldsymbol{\lambda}_1 = \lambda_{12}$ for Item 2, and $\mathbf{w}_{13j}'\boldsymbol{\lambda}_1 = \lambda_{13}$ for Item 3 as required. A traditional factor model is specified by using an identity link and allowing the residual variance to differ between items. This method for incorporating factor structures in mixed models is part of the generalized linear latent and mixed model (GLLAMM) framework by Rabe-Hesketh, Skrondal, and Pickles (2004); see also Skrondal and Rabe-Hesketh (2004, 2007).

We have described how a two-level linear mixed model with factor structures can be used to specify a confirmatory factor model. By changing the link function to a logit link, we obtain an item response model, where the factor loadings are now called discrimination parameters. If persons are nested in clusters, such as schools, a three-level mixed model with factor structures can be used to specify a multilevel factor or item response model. Since factor structures are not usually accommodated in software for generalized linear mixed models, researchers wanting to use such software to estimate multilevel measurement models have had to set the factor loadings to known constants (e.g., Raudenbush, Rowan, & Kang, 1991; Raudenbush & Sampson, 1999) or specify one-parameter item response models, with discrimination parameters set to one (e.g., Kamata, 2001; Maier, 2001).

Three general software packages for multilevel and latent variable modeling allow ML estimation of mixed models with factor structures: `gllamm` (Rabe-Hesketh, Skrondal, & Pickles, 2004) in Stata (StataCorp, 2009), PROC NLMIXED in SAS (Wolfinger, 1999), and `Mplus` (Muthén & Muthén, 2008). However, none of these programs can fit models with crossed random effects. SAS PROC NLMIXED is limited to two hierarchical levels and `Mplus` is also limited to two levels unless the lowest level can be represented by different variables. The main contribution of our article therefore is to provide a method for ML estimation of models that are not available in standard software. An alternative is to use Bayesian methods as implemented in WinBUGS (Spiegelhalter, Thomas, Best, & Gilks, 1996) or in the multilevel modeling package MLwiN (Browne, 2009; Goldstein & Browne, 2005). However, switching to Bayesian methods for expediency alone is not always advisable since real differences exist between Bayesian and frequentist approaches. Bayesian methods also require

considerable expertise to specify appropriate priors and monitor convergence of the Markov chain. If vague priors are needed in order to obtain approximate ML estimates, it is known to be difficult to specify vague priors for variance–covariance parameters in mixed models (e.g., Browne & Draper, 2006; Lockwood, McCaffrey, Mariano, & Setodji, 2007; Natarajan & Kass, 2000).

We provide two examples handled by our approach. The first example is a random effects model for longitudinal data on students who are nested in middle schools for the first two occasions and high schools for the next two occasions where middle schools are cross-classified with high schools. The effects of middle schools and high schools on student outcomes are latent variables whose impacts change over time. The relative contribution of the school effects each year can be captured by associated factor loadings or persistence parameters. Such crossed random effects models with persistence parameters have been used for value-added assessments of school and teacher effects (e.g., McCaffrey et al., 2004). The second example is a multilevel two-parameter logistic (2PL) item response model for binary responses. In this example, the responses to the items in a test are viewed as nested within students (Adams, Wilson, & Wu, 1997; Mellenbergh, 1994). Since students are nested within schools, the resulting model is a three-level generalized linear mixed model with factor loadings or discrimination parameters. It is straightforward to incorporate covariates for students and schools. Such models are important for analyzing large-scale survey assessments such as the Programme for International Student Assessment (PISA) and the National Assessment of Educational Progress (NAEP) (e.g., Li, Oranje, & Jiang, 2009).

The outline of the article is as follows. In Section 2, we present our estimation approach in detail, including estimation of standard errors. The two examples follow in Section 3. For each example, a real dataset is analyzed and a small simulation study is performed. We briefly outline alternative models that can be handled by our approach in Section 4 and end with concluding remarks in Section 5. R code for the first application is presented in Appendix A (see the online Appendix, available at http://jeb.sagepub.com/supplemental).

2. Estimation

2.1. Profile-Likelihood Method

Suppose there are two sets of parameters, $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$, where $\boldsymbol{\lambda}$ represents the vector of factor loadings and $\boldsymbol{\beta}$ represents all other model parameters (regression coefficients and variance parameters). The likelihood $L(\boldsymbol{\lambda}, \boldsymbol{\beta})$ is then a function of $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$. We obtain the profile-likelihood function $L(\boldsymbol{\lambda})$ by replacing $\boldsymbol{\beta}$ by its ML estimate $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ at fixed values of $\boldsymbol{\lambda}$ (e.g., Pawitan, 2001, p. 62)

$$\begin{aligned} L(\boldsymbol{\lambda}) &= \max_{\boldsymbol{\beta}} L(\boldsymbol{\lambda}, \boldsymbol{\beta}) \\ &= L(\boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda})). \end{aligned}$$

Here,

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, L(\boldsymbol{\lambda}, \boldsymbol{\beta})$$

does not have a closed form but is found by an iterative maximization algorithm. Our method hinges on the observation that a mixed model with factor structures becomes a standard mixed model when $\boldsymbol{\lambda}$ is replaced by known values. Specifically, our method consists of two nested maximizations: $L(\boldsymbol{\lambda})$ is maximized with respect to $\boldsymbol{\lambda}$, where $L(\boldsymbol{\lambda})$ is itself obtained by maximizing $L(\boldsymbol{\lambda}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$.

We have written an R program to implement the nested maximizations described above. The `lmer` function in the R package `lme4` (Bates & Maechler, 2009) is used to find $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ by maximizing $L(\boldsymbol{\lambda}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ when $\boldsymbol{\lambda}$ is known. To maximize $L(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$, we use the limited memory quasi-Newton algorithm implemented in the R function `optim` (Byrd, Lu, Nocedal, & Zhu, 1995). The function computes the Hessian matrix numerically, requires only modest storage, and is relatively fast.

2.2. Standard Errors for $\widehat{\boldsymbol{\lambda}}$

The covariance matrix of the parameter estimates is usually estimated by the inverse of the observed Fisher information matrix (minus the Hessian of the log-likelihood function evaluated at the ML estimates).

The required information matrix $I(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\beta}})$ for the entire parameter vector $(\boldsymbol{\lambda}', \boldsymbol{\beta}')'$ can be partitioned as

$$I(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\beta}}) \equiv \begin{pmatrix} I_{(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\lambda}})} & I_{(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\beta}})} \\ I_{(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\lambda}})} & I_{(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}})} \end{pmatrix},$$

and its inverse can be written as

$$I^{-1}(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\beta}}) \equiv \begin{pmatrix} C_{(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\lambda}})} & C_{(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\beta}})} \\ C_{(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\lambda}})} & C_{(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}})} \end{pmatrix}.$$

Note that $C_{(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\lambda}})}$, the estimated covariance matrix of $\widehat{\boldsymbol{\lambda}}$, is not equal to the inverse $I^{-1}_{(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\lambda}})}$ of the corresponding part of the information matrix. The latter represents the estimated covariance matrix for $\widehat{\boldsymbol{\lambda}}$ if the parameters $\boldsymbol{\beta}$ were constrained equal to their estimates since in this case, $I_{(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\lambda}})}$ would be the full information matrix. It turns out that the inverse of minus the Hessian of the log profile-likelihood is equal to the required covariance matrix $C_{(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\lambda}})}$ that takes parameter uncertainty for $\boldsymbol{\beta}$ into account (see Pawitan, 2001, p. 63).

