An Aggregate IRT Procedure for Exploratory Factor Analysis

Gregory Camilli
Rutgers, The State University of New Jersey

Jean-Paul Fox
University of Twente

An aggregation strategy is proposed to potentially address practical limitations related to computing resources for two-level multidimensional item response theory (MIRT) models with large data sets. The aggregate model is derived by integration of the normal ogive model, and an adaptation of the stochastic approximation expectation maximization algorithm is used for estimation. This methodology is used to conduct an exploratory factor analysis of the 2007 mathematics data from Trends in International Mathematics and Science Study (TIMSS) fourth grade to illustrate potential uses. A comparison to flexMIRT and two brief simulations indicate the aggregate model provides accurate estimates of Level 2 parameters despite loss of information ensuing from key assumption.

Keywords: multilevel model; IRT; MIRT; data aggregation; exploratory factor analysis; TIMSS mathematics

In this article, a method for investigating higher level effects is investigated. Specifically, an aggregate two-level multidimensional item response theory (MIRT) model is proposed to address a number of practical limitations related to computing resources in estimating factor structure and scoring of higher order units in sparse data sets. After description of the model, an empirical application is carried out on a large-scale data set from Trends in International Mathematics and Science Study (TIMSS) 2007 fourth-grade mathematics. The results are compared to that of a full three-level model (items nested within students nested within clusters) estimated with the Metropolis-Hastings-Robbins-Monro (MHRM) algorithm. Two brief but realistic simulations are also undertaken to investigate the accuracy of the estimation procedure. This work is related to recent developments in multilevel item response theory (IRT) models (e.g., L. Cai, 2010; Johnson & Jenkins, 2005; Johnson, Sinharay, & Bradlow, 2007) and to previous work in the use of multilevel IRT for diagnostic purposes (Camilli et al., 2008; Mislevy & Bock, 1989; Park & Bolt, 2008).
In many large-scale assessment programs, a single number-correct score for reading or mathematics is reported. However, there is no reason to believe that a single global measure is sensitive to school effects. The case for a multidimensional profiling of schools is based on the expectation that students within a school develop both specific skills at a particular grade in addition to the ability that develops over time. In a comprehensive achievement test, certain sets of items may be more relevant to some skills than others and consequently may be more useful in discovering school influences on student outcomes. For example, the National Assessment of Educational Progress (NAEP) incorporates numerous items measuring proficiency in the five key areas of number properties and operations; measurement; geometry; data analysis, statistics, and probability; and algebra (National Assessment Governing Board, 2012). Based on the notion that different subsets of items may be sensitive to the strengths and weaknesses of different schools, this provides an ideal opportunity to exploit item multidimensionality.

While traditional assessment programs disaggregate item performance by subtest, there are two key problems with this approach to providing diagnostic information. First, many assessments are developed in a way that are not conducive to producing diagnostic information. Stone, Zhu, and Lane (2010) demonstrated that with a test developed by IRT methods, the item structure is essentially unidimensional. In this case, subtest scores based on test specifications may have no diagnostic value beyond a total test score. No scoring method can overcome this obstacle. However, a different possibility is that even in such tests, some degree of multidimensionality is present. The difficulty arises from classifying items discretely into subscore domains whereas, in reality, any given item may tap a number of dimensions. Unrecognized combinations of dimensions required for an item response may predispose subscores to be highly correlated. Despite the cause of correlation among subtest scores, the effect is a familiar issue. For example, TIMSS scores at the jurisdiction level (e.g., country) are reported for the content domains number, geometric shapes and measures, and data display. The domain scores are correlated at $r \approx .96$ (Mullis, Martin, Foy, & Arora, 2012).

Item dimensionality may be considered either from a constructivist or empirical perspective. According to the constructivist approach, content classes such as “fractions and decimals” are created logically in the minds of mathematicians, mathematics educators, and test developers. Strands of problem type (e.g., algebra or geometry) might be based on research about how students learn mathematics or the structure of mathematics per se. Yet such nominal item classifications may not provide adequate profile scores for understanding the effects of instruction or learning (Clopton, Bishop, & Klein, 1997; Hiebert et al., 2005).

From the empirical perspective, nominal item groupings may not be consistent with empirical item factor structures at the individual level, and
dimensionality may not be the same at the individual and cluster levels. As noted by B. O. Muthén (1991):

[I]t is . . . of interest to separate within-class and between-class variation of student achievement, to relate between-class achievement variation to class-level information on teacher and teaching characteristics, and to contrast different educational systems. (p. 339)

This discussion leads to the second problem in providing diagnostic information: Traditional subtest scores may have inadequate empirical variance across the units to be provided with diagnostic profiles. More specifically, traditional subtest scores may contribute little empirical variance across units after conditioning on the total test score.

The purpose of this article is to demonstrate an aggregate approach to two-level factor analysis in large-scale data sets for discovering and estimating scores for higher order (or Level 2) factor structures. Although school is a Level-2 unit of particular interest, the method can apply to any high-order unit (e.g., class or district), which will be designated simply as the cluster level in this article. The approach to factor analysis technique is exploratory rather than confirmatory. Factor scores are subsequently estimated and examined for clusters. The procedure is relevant to situations in which there are many clusters in a data set, with a modest number of observations per cluster. In large-scale assessments, for example, there are thousands of schools with students nested within schools. Aggregation of item responses within schools offers a means to increase computational efficiency. An example of an operational program that used the idea of aggregation was the California Assessment program (CAP) of the late 1970s and early 1980s. In CAP, a school level (Level 2) was expressly model for finer grained, more sensitive, inferences about curricular effects (Bock & Mislevy, 1981; Bock, Mislevy, & Woodson, 1982). The focus in the next sections is on estimation and model description. Ensuing sections provide empirical demonstrations followed by two simulation studies.

Key assumptions of this approach at Level 1 are that item responses are unidimensional or nearly so, and that item parameters \((a, b, c)\) are available, which is typically the case in large-scale assessment programs. In response to the first assumption, it might be argued that Level-1 IRT parameters are in fact multidimensional. While the aggregate model proposed in this article can be designed to be consistent with either unidimensional or multidimensional calibration, Level-1 item responses in many large-scale tests are characterized by only one dominant dimension. The simulation study reported below suggests that individual-level multidimensionality does not strongly affect the estimation of Level-2 parameters, at least in the TIMSS data to which the aggregate model is applied. In regard to the second assumption, Level-1 item parameters can be estimated in a unidimensional framework if they are not available. This is less desirable than
using operational parameters because item calibrations in large-scale programs such as TIMSS and NAEP are performed with expertise and tools that would be difficult for others to implement. In any case, some literature (i.e., NAEP, 2008) suggests that performance at the student level does not depend on cluster membership, so that Level-1 and Level-2 parameters can be estimated independently.