2.3. Standard Errors for $\widehat{\boldsymbol{\beta}}$

Unlike for $\widehat{\boldsymbol{\lambda}}$, the standard errors for $\widehat{\boldsymbol{\beta}}$ are not a by-product of the profile-likelihood approach. Because we assume known values of $\widehat{\boldsymbol{\lambda}}$ when finding $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\lambda}})$, the estimated covariance matrix obtained from `lmer` corresponds to $I^{-1}_{(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}})}$. We will call this the naive covariance matrix. The standard errors will be underestimated because uncertainty involved in estimating $\widehat{\boldsymbol{\lambda}}$ is not taken into account.

In this section, we discuss three ways of estimating adjusted standard errors for $\widehat{\boldsymbol{\beta}}$ taking into account the uncertainty regarding $\boldsymbol{\lambda}$. In the following, we let the dimensions of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ be denoted p and q, respectively.

2.3.1. Hessian Matrix

The most straightforward way to obtain the correct standard errors for $\widehat{\boldsymbol{\beta}}$ is using the Hessian matrix obtained by numerical differentiation. The Hessian consists of $(p+q)(p+q+1)/2$ second derivatives and each derivative requires multiple evaluations of the log likelihood. Unfortunately, it is not easy to obtain the Hessian in our implementation because `lmer` can, at the time of writing, not be used to just evaluate the log likelihood at a given parameter vector.

2.3.2. Likelihood Ratio

The second method is based on the fact that the difference in twice the log likelihood between two nested models approximately equals the corresponding Wald statistic. In models where the log likelihood is quadratic in the parameters, such as linear mixed models with known variance parameters, the approximation becomes exact.

Suppose L and L_0 are the log likelihoods of two nested models where the reduced model has a restriction on $\boldsymbol{\beta}$ such as $\beta_1 = \beta_1(0)$. The likelihood ratio statistic, $g^2 \equiv 2(L - L_0)$ is assumed to be approximately equal to the Wald statistic for the null hypothesis, $H_0 : \beta_1 = \beta_1(0)$ as

$$g^2 \approx \left[\frac{\widehat{\beta}_1 - \beta_1(0)}{SE(\widehat{\beta}_1)}\right]^2.$$

Then, the standard error $SE(\widehat{\beta}_1)$ is approximately

$$SE(\widehat{\beta}_1) \approx \frac{\widehat{\beta}_1 - \beta_1(0)}{g} \tag{3}$$

The quadratic approximation is likely to work better for smaller differences between the estimated and hypothesized values of β_1. We therefore suggest using the null value $\beta_1(0) = \widehat{\beta}_1 - d$, where d is some small value close to zero (e.g., 0.1 or 0.01).

A similar method was proposed by Miettinen (1976) who used any chi-square statistic (such as a Mantel–Haenszel statistic) for g in Equation (3). This approach was criticized by Halperin (1977) and Greenland (1984) because, for many parameters, the approximation works only when the null hypothesis is true. Here we assume, as is typically done in generalized linear mixed models, that the standard errors of regression coefficients do not depend on the true values of the coefficients (we will not apply this method to variance–covariance parameters). Also, by specifying a null value near the estimate (instead of a null hypothesis of "no association" as in Miettinen, 1976), we assume only that the log likelihood is quadratic near the mode, an assumption inherent in the estimation of asymptotic standard errors.

Notice that the likelihood ratio method requires p maximizations.

2.3.3. Delta Method

It follows from Parke (1986) that the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ is given by

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = I^{-1}_{(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}})} + \left(\frac{\partial \widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\right) C_{(\hat{\boldsymbol{\lambda}}\hat{\boldsymbol{\lambda}})} \left(\frac{\partial \widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\right)',$$

which is the naive covariance matrix treating $\boldsymbol{\lambda}$ as known plus a correction term, where $\dfrac{\partial \widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}$ is the Jacobian matrix of partial derivatives of $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ evaluated at $\widehat{\boldsymbol{\lambda}}$ and $C_{(\hat{\boldsymbol{\lambda}}\hat{\boldsymbol{\lambda}})}$ is the covariance matrix of $\widehat{\boldsymbol{\lambda}}$. We estimate this covariance matrix by plugging in estimates for $I^{-1}_{(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}})}$ and $C_{(\hat{\boldsymbol{\lambda}}\hat{\boldsymbol{\lambda}})}$, from `lmer` and `optim`, respectively, and estimating the elements of the Jacobian, evaluated at $\widehat{\boldsymbol{\lambda}}$, using

$$\frac{\partial \widehat{\beta}_r}{\partial \widehat{\lambda}_s} \approx \frac{\widehat{\beta}_r(\widehat{\lambda}_s + d, \widehat{\boldsymbol{\lambda}}_{-s}) - \widehat{\beta}_r(\widehat{\boldsymbol{\lambda}})}{d}. \tag{4}$$

Here, $\widehat{\beta}_r(\widehat{\lambda}_s + d, \widehat{\boldsymbol{\lambda}}_{-s})$ is the estimate of the rth element of $\boldsymbol{\beta}$ when a small constant d (close to zero) is added to the sth element λ_s of $\widehat{\boldsymbol{\lambda}}$ and all other elements of $\widehat{\boldsymbol{\lambda}}$ remain the same. Instead of a constant d, the step size is often chosen to be a fraction of λ_s, such as $\sqrt{\varepsilon}\lambda_s$, where ε is the machine precision.

This method requires q maximizations if we use $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\lambda}})$ and all elements of $\widehat{\boldsymbol{\beta}}(\widehat{\lambda}_s + d, \widehat{\boldsymbol{\lambda}}_{-s})$ to obtain an entire column of the Jacobian matrix for each maximization.

Both the likelihood ratio and delta methods produce similar results in similar computation time. They both also involve a choice of d in Equations (3) and (4) that could affect the results. For the delta method, we can avoid an arbitrary

choice of d using an established method for numerical derivatives (e.g., R package `numDeriv`) making that method easier to implement.

3. Applications

The profile-likelihood method allows us to estimate a wide range of generalized (crossed) random effects models with factor structures. In this section, we describe two applications of this method.

3.1. Crossed Random Effects Model With Persistence Parameters

The first example is a study of school effects using longitudinal data where students are in middle school during the first two waves and in high school in Waves 3 and 4. The response variable could be vertically scaled achievement scores or psychological scales such as self-esteem where the questions do not change over time. Figure 1 shows the structure of the data as a diagram similar to those suggested by Browne, Goldstein, and Rasbash (2001).

In the figure, rectangles represent units or clusters. A single arrow represents a nested relationship and unconnected rectangles at the same height represent a cross-classified relationship. The four repeated measures (Time 1 to Time 4) at the bottom level are nested within students. Students are nested both within middle school and high school but middle and high schools are crossed. The dashed arrows indicate that the first two measures (Time 1 and Time 2) are nested within middle school but the last two measures (Time 3 and Time 4) are nested within high school.

The required model is a crossed random effects model because students belong to a cross-classification of middle school and high school across time (e.g., see Goldstein, 1987; Raudenbush, 1993). We will fit the model

$$Y_{tsmh} = \beta_1 + \beta_2 \text{time2}_t + \beta_3 \text{time3}_t + \beta_4 \text{time4}_t + \delta_s + \delta_m \mu_t + \delta_h \eta_t + e_{tsmh}, \qquad (5)$$

where Y_{tsmh} is a continuous score at time t for student s who attended middle school m and high school h. β_1 is an intercept and β_2, β_3, and β_4 are coefficients for Time 2, Time 3, and Time 4 dummy variables. The random part of the model consists of a student-level random effect $\delta_s \sim N(0, \sigma_s^2)$, a middle school random effect $\delta_m \sim N(0, \sigma_m^2)$, a high school random effect $\delta_h \sim N(0, \sigma_h^2)$, and an occasion- and student-specific residual $e_{tsmh} \sim N(0, \sigma_e^2)$.