Parameter Estimation in MIRT Models

Bock and Aitken (1981) introduced full information item factor analysis (FI-IFA) via marginal maximum likelihood (MML) estimation, an approach that has continued to have important applications (L. Cai, 2010). This approach to exploratory factor analysis (EFA) can be carried out with a number of software packages. Early software applications include TESTFACT (Bock et al., 2003) and NOHARM (C. Fraser, 1998). The program TESTFACT performs both limited information item factor analysis (LI-IFA) based on tetrachoric correlations and FI-IFA analyses based on the Bock-Aitken model. NOHARM fits an approximation to the multivariate normal distribution, based on first and second-order marginals of the item contingency table (A. Maydeu-Olivares, 2001). Compared to MML, estimation of the FI-IFA approach has recently been improved dramatically by applying the MHRM, for example, IRTPRO (Cai, Thissen, & du Toit, 2011) and flexMIRT (Houts & Cai, 2013). The proprietary program Bayesian Multivariate IRT (Yao, 2003) has also become available for EFA, which presumably takes a Gibbs sampling approach to FI-IFA (Reckase, 2009). Other approaches to estimation are provided by Asparouhov and Muthén (2009) and Muthén and Asparouhov (2013) who use a number of approaches in Mplus to estimation including maximum likelihood (ML), weighted least squares, and Markov chain Monte Carlo (MCMC). Finally, a generalized linear mixed model approach for analyzing latent variables is implemented in the software package GLLAMM (Rabe-Hesketh, Skondal, & Pickles, 2004, 2005). A number of these programs provide options to rotate an initial factor structure to a simple solution (e.g., varimax, oblimin). Recent summaries of the literature on MIRT estimation are provided by D. M. Bolt (2005) and T. A. Ackerman (2005).

Estimation procedures for MIRT have been developed recently for multilevel applications. A brief sketch of several interesting approaches is provided subsequently. F. Rijmen (2010) demonstrated how a two-level MIRT model can be applied to the bifactor model of Gibbons and Hedeker (1992). Cai, Yang, and Hansen (2011) generalized the bifactor model to multiple groups and a range of IRT models. While computationally efficient in terms of the expectation maximization (EM) algorithm, the bifactor model requires a confirmatory item loading structure. Recently, the MHRM algorithm has been applied to multilevel MIRT models (Houts & Cai, 2013; L. Cai, 2010). This algorithm can be used to obtain either exploratory or confirmatory structures. Fox (2003) and Fox and

An Aggregate IRT Model

Downloaded from http://jebs.aera.net at RUTGERS UNIV on July 14, 2015
Glas (2003) also developed algorithms for measurement models in a multilevel framework.

A number of key works underlie the proposed algorithm for factor analysis in this article. First, the seminal work of Rubin and Thayer (1982) provided the foundation for EM factor analysis. This approach was extended to dichotomous variables by Meng and Schilling (1996), who used augmentation to construct a complete-data log likelihood. As they noted “Implementing the M step in this setting is almost trivial” (p. 1257), and this will become evident below. This approach to augmentation was similar to that of J. H. Albert (1992) in his Gibbs sampling approach to unidimensional IRT estimation. More recently, a related approach (through a general framework) was developed by Conti, Fruhwirth-Schnatter, Heckman, and Piatek (2014). The proposed estimation procedure for the aggregate model combines elements of the Robbins–Monro algorithm (Robbins & Monro, 1951) and the stochastic EM algorithm. This synthesis was developed by Delyon, Lavielle, and Moulines, (1999), and their procedure is known as the SAEM algorithm. Other points of similarity include the algorithms proposed by Béguin and Glas (2001) and D. O. Segall (2002). Finally, many of the ideas in this article are natural extensions of the aggregate models developed by Mislevy and Bock (1989).

Aggregate IRT Model

An aggregate MIRT model is proposed in this article for item response data in which there are large numbers of items and clusters (Level 2 units), with sparse data for individuals (Level 1 units). Let $i = 1, 2, \ldots, N$ be the cluster index, $j = 1, 2, \ldots, n_i$ the person index, and $k = 1, 2, \ldots, K$ the item index. Assume an individual-level latent model represented as

$$Z_{ijk} = \eta_{ijk} - b_k + \varepsilon_{ijk},$$

where $Z_{ijk}$ is an item propensity, $\eta_{ijk}$ represents a person parameter as defined subsequently, $b_k$ is the difficulty parameter, and $\varepsilon_{ijk} \sim N(0, 1)$ is a random measurement error. A multivariate elaboration (using the notation of Houts & Cai, 2013) of the person parameter can be expressed as

$$\eta_{ijk} = a_{k1}^{B}0_{i1}^{B} + a_{k2}^{B}0_{i2}^{B} + \ldots + a_{kQ}^{B}0_{iQ}^{B} + a_{k1}^{W}0_{i1}^{W} + a_{k2}^{W}0_{i2}^{W} + \ldots + a_{kR}^{W}0_{iR}^{W} = \mathbf{A}^{B}_{k}0_{i}^{B} + \mathbf{A}^{W}_{k}0_{i}^{W}.$$  

Here, $\eta_{ijk}$ is a random variable composed of $Q$ and $R$ factors at Level 2 (B) and Level 1 (W). In this expression, between-unit cluster scores are subsumed into the matrix $\mathbf{A}^{B}$ and the within-unit factor scores into $\mathbf{A}^{W}$. In parallel, between-unit factor loadings (or item slopes) are subsumed into the matrix $\mathbf{A}^{B}$ and within-unit loadings into $\mathbf{A}^{W}$. In a unidimensional IRT model, $\mathbf{A}^{B} = 0$ and $\mathbf{A}^{W}$ consists of a single column of slope or item discrimination parameters.
The goal of aggregate IRT estimation is to model $p_{ik}$, the observed proportion of correct responses for item $k$ in cluster $i$. Because individual responses to an item within a cluster are aggregated, this can substantially reduce both the size of the data set and processing time. For a two-parameter normal ogive model, this approach ideally requires the probability

$$
\bar{P}_{ik} = \frac{1}{n_i} \sum_j P_{ijk},
$$

where

$$
P_{ijk} = \Phi(\eta_{ijk} - b_k).
$$

While the function in Equation 3 is based on the actual individuals of a cluster (though this approach is also consistent with a superpopulation model), an objective of the aggregate approach is to bypass individual computations. Accordingly, an approximation to Equation 3 can be obtained as the population proportion correct within a cluster via the integral

$$
\bar{P}_{ik} \approx \int W \int_E P_{ijk} \left( \theta^W_i, \theta^W_j \right) f \left( \theta^W_j, \varepsilon_{ijk} \right) d\theta^W_j d\varepsilon_{ijk},
$$

where $f$ the joint probability density of $\left( \theta^W_j, \varepsilon_{ijk} \right)$ is assumed to be $MVN(0, I)$. The integral can then be obtained analytically for uncorrelated Level-1 factors along the lines of R. J. Mislevy (1983) as

$$
\Phi \left( \frac{\Lambda^B_i \theta^B_j - b_k}{\sqrt{1 + \Lambda^W_k (\Lambda^W_k)^T}} \right),
$$

where $\Phi$ is the cumulative normal distribution function. Note that Equation 6 represents a normal approximation to $P_{ik}$ in Equation 3. This result bears a resemblance to the school-level model proposed by Mislevy and Bock (1989, pp. 63–65).