The model has occasion-specific parameters, $\boldsymbol{\mu} = (1, \mu_2, \mu_3, \mu_4)'$ and $\boldsymbol{\eta} = (0, 0, 1, \eta_4)'$, which represent the relative contribution of school effects on student outcomes at each time point. μ_1 and η_3 are set to one for model identification (since the middle and high school variances are free parameters) and η_1 and η_2 are set to zero because the future high school is assumed not to affect students while they are still in middle school. The model assumes that the effects of a school on the response variable at different times are perfectly correlated.

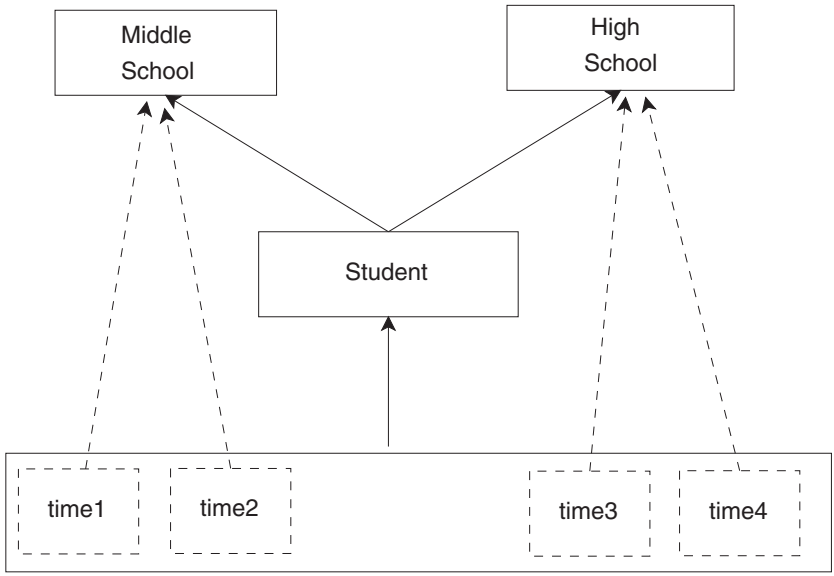

FIGURE 1. *Diagram for the longitudinal cross-classified data.*

Furthermore, the within-student covariance structure is exchangeable with constant variances and correlations.

Figure 2 represents the model by adapting the diagram notation of Rabe-Hesketh et al. (2004) to models with nonnested random effects. Circles represent latent variables and rectangles observed variables. The frames represent the levels represent at which variables within them vary. Variables located within the frame labeled "MS" vary between middle schools and variables located within the frame labeled "HS" vary between high schools. The latent variable δ_s for students is located in the intersection of these frames because subjects are cross-classified by middle schools and high schools. Arrows connecting latent and/or observed variables represent linear relations. We see that the middle school latent variable δ_m affects all four responses, whereas the high school latent variable affects only Responses 3 and 4. The short arrows pointing at the responses from below represent the residuals e_{tsmh}.

Using the framework of Equation (2), we can write the factor structures as

$$\delta_m \mu_t + \delta_h \eta_t = \delta_m \mathbf{d}_t' \boldsymbol{\mu} + \delta_h \mathbf{d}_t' \boldsymbol{\eta}, \tag{6}$$

where $\mathbf{d}_t$ is a four-dimensional vector of dummy variables with tth element equal to 1 and other elements equal to 0.

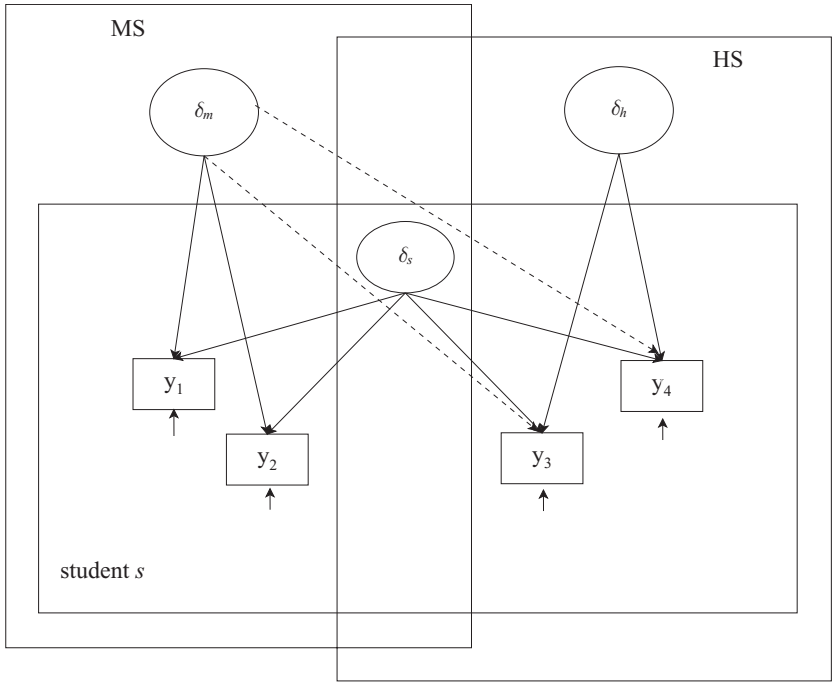

FIGURE 2. *Diagram for the crossed random effects model with persistence parameters.*

This model is similar to multiple membership models that include weights for school effects in proportion to the amount of time spent in each school (e.g., Goldstein, Burgess, & McConnell, 2007; Hill & Goldstein, 1998). Unlike multiple membership models, however, we treat the weights (or impacts) of school effects as parameters to be estimated.

McCaffrey et al. (2004) employed a similar model to investigate teacher contributions to student achievement. They considered situations where students were cross-classified by the different teachers who taught them in different grades. Teacher effects were modeled as random effects and the varying impacts of the teacher effects (called persistence parameters) were estimated by ML. They used an unstructured covariance matrix for the Level-1 residuals e_{tsmh}. While such a specification is currently not possible in `lmer`, we could specify random slopes for two functions of time, such as time and time squared, to use 7 free parameters for the 10 unique variances and covariances of the within-subject covariance matrix, as suggested by Maas and Snijders (2003). Mariano, McCaffrey, and Lockwood (2010) relaxed the assumption of perfect correlations among teacher effects across time by specifying an unstructured covariance matrix. Such a "generalized

persistence model'' can be fitted in `lmer`, but here we are interested in illustrating the profile-likelihood approach for a model with persistence parameters that cannot be fitted in `lmer`. The purpose-written ML estimation software for models with persistence parameters by McCaffrey et. al. (2004) was practically constrained to small datasets with continuous outcomes.

Because of the computational difficulty with ML estimation, Lockwood et al. (2007) suggested a Bayesian formulation of their earlier model (McCaffrey et al., 2004) and its extensions and used Markov chain Monte Carlo (MCMC). Lockwood et al. (2007) developed an MCMC algorithm in C since WinBUGS (Spiegelhalter et al., 1996) was prohibitively slow for their large datasets.