If factor scores are correlated, Equation 6 can be expressed in terms of an orthonormal transformation which does not change the value of the denominator. However, one simplifying assumption is unidimensionality at Level 1. This results in:

$$
\bar{P}_{ik} \approx \Phi_{ik} = \Phi \left( \frac{\Lambda^B_i \theta^B_j - b_k}{\sqrt{1 + a_k^2 \sigma^2_i}} \right)
$$

The assumption $\sigma^2_i = 1$ signifies that the within-cluster variances are identically equal to 1.0. The latent linear model at the aggregate level corresponding to Equation 7 can then be conceptualized as

$$
Z_{ik} \approx \eta_{ik} - b_k + \bar{e}_{ik},
$$
where

\[ \eta_{ik} = A_k^B g_f^B \]
\[ \tilde{\varepsilon}_{ik} \sim N(0, n_{ik}^{-1} \lambda_k), \] (9)

where \( \lambda_k = 1 + a_k^2 \), and \( a_k \) is the item’s factor loading (or discrimination) at Level 1. While this unidimensionality assumption is strong, it may nonetheless be useful because many major achievement batteries are constructed to have a strong first factor at the individual level. In the empirical analysis and simulation studies discussed subsequently, the effects of violating this assumption are also shown to be negligible. Another interesting consequence of this decision is that if univariate scaling is used to calibrate and score items at Level 1, then the multivariate structure at Level 2 is consistent with the Level 1 scaling.

Estimation of Model Coefficients

This procedure is proposed as an alternative to both the EM and MHRM algorithms in data sets with extremely large numbers of clusters. For multilevel MIRT models, EM is rarely practical with more than a handful of quadrature points for more than five dimensions (within plus between). The MHRM algorithm is more efficient than EM, and preliminary experience shows the SAEM approach is comparable to that of MHRM in terms of computing time. However, the SAEM algorithm provides at least three interesting possibilities relative to MHRM. First, some functions of components of the SAEM algorithm presented subsequently have an “embarrassingly parallel” structure and suggest computational speeds can be improved substantially. Second, basic code of the major program functions is extremely compact. For example, using simple prepackaged functions, for example, five lines of code are used for estimating \( A^B \), and an alternative approach requires only one line. Software packages using the Gibbs sampling approach to EFA would require only slight modification to perform the SAEM algorithm. Third, the algorithm relies on missing value imputation, and this provides a means for incorporating auxiliary information for imputing structurally missing item responses due to matrix sampling. Fourth, the algorithm can also incorporate observed quantitative variables directly into the factor analysis; thus, binary and quantitative variables can be combined. Fifth, the estimation approach also works in cases with one unit at each Level 1/Level 2 combination. In this case, the SAEM approach can very easily be applied to Level-1 IRT calibrations. In short, the SAEM approach should not be viewed as a better or faster alternative to current MIRT estimation algorithms but rather as an approximation that may allow for creative solutions in current IRT and MIRT applications. Finally, the SAEM algorithm is a popular estimation approach in other fields (e.g., see Savic, Mentre, & Lavielle, 2011).
The applications discussed subsequently require items to be dichotomized, though in principle, the aggregate model can also be applied to polytomous data. No doubt, some information is lost in converting to binary responses, but the virtue of the aggregate approximation is speed and simplicity. For each item, it is assumed that a set of unidimensional IRT parameters are available (as they typically are in a large-scale assessment). Subsequently, $A^W_k$ has been abbreviated to $a_k$, which reflects the assumption of unidimensionality at Level 1. Guessing is incorporated into the model as

$$\hat{P}_{ik} \approx c_k + (1 - c_k)\Phi_{ik},$$

where $c_k$ is the lower asymptote of the normal ogive.

**General Description of the SAEM Algorithm**

The SAEM algorithm (Delyon, Lavielle, & Moulines, 1999), which bears a similarity to the SAEM method of Celeux and Diebolt (1992), is a combination of stochastic EM and the Robbins–Monro method. In this approach, the E step is replaced by a stochastic approximation, which can be based on a single draw of the missing or augmented data. Let $y$ represent the data, $\psi$ the augmented data, and $\log f_0(y; \xi)$ the complete data log likelihood. At iteration $t$ of SAEM:

1. **Augmented data.** Obtain one draw of the missing data $\psi$ from the conditional distribution $p(\psi|y, \xi^{(t)})$.

2. **SA-step (stochastic approximation).** Update $Q_t = Q_{t-1} + \gamma_t \{\log f_0(\psi; \xi) - Q_{t-1}\}$, where $\gamma_t$ is a positive sequence such that $\gamma_t = 1$, $\sum_{t=1}^{\infty} \gamma_t \rightarrow \infty$, and $\sum_{t=1}^{\infty} \gamma_t^2$ is finite.

3. **M-step.** Update $\xi^{(t+1)} = \arg \max_{\xi} Q_{t+1}$.

Implementation of SAEM (as well as EM) is highly simplified when the complete likelihood $f_0(y; \xi)$ belongs to a curved exponential family. In this case (in Step 2 of the SAEM algorithm), the vector of sufficient statistics $s(y, \psi)$ is computed at iteration $t$ and then updated according to

$$s^{(t+1)} = s^{(t)} + \gamma_k \{s^{(t)} - s^{(t)}\}.$$  

In the M step, $\xi$ is updated as a function of $s$ at iteration $t + 1$. Any maximizing function of $\xi|s$ can be used.

The SAEM algorithm has been shown to converge to the maximum likelihood estimation under general conditions (Delyon et al., 1999). Generally, SAEM provides better results than EM because SAEM usually does not terminate in the first local maximum encountered. Celeux, Chauveau, and Diebolt (1996, p. 2) explained that random perturbations prevent the parameter search from getting bogged down in unstable or hyperbolic fixed points, or even stable but insignificant local maxima, to which EM is susceptible.
Aggregate Model Estimation

For estimation, the SAEM algorithm is applied to aggregate data. In this case, the observed data consist of the proportions correct for items within a cluster, the augmented data correspond to the set $\psi = (Z, V, \Theta)$, where $V$ is defined subsequently. The fixed parameters of interests are contained in $A$. These augmented data are identical in form to those described by Béguin and Glas (2001). Combined with augmented data for obtaining stochastic expectations, the SAEM algorithm is based on the expected value of the log likelihood assuming the prior distribution $\Theta \sim N(0, I)$. Let $D_i = \text{diag}(n^{-1}_k \lambda^2_k)$, $k = 1, 2, ..., K$. The complete-data log likelihood is then given by

$$
\log L = -\frac{1}{2} \left[ \sum_i \{z_i - (A\Theta_i - b)\}^T D_i^{-1} (z_i - (A\Theta_i - b)) + \Theta_i^T \Theta_i \right].
$$

(12)

The steps of the algorithm are as follows:

1. Draw dichotomous item knowledge $V$ conditional on $A, \Theta$ (first augmented data step).
2. Draw continuous item knowledge $Z$ conditional on $A, \Theta, V$ (second augmented data step).
3. Estimate the covariance $\Sigma_z = \text{Cov}(Z)$ with the sufficient statistic $s = \text{cov}(z)$.
4. Estimate $A$ by with standard ML analysis using a sufficient statistic $s$.
5. Draw the factor scores $\Theta$ conditional on $A, Z$ (third augmented data step).

The Steps 1 to 5 are continued until $\max|A_k - A_{k-1}| < .0001$ for three successive iterations. In models with many parameters, a small convergence criterion (say 10E-8) may dramatically increase computation time.

Each step is explained in more detail subsequently:

**Step 1. Dichotomous item knowledge.** Based on Equation 7, the conditional probability of a correct response without guessing (e.g., J. H. Albert, 1992; D. O. Segall, 2002) is given by

$$
R_{ik} = \frac{\Phi_{ik}}{c_k + (1 - c_k)\Phi_{ik}},
$$

(13)

where $c_k$ is the guessing parameter for item $k$. That is, each of $n_{ik}$ pseudo draws from the marginalized distribution (defined in the previous section) has a probability that the associated response was not due to guessing. The number of correct responses $r_{ik}$ can be adjusted for guessing by sampling $v_{ik}$ from the distribution

$$
v_{ik} \sim \text{Bin}(r_{ik}, R_{ik}).
$$

(14)

The quantity $v_{ik}$ denotes the number of students knowing the answer when taking guessing into account. When guessing is present, $v_{ik}$ is substituted for $r_{ik}$, and the number of incorrect responses is obtained as $n_{ik} - v_{ik}$. Recall that in the present
case, the $c_k$ parameters (as well as $a_k, b_k$) are assumed known. Define $V$ as the $N \times K$ matrix of dichotomous knowledge values $v_{ik}$.

**Step 2. Continuous item knowledge.** The goal here is to generate continuous latent propensities for each item at the cluster level. Given the latent knowledge indicator $v_{ik}$, draw random values from $N(\eta_{ik} - b_k, \lambda_k)$ truncated from below with frequencies

$$Z_{ijk} : \begin{cases} v_{ik} & Z_{ijk}^{(1)} > 0 \\ n_{ik} - v_{ik} & Z_{ijk}^{(0)} \leq 0. \end{cases}$$

Aggregate augmented values are then computed as

$$\bar{Z}_{ik} = n_{ik}^{-1} \left[ \sum_{v_{ik}} Z_{ijk}^{(1)} + \sum_{n_{ik} - v_{ik}} Z_{ijk}^{(0)} \right].$$

For clusters with missing item responses for item $k$, a single value of $Z_{ijk}$ is imputed from the full distribution $N(\eta_{ik} - b_k, \lambda_k)$. For convenience, lower case $z$ is used to designate actual values of the vector of mean item propensities.

**Step 3. Sufficient statistics.** Estimate $\Sigma_z$ in the framework of stochastic approximation. On step $k$, compute

$$s_t = s_{t-1} + \gamma_t \text{cov}(z_t) - s_{t-1},$$

where $\text{cov}(z_t)$ is the weighted empirical covariance matrix (where cluster weights are obtained by summing individual unconditional weights) for iteration $t$, and $\{\gamma_t; t \geq 1\}$ is a sequence of gain constants (Robbins & Monro, 1951). Typically, $\gamma_t = t^{-1}$ or $\gamma_t = t^{-2/3}$. In the SAEM algorithm of Delyon et al. (1999), sufficient statistics for the complete data likelihood are updated, and estimates of fixed coefficients are obtained directly from these updates. The proposed procedure bears a resemblance to the method of Gu and Kong (1998) and L. Cai (2010), but in those cases stochastic approximation is applied to parameters rather than sufficient statistics. Delyon et al. (1999) provided a proof of the convergence of the SAEM algorithm. As shown subsequently, estimates from the algorithm also agree closely with existing MIRT software and perform adequately in simulations.

**Step 4. Factor loadings.** A number of approaches for obtaining the solution for $A$, the $K \times Q$ factor loading matrix, are given subsequently. First, the unconditional ML estimate in the M step is obtained by taking the derivative of log $L$ with respect to $A$ and setting the result to zero:

$$A_k^T = \left( \theta^T D_k^{-1} \theta \right)^{-1} \theta^T D_k^{-1} (z_k + 1b_k),$$

where $1$ is a $N \times 1$ vector of 1s, $D_k = \text{diag}(n_{ik}^{-1} \lambda_k^2)$ for $i = 1, 2, ..., N$, and $z_k$ is a $N \times 1$ vector of propensities for item $k$. For estimation in the Robbins–Monro framework, gain constants could be applied to $\theta^T D_k^{-1} \theta$ and $\theta^T D_k^{-1} z_k$ on each
iteration. In this case, the latter two functions of the complete data are the sufficient statistics. Rubin and Thayer (1982) and Meng and Schilling (1996) examined issues of sufficiency in the context of EM factor analysis, and this may provide an alternative means of estimation.

A different approach is obtained by direct factoring of augmented values. Let $w_i, i = 1, 2, ..., N$ define sampling weights for clusters. The weighted covariance matrix is then obtained as

$$
\text{cov}(z) = \frac{\sum w_i}{\left(\sum w_i\right)^2 - \sum w_i^2} \sum w_i (z_i + b)(z_i + b)^T.
$$

As indicated earlier, once item propensities are generated for clusters, the weighted covariance matrix is obtained solely by sampling weights. This approach is approximate due to the difficulties of using sampling weights in conjunction with item response weights that are heterogeneous within and across clusters. However, it should be noted that within a cluster, item weights are generally (though not always) similar for assessment program implementing balanced incomplete block booklet spiraling. Use of item error variances (inverse of the weights) is a more important issue in estimating and sampling.