3.1.1. Empirical Study

We applied this model to the Korea Youth Panel Survey (KYPS) that sampled middle schools in the first stage and then randomly selected one class per school in the second stage. Students and parents were interviewed every year from 2003 to 2008. We analyzed data on 3,281 students observed in 104 middle schools at Waves 1 and 2 and 2,924 students followed up after dispersing into 860 high schools at Waves 3 and 4.

About 2.7% students switched their school membership during the middle school or high school years. Including them would necessitate making assumptions regarding the effects of the first and second middle schools in the second and subsequent waves of data and about the effects of the first and second high schools in the fourth wave of data. Since the portion of those students was small, we excluded them from the data for simplicity. In addition, 559 students with missing school identifiers were deleted. We handled the drop in student numbers after transition to high school by analyzing all available data under the missing at random (MAR) assumption (except for deleting 31 students who dropped out in Waves 2 and 4, although they could have easily been included). Missing school identifiers could alternatively be handled by assigning students to dummy schools as suggested by Lockwood et al. (2007).

There are an average of 34 students per middle school and 4 per high school, implying a highly sparse cross-classification between middle school and high school. The number of middle schools per high school ranges from 1 to 5, whereas the number of high schools per middle school ranges from 2 to 17. The response variable is self-esteem which is a mean-composite variable computed from six 5-point Likert-type scale items. The mean (standard deviation) of self-esteem is 3.16 (0.62), 3.26 (0.62), 3.31 (0.60), and 3.33 (0.61) at Waves 1, 2, 3, and 4, respectively. The internal consistency of the measures (Cronbach's α) is on average 0.734.

Table 1 lists the result of fitting the model in Equation (5) to the data. All parameters were estimated using the profile-likelihood method described in the

TABLE 1
Parameter Estimates and Standard Errors for Crossed Random Effects Model With Persistence Parameters

Parameters	Estimate	Standard Error
Fixed part		
β_1	3.162	0.015
β_2	0.109	0.012
β_3	0.159	0.016
β_4	0.176	0.018
Random part		
σ_s^2	0.152	—
σ_m^2	0.011	—
σ_h^2	0.008	—
σ_e^2	0.213	—
Factor loading		
μ_2	0.787	0.166
μ_3	0.003	0.288
μ_4	−0.180	0.352
η_4	1.509	0.285
Log likelihood	−9509.0	

previous section. We used the delta method described in Section 2.3 to estimate the adjusted standard errors for the regression coefficients $\widehat{\boldsymbol{\beta}}$. Differences between the delta method and likelihood ratio method for estimating standard errors were less than 10^{-6}. The naive, unadjusted standard errors from `lmer` were generally smaller than the adjusted standard errors as expected, but the difference was less than 10^{-5} in this application. As for computation time, it took about 160 seconds in total to fit the model and obtain the adjusted standard errors using the delta method on an Intel Pentium Dual-Core 2.5-GHz processor computer with 3.2 GB of memory.

In the fixed part, $\widehat{\beta}_1$ is the estimated mean self-esteem of students at Wave 1 (second grade in middle school). The coefficients for the time dummy variables, $\widehat{\beta}_2$, $\widehat{\beta}_3$, and $\widehat{\beta}_4$ represent the estimated differences in the mean self-esteem between each wave and Wave 1. Therefore, the mean growth Waves 1 to 2 is estimated as 0.11, the mean growth from Waves 2 to 3 is estimated as 0.05, whereas the mean growth from Waves 3 to 4 is estimated as 0.02. Self-esteem tends to increase, but the rate of growth decreases somewhat over time.

In the random part, we observe that the estimated within-student and between-student variation ($\widehat{\sigma}_e^2, \widehat{\sigma}_s^2$) is greater than the between-school variation ($\widehat{\sigma}_m^2, \widehat{\sigma}_h^2$). The estimated factor loadings suggest that school effects on the students' self-esteem change over time. To be specific, first recall that the middle school

loading at Wave 1 was set to one. Hence, the estimate $\widehat{\mu}_2 = 0.79$ implies a decline in the effect of middle school. However, comparing the estimate with its standard error, we see that the loading does not differ significantly from one at the 5% level. It does, however, differ significantly from zero (at the 5% level), so we can reject the hypothesis that the middle school effect vanishes in Wave 2. In contrast, the Waves 3 and 4 loadings do not differ significantly from zero, suggesting that middle school effects do not persist after students have moved into high school. For the high school random effect, the loadings for Waves 1 and 2 were set to zero and the loading for Wave 3 was set to one. The estimate $\widehat{\eta}_4 = 1.5$ suggests that high school effects increase over time, but the increase is not significant at the 5% level.

We do not provide the standard errors for variance parameters because the use of the standard errors for Wald-type tests and confidence intervals is inappropriate for these parameters (e.g., Berkhof & Snijders, 2001).

3.1.2. Simulation Study

Keeping the structure of the data as in the empirical application, we simulated new responses from the model fitted in the previous section with parameters set to the estimates in Table 1. This parametric bootstrapping procedure allows us to assess properties of the estimator. Table 2 summarizes the results for 100 replicates. The standard errors for the regression coefficients were obtained using the delta method.

The estimated bias ($\widehat{\text{Bias}}_B$) is negligible except for the factor loadings. Using one-sample t tests, we did not find that any of the bias estimates differed significantly from zero at the 5% level. The mean standard error estimates ($\overline{\widehat{SE}}$) were quite close to the standard deviations of the estimates or bootstrap standard errors ($\widehat{SE}_B$). These results suggest that the estimates are approximately unbiased and the estimated standard errors are approximately correct.

3.2. Multilevel 2PL Item Response Model

Another useful application is the multilevel 2PL item response model. One-parameter item response models can be viewed as two-level logistic regression models for binary responses Y_{ip} to item i by person p, nested in persons, where the person ability is a random intercept and item difficulties are regression coefficients of item dummy variables (e.g., Adams et al., 1997; Mellenbergh, 1994)

$$\text{logit}[\Pr(Y_{ip} = 1|\theta_p)] = \sum_{r=1}^{I} \beta_r d_{ri} + \theta_p.$$

Here d_{ri} is a dummy variable taking the value 1 when $r = i$ and 0 otherwise, $-\beta_i$ are item difficulties, and θ_p are person abilities. θ_p is a random effect (or latent variable) with $\theta_p \sim N(0, \sigma_p^2)$. Multilevel versions of this model, for students

TABLE 2
Results of the Simulation Study for Crossed Random Effects Model With Persistence Parameters Estimated Using the Profile-Likelihood Method

Parameters	θ	$\bar{\hat{\theta}}$	$\widehat{\text{Bias}}_B$	$\widehat{SE_B}$	$\overline{\widehat{SE}}$	$\widehat{\text{RMSE}}_B$
Fixed part						
β_1	3.162	3.160	−0.002	0.014	0.016	0.014
β_2	0.109	0.110	0.001	0.012	0.013	0.012
β_3	0.159	0.159	0.000	0.014	0.017	0.014
β_4	0.176	0.177	0.001	0.018	0.018	0.018
Random part						
σ_e^2	0.213	0.213	0.000	0.003	—	0.003
σ_s^2	0.152	0.152	0.000	0.006	—	0.006
σ_m^2	0.011	0.011	−0.000	0.003	—	0.003
σ_h^2	0.008	0.008	0.000	0.004	—	0.004
Factor loading						
μ_2	0.787	0.792	0.005	0.141	0.144	0.141
μ_3	0.003	0.024	0.021	0.163	0.155	0.165
μ_4	−0.180	−0.166	0.014	0.165	0.177	0.166
η_4	1.509	1.538	0.029	0.374	0.373	0.375

Note. θ are the true parameters, $\bar{\hat{\theta}}$ is the mean of the parameter estimates, $\widehat{\text{Bias}}_B$ is the estimated bias, $\widehat{SE_B}$ is the standard deviation of the parameter estimates, $\overline{\widehat{SE}}$ is the mean of the standard error estimates, and $\widehat{\text{RMSE}}_B$ is the root mean square error of the parameter estimates.

nested in schools, can be specified by simply adding higher-level random intercepts.