The estimated covariance is then treated as a sufficient statistic and updated on each iteration according to the Robbins–Monro procedure. After updating, factoring of $s_t$ is carried out according to the model

$$
\Sigma_z = A A^T + U.
$$

Using standard ML analysis, both $A$ and $U$ are estimated in each iteration $t$. However, in the presented context, uniqueness score components are generated in the Gibbs sampling process, and this knowledge can be used to bypass estimation of $U$. Let $\tilde{n}_k = N^{-1} \sum_i n_{ik}$ be the average number of responses to item $k$ across clusters. According to Equation 9, the unique variance at the $k$th diagonal element of $U$ can be determined as

$$
u_k = \tilde{n}_k^{-1} \lambda_k = \tilde{n}_k^{-1} (1 + \alpha_k^2).
$$

Estimation of $A$ on iteration $t$ is then conducted by a standard eigenvalue analysis of $s_t - U_k$. Note the covariance matrix $s_t$ is analyzed, not the correlation matrix.

**Step 5. Cluster Scores.** Recall the $K \times 1$ vector of item difficulties is $b$. The latent linear model for cluster $i$ is

$$
z_i^T = A \theta_i - b + \varepsilon_i,
$$

where $z_i^T$ is the $K \times 1$ vector of item propensities for cluster $i$. Let $D_i = \text{diag}(\tilde{n}_{ik}^{-1} \lambda_k^2)$ for $k = 1, 2, ..., K$. The ML solution for cluster score is given by
\[ \hat{\theta}_i \sim (1 + A^T D_i^{-1} A)^{-1} A^T D_i^{-1} (z_i^T + b), \]  

(22)

upon applying the prior \( \Theta \sim MVN(0, I) \). A Gibbs step on each iteration is used to draw augmented data as

\[ \Theta_i \sim MVN\left(\hat{\Theta}_i, (1 + A^T D_i^{-1} A)^{-1}\right), \]

(23)

which also specifies the latent scale. Because the items are weighted in this case according to the inverse of their residual variance, the posterior estimate is a type of Bartlett factor score. Such a score has the property of being strongly correlated to its corresponding factor but not other factors. The last step of the algorithm can also be viewed as the first step of the next iteration due to the cyclical nature of the algorithm (so that augmented variables are generated sequentially).

**Empirical Example**

Data were taken from the TIMSS 2007 mathematics for grade 4 (Williams et al., 2009; for updated information, see Martin & Mullis, 2012). In TIMSS 2007, there are 59 jurisdictions with about 150 schools per jurisdiction and over 200 items matrix sampled by student within school (who each typically take about 25 items). Here, jurisdictions can be countries, states, or provinces within a country. The number of jurisdictions differs by both grade and topic (science or mathematics). For Grade 4, there are 41 jurisdictions, 177 items, and about 194,000 students. Initial analyses suggested that school sizes in many cases are very small (e.g., in the Australian outback), and a strategy was developed for aggregating across two similar schools belonging to a jackknife zone (JKZ), which are clusters formed by sorting schools on both explicit and implicit background variables and then pairing adjacent schools within jurisdictions (P. Foy, 2012). This resulted in 3,133 clusters (pairs of schools), and “house weights” (Rutkowski, von Davier, & Rutkowski, 2013) from the TIMSS survey were also aggregated to the cluster level. All items were dichotomized and a sample weighted IRT analysis (one dimensional) was run to obtain estimates of Level-1 IRT parameters. These were subsequently incorporated into Equation 10.

**Estimation results for TIMSS data**

First, the individual-level item parameters were obtained; second, these parameters were incorporated into the aggregate model algorithm. The results consist of estimates obtained after the absolute maximum parameter change in A was .0001 across three consecutive iterations. Estimation of a \( Q = 5 \) factor solution was obtained with a custom algorithm written in R in which generation of augmented values and several other functions were parallelized. Altogether, 150 SEM iterations were performed, followed by 200 SAEM iterations.
At convergence, raw eigenvalues for 177 items of 147.14, 10.89, 7.82, 2.88, and 2.63 for the covariance matrix were obtained, indicating three strong and two weak factors. Note these eigenvalues are not scaled: They represent the sum of squared loadings on factors, and loadings (or discriminations) in the covariance metric can exceed 1.0. Given a fixed number of items, their relative sizes are informative rather than their absolute sizes. To aid interpretation, eigenvalues are also given in the correlation metric (scaled eigenvalues), which represent the proportion of between-cluster variance accounted for. Initial factor loadings (upon extraction) were then rotated to a simple solution with the varimax criterion (not normalized). The varimax transformation matrix was then applied to factor score estimates of \( \mathbf{H} \). Transformed factor score correlations were expectedly low, given the orthogonal rotation; however, an oblique rotation obtained highly similar results. The highest off-diagonal correlation at the cluster level was \( r = -0.055 \).

In Table 1, the items comprising the two highest loadings are given with brief descriptions, though the discussion presented subsequently is based on the item highest loadings. The factors are highly interpretable. Factor 1 involves graphical

<table>
<thead>
<tr>
<th>Factor</th>
<th>Content Domain</th>
<th>Topic</th>
<th>Cognitive Domain</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Geom Shapes and Meas Data Display</td>
<td>Location and Movements</td>
<td>Applying</td>
<td>Mark Lucy’s house</td>
</tr>
<tr>
<td>2</td>
<td>Number</td>
<td>Whole Numbers</td>
<td>Knowing</td>
<td>762 cars parked in 6 equal rows</td>
</tr>
<tr>
<td>3</td>
<td>Number</td>
<td>Fractions and Decimals</td>
<td>Knowing</td>
<td>4/5 minus 1/5</td>
</tr>
<tr>
<td>4</td>
<td>Geom Shapes and Meas</td>
<td>2-D and 3-D shapes</td>
<td>Applying</td>
<td>Make a 6-sided figure</td>
</tr>
<tr>
<td>5</td>
<td>Number</td>
<td>Fractions and Decimals</td>
<td>Knowing</td>
<td>Make a 4-sided figure</td>
</tr>
<tr>
<td></td>
<td>Number</td>
<td>Number Sentence</td>
<td>Applying</td>
<td>Equivalent fractions</td>
</tr>
</tbody>
</table>

Note. TIMSS = Trends in International Mathematics and Science Study.
or visually assisted reasoning related to data display and geometric shapes and measures. While there are whole number problems loading on this factor, these tend to have a strong visual orientation (e.g., M031173: “Maria has 6 red boxes. Each red box has 4 pencils inside. She also has 3 blue boxes. Each blue box has 4 pencils inside. How many pencils does Maria have altogether?”). Factor 2 involves whole numbers and number sentences. These items all involve solving for an unknown rather than counting and conspicuously lack a robust graphical representation. Factor 3 clearly consists of items concerning fractions and decimals. Factor 4 is a relatively weak factor that involves primarily two open-ended items involving spatialization with multiple questions. This appears to be a case of local dependence within the test, but these two sets of items have interesting properties as explained subsequently. Factor 5 is also a relatively weak factor that involves symmetry and equivalence.

The TIMSS assessment is comprised of three content domains and nine topics. The 20 highest loading items for each factor can be broken out as follows. Factor 1 contains items from all three content domains and six topics. Factor 2 is comprised primarily of items from the number domain and two topics: whole numbers and number sentences. Factor 3 is marked by items from the content domains fractions and decimals, and geometric shapes and measures. A majority of these items come from the topic factions and decimals, but other six topics are also represented. Factor 4 contains items from all three content domains, and Factor 5 contains items from two content domains (number, and geometric shapes and measures). Factors 4 and 5 both contain items from seven different topics.

The 2007 TIMMS assessment also contains three cognitive domains. The Knowing domain comprises knowing basic facts, procedures, and concepts that bridge the elements of basic knowledge. Behaviors include recall, recognition, computation, extraction of information (e.g., from graphs), measurement, knowing (e.g., place value), and classification/ordering. The Applying domain comprises the application of knowledge and concepts to create representation and solve problems. Behaviors include selecting, representing, modeling, implementing, and solving routine problems. The Reasoning domain involves the capacity for inductive and deductive reasoning to obtain solutions to nonroutine problems. Behaviors include analyzing, generalizing, synthesizing, justifying, and solving

<table>
<thead>
<tr>
<th>Factor</th>
<th>Knowing</th>
<th>Applying</th>
<th>Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>9</td>
<td>7</td>
</tr>
</tbody>
</table>
nonroutine problems. In terms of the 20 highest loading items on each factor, the factors varied with respect to the proportion of cells of the cognitive domain. These frequencies are shown in Table 2. As can be seen, Factors 3 (Fractions and Decimals) and 5 (Symmetry) are primarily composed of Knowing items, whereas other factors are mainly composed of Applying and Reasoning items. Factor 4 is a weak factor that was indicated by 2 open-ended items, yet 7 items are tagged as Reasoning and 9 as Applying.

Comparison to flexMIRT

One approach to model evaluation is to compare the aggregate procedure to a multilevel MIRT program that makes fewer assumptions. The flexMIRT software (Houts & Cai, 2013) is ideal for this purpose because it offers both the EM and MHRM algorithms. For both flexMIRT and the aggregate approaches, Level-1 item parameters are first obtained and fixed in order to estimate Level-2 loadings. It should be recognized that Level-1 parameters are independent of Level-2 parameters because “the distribution of the item responses conditional on the underlying performance level of the student, does not depend on group membership at all, because IRT models require the assumption that underlying performance level only depends on the item responses for a student and the item parameters” (National Center for Education Statistics, 2008). Below, the comparison is made based on the TIMSS 2007 mathematics for Grade 4 data, which included 187,673 students and 177 items.

For flexMIRT, the setup requires five between factors and one within factor, with Level-1 parameters (read in and) fixed. With a standard MHRM configuration, estimates converged at 1E-4 in 1 hour and 49 minutes (210 Stage 3 iterations were required). The flexMIRT loadings were then rotated to the target of loadings from the aggregate model using the TargetT function (Bernaards & Jennrich, 2005) of the GPArotation package in R (R Core Development Team, 2012). Note that without rotation, the comparison of factor loadings is obscured.

**TABLE 3.**
Comparison of Factor Loadings From Aggregate Model and flexMIRT

<table>
<thead>
<tr>
<th>Factor</th>
<th>Correlation</th>
<th>Difference SD</th>
<th>Min^a Loading</th>
<th>Max^a Loading</th>
<th>Multiple R</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.90</td>
<td>.15</td>
<td>-.28</td>
<td>1.94</td>
<td>.90</td>
</tr>
<tr>
<td>2</td>
<td>.98</td>
<td>.08</td>
<td>-.48</td>
<td>1.13</td>
<td>.96</td>
</tr>
<tr>
<td>3</td>
<td>.98</td>
<td>.07</td>
<td>-.30</td>
<td>1.26</td>
<td>.89</td>
</tr>
<tr>
<td>4</td>
<td>.88</td>
<td>.10</td>
<td>-.53</td>
<td>.40</td>
<td>.97</td>
</tr>
<tr>
<td>5</td>
<td>.85</td>
<td>.10</td>
<td>-.13</td>
<td>.61</td>
<td>.97</td>
</tr>
</tbody>
</table>

Note. MIRT = multidimensional item response theory.
^aLoadings from aggregate model.
by rotational indeterminacy. For the five factors, summary statistics are given in Table 3. As can be seen, the factor loadings from the two procedures are highly correlated (Column 1) and have a relatively low variation (SD) in absolute difference (Column 2), though some differences are evident for Factors 1, 4, and 5. However, another perspective is provided by multiple correlations predicting each set of aggregate loadings from the five loadings produced by flexMIRT, as shown in Column 5 of Table 3. It appears that some of the differences in loadings for Factors 4 and 5 may be due to rotational issues.

Simulation

Two brief simulations were carried out to see whether the aggregate model could recover cluster-level information accurately. Data generation was partially designed according to the matrix sampling plan of TIMSS (Jonas & Foy, 2012) but did not take into account the complex sampling structure of JKZ clusters within country. Recall that the empirical models are estimated from item response data aggregated to the JKZ; that is, the JKZ strata are the clusters upon which $A$ and factor scores are estimated. Actual JKZs are created by sorting on explicit and implicit criteria that differ by jurisdiction, and many of these indicators are not easily accessed. For this reason, it is not feasible to conduct simulations at this level of detail. Consequently, a simplified JKZ structure was implemented in the simulation studies, as described subsequently that focused on a reduced number of key elements of the TIMSS sampling design.

First simulation

A simulation was run with $Q = 5$ dimensions at Level 2, $m = 177$ items, and $n = 3,133$ clusters with 25 students each. Individual ($a, b, c$) and group-level ($A$) IRT parameter estimates from the empirical analysis mentioned earlier were used to generate item response data. A multinomial distribution was constructed to represent the range of JKZ cluster sizes observed in the actual data. Sizes varied from 2 to 80, though extreme values were rare. A staggered matrix sampling design was created with students responding to a maximum of 26 items. Sequential students had a 50% item overlap in the matrix sampling design, and about 7% of item responses to randomly flipped to “not presented” (a feature noted in the actual data). The estimated factor loading matrix was rotated orthogonally (not normalized) to the target $A$ to resolve factor indeterminacy using the TargetT function. Note the assumption $\sigma_i^2 = 1$ was used to generate data, but actual values of $s_i$ varied due to sampling.