Unlike one-parameter models, two-parameter models are no longer standard generalized linear mixed models due to the item discrimination parameters multiplying the latent variable. The two-parameter model is usually written as

$$\text{logit}[\Pr(Y_{ip} = 1|\theta_p)] = \alpha_i(\theta_p - b_i) = \alpha_i\theta_p - \alpha_i b_i,$$

where α_i is the discrimination parameter and b_i is the difficulty parameter for item i. Defining intercept parameters as $\beta_i = -\alpha_i b_i$, the model can be written as a two-level logistic mixed model with a factor structure

$$\text{logit}[\Pr(Y_{ip} = 1|\theta_p)] = \sum_{r=1}^{I} \beta_r d_{ri} + \theta_p \sum_{r=1}^{I} \alpha_r d_{ri} = \mathbf{d}_i'\boldsymbol{\beta} + \theta_p \mathbf{d}_i'\boldsymbol{\alpha},$$

where $\mathbf{d}_i$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$ are vectors with elements d_{ri} $(r = 1, \ldots, I)$, β_i, and α_i $(i = 1, \ldots, I)$, respectively. For model identification, a factor loading (α_i) for one

item is typically constrained to one or the variance of the latent variable (σ_p^2) is constrained to one as in conventional factor models.

Multilevel versions of the two-parameter model for student p nested in school k have been suggested by Fox & Glas (2001) and Rabe-Hesketh et al. (2004). The model we will consider here can be written as

$$\begin{aligned} \text{logit}[\Pr(Y_{ipk} = 1|\theta_{pk})] &= \mathbf{d}_i'\boldsymbol{\beta} + \theta_{pk}\mathbf{d}_i'\boldsymbol{\alpha}, \\ \theta_{pk} &= \delta_{pk}^{(2)} + \delta_k^{(3)}, \end{aligned} \tag{7}$$

where k indicates schools, $\delta_{pk}^{(2)} \sim N(0, \sigma_p^2)$ is a person-level random intercept and $\delta_k^{(3)} \sim N(0, \sigma_s^2)$ is a school-level random intercept. The latter represents school mean ability, whereas the former represents the deviation of the student's ability from the school mean. Substituting the model for θ_{pk} into the model for y_{ipk}, we obtain a three-level logistic random intercept model with factor structures

$$\text{logit}[\Pr(Y_{ipk} = 1|\delta_{pk}^{(2)} + \delta_k^{(3)})] = \mathbf{d}_i'\boldsymbol{\beta} + \delta_{pk}^{(2)}\mathbf{d}_i'\boldsymbol{\alpha} + \delta_k^{(3)}\mathbf{d}_i'\boldsymbol{\alpha}. \tag{8}$$

The factor loadings ($\boldsymbol{\alpha}$) are assumed to be the same at Levels 2 and 3. For model identification, we can set the factor loading of one item to one or the variance of $\delta_{pk}^{(2)}$ to one.

Figure 3 shows a path diagram for the two-parameter multilevel item response model in Equation (8) using the path diagram conventions from Rabe-Hesketh et al. (2004). Here, the paths no longer represent linear relations and the short arrows represent Bernoulli variability instead of additive errors.

Multilevel two-parameter item response models have seldom been used due to computational obstacles. Standard software for item response models cannot handle latent variables at higher levels, whereas software for mixed models cannot estimate factor loadings or discrimination parameters. The general packages `Mplus` and `gllamm` (Zheng & Rabe-Hesketh, 2007) can be used to fit multilevel two-parameter item response theory (IRT) models as can various purpose-written programs including the Bayesian package `mlirt` in R (Fox, 2007).

In this section, we show how a multilevel two-parameter item response model can be fitted using the profile-likelihood method. Unlike linear mixed models for continuous responses, generalized linear mixed models for binary responses do not have a closed-form likelihood because the integrals over the random effects or latent variables are intractable. Whereas some software uses adaptive quadrature to evaluate the integrals numerically, the `lme4` package did not allow this option for models with more than two levels at the time of writing this article. Our profile-likelihood method therefore had to rely on the Laplace approximation of the integrals. This method is faster than numerical integration but is known to produce downward bias for variance parameters for binary data with small cluster sizes (here small number of items) and large variance components (Cho & Rabe-Hesketh, 2011; Joe, 2008). The `lme4` package is nevertheless a

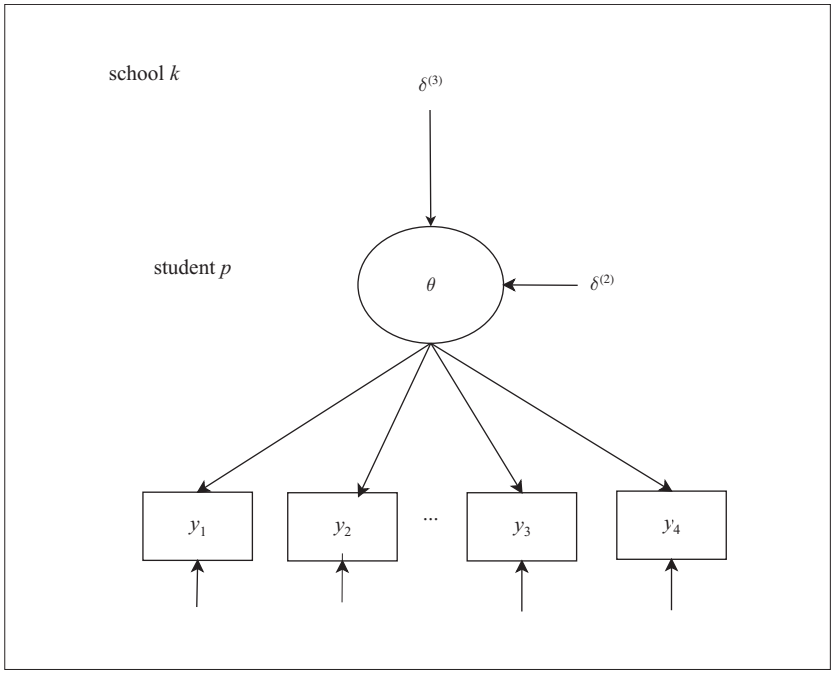

FIGURE 3. *Diagram for the multilevel two-parameter item response model.*

popular choice for item response modeling (e.g., De Boeck et al., 2011; Doran, Bates, Bliese, & Dowling, 2007).

As described in Section 4, more complex measurement models can also be estimated using the profile-likelihood approach.

3.2.1. Empirical Study

We analyzed data collected by Doolaard (1999), and previously analyzed by Cho and Rabe-Hesketh (2011), Fox and Glas (2001), and Vermunt (2007). (The data can be downloaded from http://www.statisticalinnovations.com/products/latentgold_datasets.html.) The data are from a Dutch primary school mathematics test that includes 18 items taken by 2,156 students who attended 97 schools in the Netherlands. Equation (8) was fitted using the profile-likelihood method. To estimate the adjusted standard errors for the fixed-effect parameters, the delta method was used. Differences between the delta and likelihood ratio methods were smaller than 0.001. The naive standard errors from `lmer` were about 0.1 to 0.3 smaller than the adjusted standard errors.

For comparison, the model was also fitted using `gllamm` whose adaptive quadrature method has been shown to be very accurate with a sufficient number

of quadrature points (Rabe-Hesketh, Skrondal, & Pickles, 2005). For this application, we used eight quadrature points at each level. Differences in the parameter estimates and standard errors with 6 and 9 quadrature points were at most 0.03. The computation time to fit this model using the profile-likelihood approach and obtain adjusted standard errors was about 3 days which was about the same as with `gllamm`. The time for a single maximization in `lmer` with known $\boldsymbol{\alpha}$ was 418 seconds. Table 3 shows the results.

The last columns of Table 3 show the relative difference for all parameters and standard errors defined as the estimate using profile-likelihood minus the estimate using `gllamm` divided by the estimate using `gllamm`. We see that agreement between the two methods is quite good both for the point estimates and for the standard errors.

3.2.2. Simulation Study

A small simulation study was carried out to evaluate the profile-likelihood method for Equation (8). We generated 50 datasets in which there are 10 dichotomous items for 1,000 students in 50 schools with 20 students per school. For each dataset, we fitted the model using both the profile-likelihood method and the `gllamm`. Table 4 shows the results.

The results for the profile-likelihood method and `gllamm` are remarkably similar. In both cases, the estimated bias ($\widehat{\text{Bias}}_B$) is quite small. Using one-sample t tests for each parameter, we did not find any bias to be significant at the 5% level except for α_5 ($t = -2.60$, $df = 49$, $p = 0.01$). The bias estimates using `gllamm` were very similar. The mean standard error estimates ($\overline{\widehat{SE}}$) were quite close to the standard deviations of the estimates ($\widehat{SE_B}$), considering that there are only 50 replicates. These results suggest that the estimates are approximately unbiased and the estimated standard errors are approximately correct.

4. Other Models That Can Be Estimated Using the Profile-Likelihood Approach

Any model that becomes a standard model when some parameters are set to known constants can be fitted using the profile-likelihood idea. Most of the models discussed in this section include factor structures and become standard generalized mixed models when the factor loadings are known.

An obvious extension of the multilevel two-parameter item response model discussed in Section 3 is to include more hierarchical levels and cross-classified levels, such as students nested in schools cross-classified with neighborhoods. To our knowledge, such a model has not been fitted before. It would also be straightforward to estimate *multidimensional* multilevel item response models (see, e.g., Goldstein, Bonnet, & Rocher, 2007) as outlined for the single-level case in the introduction. An example would be a bifactor model (Cai, Yang, & Hansen,

TABLE 3
Parameter Estimates and Standard Errors for Multilevel Two-Parameter Item Response Model Using `gllamm` *and the Profile-Likelihood Method*

	`gllamm` Eight Quadrature Points		Profile-Likelihood With Laplace		Relative Difference	
Parameters	$\widehat{\theta}_g$	$\widehat{SE}_g$	$\widehat{\theta}_l$	$\widehat{SE}_l$	$\frac{\widehat{\theta}_g - \widehat{\theta}_l}{\widehat{\theta}_g}$	$\frac{\widehat{SE_g} - \widehat{SE_l}}{\widehat{SE}_g}$
Fixed part						
β_1	−0.541	0.097	−0.545	0.099	0.001	−0.002
β_2	−1.438	0.103	−1.441	0.111	0.000	−0.008
β_3	−0.185	0.087	−0.188	0.088	0.003	−0.002
β_4	−1.634	0.098	−1.635	0.107	0.000	−0.012
β_1	−0.885	0.067	−0.886	0.068	0.001	−0.004
β_6	0.064	0.069	0.061	0.069	0.004	−0.001
β_7	−0.723	0.086	−0.726	0.088	0.001	−0.011
β_8	−2.300	0.098	−2.297	0.116	0.000	0.000
β_9	−0.448	0.077	−0.451	0.078	0.002	−0.006
β_{10}	−1.293	0.078	−1.294	0.082	0.000	−0.002
β_{11}	−1.391	0.084	−1.392	0.090	0.000	−0.002
β_{12}	0.007	0.062	0.005	0.062	0.065	−0.003
β_{13}	−2.941	0.112	−2.931	0.143	0.000	−0.002
β_{14}	−1.834	0.121	−1.838	0.136	0.000	−0.013
β_{15}	−1.168	0.091	−1.171	0.095	0.001	−0.010
β_{16}	−0.874	0.070	−0.875	0.072	0.001	−0.002
β_{17}	−2.561	0.113	−2.555	0.139	0.000	−0.010
β_{18}	−0.462	0.083	−0.464	0.084	0.000	−0.004
Random part						
σ_p	1.259	—	1.283	—	−0.019	—
σ_s	0.756	—	0.767	—	−0.015	—
Factor loading						
α_2	1.031	0.092	1.027	0.088	0.000	0.009
α_3	0.873	0.081	0.872	0.073	0.001	0.005
α_4	0.937	0.088	0.931	0.082	0.000	0.001
α_5	0.551	0.061	0.546	0.053	0.001	0.005
α_6	0.622	0.061	0.616	0.055	0.001	0.001
α_7	0.835	0.079	0.832	0.072	0.001	0.012
α_8	0.842	0.088	0.830	0.084	0.000	0.009
α_9	0.726	0.067	0.722	0.063	0.000	0.001
α_{10}	0.687	0.071	0.681	0.064	0.000	0.013
α_{11}	0.768	0.073	0.762	0.070	0.001	0.006
α_{12}	0.513	0.053	0.506	0.049	0.001	0.002
α_{13}	0.879	0.096	0.860	0.096	0.000	0.002

(continued)

TABLE 3 *(continued)*

	gllamm Eight Quadrature Points		Profile-Likelihood With Laplace		Relative Difference	
Parameters	$\widehat{\theta}_g$	$\widehat{SE}_g$	$\widehat{\theta}_l$	$\widehat{SE}_l$	$\frac{\widehat{\theta}_g - \widehat{\theta}_l}{\widehat{\theta}_g}$	$\frac{\widehat{SE_g} - \widehat{SE_l}}{\widehat{SE_g}}$
α_{14}	1.237	0.118	1.233	0.108	0.000	0.003
α_{15}	0.881	0.082	0.877	0.076	0.001	0.004
α_{16}	0.603	0.060	0.597	0.057	0.000	0.002
α_{17}	1.024	0.109	1.009	0.099	0.000	0.001
α_{18}	0.811	0.076	0.809	0.069	0.001	0.009
Log likelihood	−20090.9		−20071.8		0.001	

Note. $\widehat{\theta}_g$ and $\widehat{SE}_g$ are the parameter estimates and standard errors obtained from `gllamm` and θ_1 and $\widehat{SE}_l$ are the parameter estimates and standard errors obtained from the profile-likelihood method. $\frac{\widehat{\theta}_g - \widehat{\theta}_l}{\widehat{\theta}_g}$ and $\frac{\widehat{SE_g} - \widehat{SE_l}}{\widehat{SE_g}}$ are the relative difference of the parameter estimates and standard errors between `gllamm` and the profile-likelihood method.

2011; Gibbons & Hedeker, 1992; Jeon, Rijmen, & Rabe-Hesketh, in press) where the entire test can be viewed as Level 3, whereas the item bundles or testlets are at Level 2, and the items are at Level 1.