Simulation results. The difference $D = A - \hat{A}$ was computed for the rotated solution, and average difference (bias) and standard deviation were computed for each column. These figures are given in Table 4. The largest bias (about .04) was for the first factor, which generally had the highest loadings. The difference standard deviation (about .13) was a highest for Factor 5 but also relatively high for
Factor 4. The correlations between estimated and generating factor loadings for the first three factors were high at \( r > .95 \); however, the correlations for Factors 4 and 5 were lower at .90 and .47. The eigenvalues at convergence shown in Table 4 generally matched in relative size those of the empirical analysis. In terms of absolute size, the simulated eigenvalues for Factors 1, 4, and 5 deviated somewhat from their empirical counterparts.

Cluster scores were estimated and rotated to examine correlations with the generating cluster scores. For this purpose, 25 imputations were generated post-convergence, and \( \hat{\theta} \) was obtained as the average of these values prior to rotation. The correlation matrix is given in Table 5. As can be seen, the correlations are high for Clusters 1 through 3 but moderate to low for Clusters 4 and 5. In summary, both the generating parameters and \( \theta \) values were adequately estimated for the three strongest factors, but estimates of the cluster scores for the weaker factors still held some relevant information.

**Second simulation**

A second simulation was carried out assuming a multivariate structure for \( \Theta^W \). Three Level-1 dimensions were specified: The first dimension was strong and the

---

**TABLE 4.**

Statistics for Evaluation of the First Simulation

<table>
<thead>
<tr>
<th>Factor</th>
<th>Raw Eigenvalues</th>
<th>Scaled Eigenvalues</th>
<th>( D = \hat{A} - A )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Empirical</td>
<td>Simulated</td>
<td>Empirical</td>
</tr>
<tr>
<td>1</td>
<td>147.14</td>
<td>129.97</td>
<td>110.83</td>
</tr>
<tr>
<td>2</td>
<td>10.89</td>
<td>9.32</td>
<td>7.52</td>
</tr>
<tr>
<td>3</td>
<td>7.82</td>
<td>7.32</td>
<td>4.93</td>
</tr>
<tr>
<td>4</td>
<td>2.88</td>
<td>5.86</td>
<td>2.33</td>
</tr>
<tr>
<td>5</td>
<td>2.63</td>
<td>4.42</td>
<td>2.07</td>
</tr>
</tbody>
</table>

**TABLE 5.**

Correlations of Estimated and Actual Cluster Scores for the First Simulation

<table>
<thead>
<tr>
<th>Actual ( \Theta )</th>
<th>( \theta_1 )</th>
<th>( \theta_2 )</th>
<th>( \theta_3 )</th>
<th>( \theta_4 )</th>
<th>( \theta_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta_1 )</td>
<td>.95</td>
<td>.01</td>
<td>-.03</td>
<td>-.04</td>
<td>-.03</td>
</tr>
<tr>
<td>( \theta_2 )</td>
<td>.00</td>
<td>.90</td>
<td>-.04</td>
<td>.04</td>
<td>.03</td>
</tr>
<tr>
<td>( \theta_3 )</td>
<td>-.03</td>
<td>-.04</td>
<td>.89</td>
<td>.01</td>
<td>.01</td>
</tr>
<tr>
<td>( \theta_4 )</td>
<td>-.04</td>
<td>-.02</td>
<td>.01</td>
<td>.54</td>
<td>.03</td>
</tr>
<tr>
<td>( \theta_5 )</td>
<td>-.02</td>
<td>-.00</td>
<td>-.06</td>
<td>-.02</td>
<td>.43</td>
</tr>
</tbody>
</table>
next two were weaker. The first dimension was defined by a set of loadings equal to the $a$ parameter from the univariate IRT calibration. For the second and third Level-1 dimensions, a set of loadings were randomly generated with log normal distribution $a \sim \text{lnorm}(-1, .25)$ and offset downward by .25 units. This produced many loadings close to zero and a few larger loadings. Other aspects of the simulation were identical to those of the first simulation. In particular, the univariate Level-1 assumption was maintained for analysis, though not for data generation.

Simulation results. The difference $D = \hat{A} - A$ was computed for the rotated solution, and average difference (bias) and standard deviation were computed for each column. These figures are given in Table 6. The biases, standard deviations, factor correlations, and eigenvalues generally matched those of the first simulation. Again, those results deviated moderately from the empirical solution that was used to design the simulations. Cluster scores were estimated and rotated to examine correlations with the generating cluster scores. The correlation matrix is given in Table 7. As can be seen, the correlations are high for Factors 1 through 3, but moderate to low for Factors 4 and 5. These results are also similar to those of the first simulation.

### Table 6.
Statistics for Evaluation of the Second Simulation

<table>
<thead>
<tr>
<th>Factor</th>
<th>Raw Eigenvalues</th>
<th>Scaled Eigenvalues</th>
<th>$D = \hat{A} - A$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Empirical</td>
<td>Simulated</td>
<td>Empirical</td>
</tr>
<tr>
<td>1</td>
<td>147.14</td>
<td>123.86</td>
<td>110.83</td>
</tr>
<tr>
<td>2</td>
<td>10.89</td>
<td>8.69</td>
<td>7.52</td>
</tr>
<tr>
<td>3</td>
<td>7.82</td>
<td>7.12</td>
<td>4.93</td>
</tr>
<tr>
<td>4</td>
<td>2.88</td>
<td>5.46</td>
<td>2.33</td>
</tr>
<tr>
<td>5</td>
<td>2.63</td>
<td>4.22</td>
<td>2.07</td>
</tr>
</tbody>
</table>

### Table 7.
Correlations of Estimated and Actual Cluster Scores for the Second Simulation

<table>
<thead>
<tr>
<th>Actual $\theta$</th>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>$\theta_3$</th>
<th>$\theta_4$</th>
<th>$\theta_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td>.96</td>
<td>.02</td>
<td>-.05</td>
<td>-.06</td>
<td>-.00</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>-.00</td>
<td>.89</td>
<td>-.04</td>
<td>.06</td>
<td>.05</td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>-.04</td>
<td>-.06</td>
<td>.89</td>
<td>.01</td>
<td>-.02</td>
</tr>
<tr>
<td>$\theta_4$</td>
<td>-.04</td>
<td>.01</td>
<td>-.04</td>
<td>.54</td>
<td>.01</td>
</tr>
<tr>
<td>$\theta_5$</td>
<td>.01</td>
<td>-.04</td>
<td>-.04</td>
<td>.00</td>
<td>.41</td>
</tr>
</tbody>
</table>
Discussion

A method for estimating cluster-level factor scores was proposed, and a potential application was illustrated. Model evaluation issues were addressed by applying the model to assessment data from TIMSS 2007 Mathematics in grade 4. In addition, one comparison to an independent software package and two brief simulations were used to examine the adequacy of factor loadings and cluster scores. In the conditions examined, the method appears to be useful despite violations of local independence and the unidimensionality at Level 1.