Covariates could be added to the measurement model in several ways. Adding person or school covariates $\mathbf{x}_{pk}$ to the latent regression in Equation (7)

$$\theta_{pk} = \mathbf{x}_{pk}'\boldsymbol{\gamma} + \delta_{pk}^{(2)} + \delta_k^{(3)},$$

with regression coefficients $\boldsymbol{\gamma}$, and substituting this model into the response model, we obtain

$$\text{logit}[\Pr(Y_{ipk} = 1|\theta_{pk})] = \mathbf{d}_i'\boldsymbol{\beta} + \mathbf{x}_{pk}'\boldsymbol{\gamma}\mathbf{d}_i'\boldsymbol{\alpha} + \delta_{\text{pk}}^{(2)}\mathbf{d}_i'\boldsymbol{\alpha} + \delta_{\text{k}}^{(3)}\mathbf{d}_i'\boldsymbol{\alpha}. \quad (9)$$

When the factor loadings or discrimination parameters $\boldsymbol{\alpha}$ are known, $\mathbf{d}_i'\boldsymbol{\alpha}$ can be treated as a covariate that multiplies each of the covariates in $\mathbf{x}_{pk}$. Such a model was recently discussed by Li, Oranje, and Jiang (2009).

Interactions between person covariates and item dummies can be added to the model to accommodate and test for differential item functioning (DIF); if such an interaction is included in the $\mathbf{d}_i$ vector in Equation (9), we obtain non-uniform DIF (see e.g., Swaminathan & Rogers, 1990).

We can also structure the difficulty parameters as a linear combination of item covariates (LLTM; De Boeck & Wilson, 2004; Fisher, 1983). When there is no discrimination parameter, the model is a standard generalized linear mixed model. Structuring the discrimination parameters this way in addition to the

TABLE 4
Summary of the Simulation Study for Multilevel Two-Parameter Item Response Model Using the Profile-Likelihood Method and `gllamm`

		Profile	`gllamm`	Profile	`gllamm`	Profile	
Parameters	θ	$\widehat{\text{Bias}}_B$	$\widehat{\text{Bias}}_B$	$\widehat{\text{RMSE}}_B$	$\widehat{\text{RMSE}}_B$	$\widehat{SE}_B$	$\overline{\widehat{SE}}$
Fixed part							
β_1	0.0	−0.008	−0.007	0.119	0.119	0.119	0.096
β_2	0.5	0.024	0.023	0.086	0.086	0.083	0.092
β_3	1.0	−0.022	−0.025	0.110	0.110	0.108	0.088
β_4	−0.5	−0.010	−0.008	0.094	0.094	0.094	0.081
β_5	0.5	0.026	0.027	0.123	0.124	0.121	0.098
β_6	0.0	0.016	0.017	0.120	0.121	0.119	0.103
β_7	−0.5	0.024	0.021	0.125	0.126	0.123	0.109
β_8	1.5	0.014	0.004	0.130	0.127	0.129	0.105
β_9	−1.5	0.000	0.013	0.143	0.141	0.143	0.103
β_{10}	0.0	0.006	0.006	0.098	0.098	0.098	0.093
Random part							
σ_p	1.0	0.043	0.107	0.222	0.262	0.218	—
σ_s	0.25	0.009	0.015	0.093	0.097	0.092	—
Factor loading							
α_2	0.9	0.016	0.009	0.163	0.170	0.163	0.149
α_3	0.8	−0.019	−0.036	0.118	0.124	0.116	0.136
α_4	0.7	−0.009	−0.022	0.100	0.105	0.100	0.124
α_5	1.1	−0.057	−0.057	0.164	0.173	0.154	0.171
α_6	1.2	−0.043	−0.031	0.161	0.170	0.156	0.180
α_7	1.3	−0.032	−0.013	0.209	0.225	0.207	0.198
α_8	1.0	0.017	−0.009	0.151	0.152	0.150	0.185
α_9	1.0	−0.021	−0.051	0.122	0.132	0.120	0.165
α_{10}	1.0	−0.025	−0.026	0.142	0.150	0.140	0.162

Note. θ are the true parameters, $\widehat{\text{Bias}}_B$ is the estimated bias, $\widehat{\text{RMSE}}_B$ is the root mean square error of the parameter estimates, $\widehat{SE}_B$ is the standard deviation of the parameter estimates, and $\overline{\widehat{SE}}$ is the mean of the standard error estimates.

difficulty parameters, as suggested by Embretson (1999), we obtain (in the single-level case)

$$\text{logit}[\Pr(Y_{ip} = 1|\theta_p)] = \mathbf{x}_i'\boldsymbol{\alpha}(\theta_p - \mathbf{x}_i'\boldsymbol{\beta}).$$

The first term, $\theta_p\mathbf{x}'_i\boldsymbol{\alpha}$, is a standard factor structure (but for the first time in this article, the variables multiplying the factor loadings are not dummy variables) and the second term is analogous to the second term in Equation (9)

and can thus be handled by multiplying each element of $\mathbf{x}_i$ by the scalar $\mathbf{x}_i'\boldsymbol{\alpha}$ for known $\boldsymbol{\alpha}$.

As discussed by De Boeck et al. (2011), models for polytomous responses with a continuation ratio logit link can be estimated by software for binary responses such as `lmer` by expanding the data appropriately.

For longitudinal data, Meredith and Tisak (1988) model nonlinear growth by specifying a growth curve model where the random coefficient does not multiply the known times associated with the measurement occasions, as in linear growth curve models, but unknown factor loadings to be estimated. The model has a similar structure to our model with persistence parameters, but it also includes a random intercept. For identification, two of the factor loadings are typically set to zero and one.

Factor loadings can also be used to relax the homoscedasticity assumption for random intercepts and random coefficients. For example, to let the variance of the random coefficient δ_{1j} of a covariate z_{ij} differ between males and females, specify $\mathbf{w}_{ij} = (d_{ij}z_{ij}, (1 - d_{ij})z_{ij})'$, where d_{ij} is a dummy variable for being male. Then include the term δ_{1j},$\mathbf{w}'_{ij}$,$\boldsymbol{\lambda}$ in the linear predictor. For males, δ_{1j} is then multiplied by $\mathbf{w}_{ij}'\boldsymbol{\lambda} = \lambda_1 z_{ij}$ and for females δ_{1j} is multiplied by $\mathbf{w}_{ij}'\boldsymbol{\lambda} = \lambda_2 z_{ij}$. If δ_j has variance σ^2, then $\sigma^2\lambda_m^2$ and $\sigma^2\lambda_m^2$ can be interpreted as the variances of the random coefficient for males and females, respectively. One of the factor loadings could be set to one for identification.

Factor structures are not the only model extension that can be handled using the profile-likelihood approach. We could also estimate extensions of generalized linear (mixed) models with nonlinear terms such as $\beta_1 x_i^{\lambda}$ or $\beta_1 \exp(x_i + \lambda z_i)$ in the linear predictor. Or we could have products of regression coefficients as in the stereotype model where

$$\log\frac{\Pr(y_i = s)}{\Pr(y_i = 1)} = \alpha_s + \lambda_s(\mathbf{x}_i'\boldsymbol{\beta}).$$

When the λ_s are known, this becomes a standard conditional logistic regression model with category-specific intercept and covariates $\lambda_s\mathbf{x}_i$.

The above list of models that can be estimated using the profile-likelihood method clearly does not exhaust all possibilities but should give a flavor of the power of this approach.

5. Concluding Remarks

In this article, we have developed a simple approach for estimating complex models by ML using standard software and minimal programming. The method works whenever setting some of the parameters of the model to known constants turns the model into a standard model. An important class of models that can be estimated this way are generalized linear mixed models with factor loadings. Such models include random effects or latent variables weighted by some

unknown parameters which are called factor loadings, persistence parameters, or discrimination parameters depending on the context.