Although the cluster scores were less reliable for weaker factors in the simulations, for the real data even the weakest factors were highly interpretable. This mismatch could be due to the difficulty in simulating all aspects of the actual TIMSS sampling design. A good discussion of the issues with extracting weaker factors is provided by C. Ximenez (2009) and Briggs and MacCallum (2003), who showed that factor recovery depends on a number of variables. Examining goodness-of-fit statistics may also help to clarify the role of weaker factors. This would be useful, but the development of fit indices in the aggregate model is complicated by the issue of local independence, which is considered more fully in the concluding section. The results were demonstrated to be robust only for two limited simulations, and application of the model in very different contexts would require additional simulation. Finally, there is often not an obvious method of factor rotation with real data sets. This issue is somewhat obscured in the simulation studies where the rotation target was the set of actual loadings. The latter choice was made in order to distinguish the accuracy of estimation from the more subjective choices involved in rotation.

Areas of Further Research

Local independence

In full three-level models, local independence is achieved by conditioning on both \( \theta^W \) and \( \theta^B \). The aggregate model, however, only conditions on \( \theta^B \). The local independence holds for the Level-2 model only when each sampled subject within a school responds to only 1 item. When a student responds to more than 1 item, there is dependence among those responses due to the shared but unmodeled \( \theta^W \). The resulting violation of local independence is not serious, however, because it does not bias the percentage correct for the cluster for either census or matrix sampled items. Consequently, the aggregate data are consistent, for the same cluster-level expectations as if there had been only one response per sampled student (Mislevy & Bock, 1989). In turn, this suggests that Level-2 models will not produce biased estimates of item parameters due to this particular source of local dependence.
Standard errors

There are three central issues with regard to obtaining accurate standard errors of item and person parameters in the aggregate models. First, standard errors of item parameters are expensive to compute in large multilevel MIRT models, and complex survey data add an order of magnitude to this complexity. One approach for taking the sampling design into account would be to use stratified bootstrap sampling from individual data, where the strata are defined by jurisdiction and the JCK zones (the clusters within those strata). This would bring standard error estimates partially in line with the sampling design for the TIMSS survey but would defeat the efficiency of the aggregate model as well as dramatically increase the computational expense. The approach of Delyon et al. (1999) for computing standard errors would seem to be ideal, but this has not been adapted to survey data. Second, standard errors produced by the aggregate model are likely to be too small because the aggregate statistics are treated as though they were independent estimates of the item proportions correct. The more items a student takes, the greater the unmodeled conditional dependence. In turn, the negative bias in the standard errors increases. Further research for demonstrating the level of bias in standard errors would be useful.

A third issue arises in the computation of standard errors for Level 2 factor scores. Fast approximate standard errors for $A$ can be computed in two steps. On a single iteration post-convergence, bootstrap from $z$ and compute the bootstrap variance of $A$. Repeat this procedure for a small number of iterations and average variance errors (for each parameter separately) across iterations. Post-convergence, standard errors of a cluster score $\theta$ can be obtained as follows. First, remove the gain constant from the iterative process. Then run $m$ estimation complete estimation cycles and save estimates of $\theta$ on each cycle; these values are referred to as imputations. The final estimate of $\theta$ is the mean over imputations. This is essentially an EAP estimate based on Monte Carlo integration. The variance of $\theta$ is likewise obtained as the variance over imputations using MCMC techniques. This method can also be used in a similar fashion for computing standard errors for $A$, so that uncertainty about $A$ would be incorporated into the standard error of $\theta$. After rotation, however, standard errors would need to be transformed (see G. Zhang, 2014, for recent work on this issue).

Multilevel Measurement Invariance

For the case of MIRT, there are two separate questions of measurement invariance. First, a key assumption is that the Level-1 measurement model is sample independent with respect to clusters and any other identifiable subset of the data. If this is not the case, then test scores cannot be interpreted as revealing inter-individual differences. In fact, this is a well-known issue problem commonly addressed with differential item functioning (DIF) techniques. In large-scale assessment programs, this issue uniformly receives a great deal of attention.
Second, there is a corresponding concern at Level 2 regarding whether there are groups of clusters (say countries) for which $A^B$ is not invariant. Both assumptions of invariance are made in the empirical example to illustrate the SAEM estimation approach. However, DIF studies of $A^B$ are likely to be useful, for example, Asian versus European countries. If differences were found, this would provide limiting information for cross-country interpretations of test scores but also interesting insights to the effects of culture and policy on achievement in different educational systems.

It should also be recognized that the cluster-level model does represent a type of DIF study. That is, a nonzero $A^B$ indicates variation in item difficulty effects across groups in the sense of Rijmen and Jeon (2013). This is because Equation 1 can be reinterpreted as

$$A^W_k W_{ij}^W + A^B_k B_{ij} - b_k = A^W_k W_{ij} - (A^B_k B_{ij} + b_k). \quad (24)$$

Specifically, the last term above indicates that item difficulty (indexed by $k$) is a multivariate function that depends on cluster (indexed by $i$). The more difficult assessment of invariance concerns whether there are random item effects across groups of clusters with respect to $A^B$. Similarly, DIF studies concerning item discrimination are more complex to carry out and evaluate than DIF studies of item difficulty. Rather than a problem, differences in dimensional structure would seem to be a fertile area for further investigation.

Robustness Studies

The limited simulations provided in this study demonstrate feasibility of the approach and results that show minimally biased for main inferences despite some violations of assumptions. But the simulations do not demonstrate the procedure is robust to a wide range of violations of various kinds. Thus, additional research is clearly necessary to determine how inferences are limited, and to which violations the aggregate method is sensitive. For example, it would seem to be a strong assumption that the within-cluster variances are identically equal to unity. An ensuing simulation study might demonstrate that inferences about factor loadings or factor scores begin to break down as the (superpopulation) variances within clusters span a wider range than suggested by a simple chi-square density function. Likewise, a robustness study might assume different levels of intraclass correlations. As suggested earlier, robustness studies might also focus on factorial invariance across subpopulations or groups of clusters.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.
Funding
The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Science Foundation under Grant No. 1028230.

Supplementary Material
The online appendices are available at http://jeb.sagepub.com/supplemental

References


**Authors**

GREGORY CAMILLI is a professor at Rutgers Graduate School of Education, 10 Seminary Place, New Brunswick, NJ 08901; e-mail: greg.camilli@gse.rutgers.edu. His interests are in measurement, educational policy, and comparative education.

JEAN-PAUL FOX is a professor at the University of Twente, Faculty of Behavioral, Management, and Social Science, in the Department of Research Methodology, Measurement and Data Analysis, P.O. Box 217, 7500 AE Enschede, The Netherlands; e-mail: G.J.A.Fox@utwente.nl. His interest is focused on complex latent variable modeling for higher dimensional problems.

Manuscript received August 6, 2014
First revision received January 30, 2015
Second revision received March 20, 2015
Accepted March 27, 2015