We have described two applications in this article. Crossed random effects models with persistence parameters are useful for the assessment of value-added effects for teachers and schools. Multilevel two-parameter item response models are important for analysis of large-scale assessment studies such as NAEP and PISA.

We implemented the profile-likelihood method using the `lme4` package in R. Readers who wish to implement the methods may find it useful to refer to the code for the crossed random effects models with persistence parameters (Section 3.1) that is provided in Appendix A (see the online Appendix, available at http://jeb.sagepub.com/supplemental).

Our implementation of the profile-likelihood method shares advantages as well as disadvantages of `lme4`. For instance, we are able to handle an arbitrary number of levels, nested and fully or partially crossed structures, and different response types. For discrete data, however, the method may in some cases produce biased estimates for variance parameters since the Laplace approximation method is used for estimation. However, simulations and comparison with other software suggest that estimation of parameters and standard errors performs well in the types of applications considered here.

The profile-likelihood method could also be implemented in other programming environments such as `SAS` or `Stata` that provide a method for fitting generalized mixed models and a function for optimization.

Finally, as described in Section 4, we emphasize that the possible applications go further than the supplied cases. Much broader classes of models can be estimated using the proposed profile-likelihood approach.

References

Adams, R., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22*, 47–76.

Bates, D., & Maechler, M. (2009). *lme4: Linear mixed-effects models using S4 classes R package version* 0.999375-31. Retrieved from http://CRAN.Rproject.org/package=lme4

Berkhof, J., & Snijders, T. A. B. (2001). Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics*, *26*, 132–152.

Browne, W. J. (2009). *MCMC estimation in MLwiN.* Bristol, UK: Centre for Multilevel Modelling.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood methods for fitting multilevel models. *Bayesian Analysis*, *1*, 473–514.

Browne, W. J., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling*, *1*, 103–124.

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*, 1190–1208.

Cai, L., Yang, J. S., & Hansen, M. P. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*, 221–248.

Cho, S.-J., & Rabe-Hesketh, S. (2011). Alternating imputation posterior estimation of models with crossed random effects. *Computational Statistics and Data Analysis*, *55*, 12–25.

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*, 1–28.

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

Doolaard, S. (1999). *Schools in change or school in chain.* Enschede, Netherlands: University of Twente.

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the `lme4` package. *Journal of Statistical Software*, *20*, 1–18.

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

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

Fox, J. P. (2007). Multilevel IRT modeling in practice with the package `mlirt`. *Journal of Statistical Software*, *20*, 1–16.

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

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436.

Goldstein, H. (1987). Multilevel covariance component models. *Biometrika*, *74*, 430–431.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London, England: Arnold.

Goldstein, H., Bonnet, G., & Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioral Statistics*, *32*, 252–286.

Goldstein, H., & Browne, J. W. (2005). Multilevel factor analysis models for continuous and discrete data. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 453–475). Mahwah, New Jersey:Lawrence Erlbaum.

Goldstein, H., Burgess, S., & McConnell, B. (2007). Modelling the effect of pupil mobility on school differences in educational achievement. *Journal of the Royal Statistical Society Series A*, *170*, 841–954.

Greenland, S. (1984). A counterexample to the test-based principle of setting confidence limits. *American Journal of Epidemiology*, *120*, 4–7.

Halperin, M. (1977). Estimability and estimation in case-referent studies. Letters to the Editor. *American Journal of Epidemiology*, *105*, 496–498.

Hill, P., & Goldstein, H. (1998). Multilevel modeling of educational data with cross-classification and missing identification for units. *Journal of Educational and Behavioral Statistics*, *23*, 117–128.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (in press). Modeling differential item functioning using a generalized multiple-group bifactor model. *Journal of Educational and Behavioral Statistics*.

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

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Applied Psychological Measurement*, *38*, 79–93.

Li, D., Oranje, A., & Jiang, Y. (2009). On the estimation of hierarchical latent regression models for large-scale assessments. *Journal of Educational and Behavioral Statistics*, *34*, 433–463.

Lockwood, J., McCaffrey, D., Mariano, L., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics*, *32*, 125–150.

Maas, C. M., & Snijders, T. A. B. (2003). The multilevel approach to repeated measures for complete and incomplete data. *Quality & Quantity*, *37*, 71–89.

Maier, K. S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics*, *26*, 307–330.

Mariano, L., McCaffrey, D., & Lockwood, J. (2010). A model for teacher effects from longitudinal data without assuming vertical scaling. *Journal of Educational and Behavioral Statistics*, *35*, 253–279.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*, 67–101.

Mellenbergh, J. G. (1994). Generalized linear item response theory. *Psychological Bulletin*, *115*, 300–307.

Meredith, W., & Tisak, J. (1988). Latent curve analysis. *Psychometrika*, *55*, 107–122.

Miettinen, O. (1976). Estimability and estimation in case-referent studies. *American Journal of Epidemiology*, *103*, 226–235.

Muthén, L., & Muthén, B. (2008). *Mplus user's guide.* Los Angeles, CA: Muthen & Muthen.

Natarajan, R., & Kass, R. E. (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, *95*, 227–237.

Parke, R. W (1986). Pseudo maximum likelihood estimation: The asymptotic distribution. *Annals of Statistics*, *14*, 355–357.

Pawitan, Y. (2001). *In all likelihood: Statistical modeling and inference using likelihood.* New York, NY: Oxford.

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

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, *128*, 301–323.

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics*, *18*, 321–349.

Raudenbush, S. W., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage.

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational Statistics*, *16*, 295–330.

Raudenbush, S. W., & Sampson, R. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology*, *29*, 1–41.

Skrondal, A., & Rabe-Hesketh, S. W. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* Boca Raton, FL: Chapman & Hall/CRC.

Skrondal, A., & Rabe-Hesketh, S. W. (2007). Latent variable modelling: A survey. *Scandinavian Journal of Statistics*, *34*, 712–745.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1996). *BUGS 0.5 Bayesian analysis using Gibbs sampling. Manual (version ii).* Cambridge, UK: MRC-Biostatistics Unit. Retrieved from http://www.mrc-bsu.cam.ac.uk/bugs/documentation/contents.shtml

StataCorp. (2009). *Stata statistical software: Release 11.* College Station, TX: Author.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361–370.

Vermunt, J. (2007). Multilevel latent variable modeling: An application in education testing. *Austrian Journal of Statistics*, *38*, 285–299.

Wolfinger, R. D. (1999). *Fitting non-linear mixed models with the new NLMIXED procedure (Technical Report).* Cary, NC: SAS Institute.

Zheng, X., & Rabe-Hesketh, S. (2007). Estimating parameters of dichotomous and ordinal item response models with `gllamm`. *The Stata Journal*, *7*, 313–333.

Authors

MINJEONG JEON is a Graduate Student at the Graduate School of Education, University of California, Berkeley, CA 94720; mjj@berkeley.edu. Her primary research interests include item response modeling, multilevel modeling, and research methods.

SOPHIA RABE-HESKETH is a Professor at the Graduate School of Education and Graduate Group in Biostatistics, University of California, Berkeley, CA 94720, and at the Institute of Education, University of London; sophiarh@berkeley.edu. Her primary research interests include multilevel and latent variable modeling.

Manuscript received December 16, 2010
Revision received April 15, 2011
Accepted May 29, 2011