British Journal of Mathematical and Statistical Psychology (2015), 68, 410–433

The British Psychological Society

www.wileyonlinelibrary.com

Multilevel multidimensional item response model with a multilevel latent covariate

Sun-Joo Cho[1]* and Brian Bottge[2]

[1]Vanderbilt University, Nashville, Tennessee, USA
[2]University of Kentucky, Lexington, Kentucky, USA

In a pre-test–post-test cluster randomized trial, one of the methods commonly used to detect an intervention effect involves controlling pre-test scores and other related covariates while estimating an intervention effect at post-test. In many applications in education, the total post-test and pre-test scores, ignoring measurement error, are used as response variable and covariate, respectively, to estimate the intervention effect. However, these test scores are frequently subject to measurement error, and statistical inferences based on the model ignoring measurement error can yield a biased estimate of the intervention effect. When multiple domains exist in test data, it is sometimes more informative to detect the intervention effect for each domain than for the entire test. This paper presents applications of the multilevel multidimensional item response model with measurement error adjustments in a response variable and a covariate to estimate the intervention effect for each domain.

1. Introduction

Pre-test–post-test cluster randomized trials are common in educational intervention studies because researchers cannot control students' class assignment, although random assignment sometimes occurs at the student level as well (Raudenbush, 1997). Thus, study designs have multilevel data in which teachers, classes or schools are randomly assigned to intervention. One of the commonly used methods for detecting an intervention effect involves controlling pre-test scores and other related covariates when estimating the intervention effect at post-test (e.g., Aitkin & Longford, 1986; Goldstein, 2003, ch. 2).

Students' ability scores at pre-test and post-test are vulnerable to measurement error,[1] and ability is often measured with a set of items. It has been shown that ignoring measurement error in a response variable (i.e., post-test scores) and a covariate (i.e., pre-test scores) leads to biased parameter estimates. The bias is due to attenuation from measurement error in the response variable (e.g., Carroll, Ruppert, Stefanski, & Crainiceanu, 2006, ch. 15; Fox, 2004). Measurement error in the covariate is also responsible for biased parameter estimates and loss of power to detect relationships among variables (Bryk & Raudenbush, 1992; Carroll *et al.*, 2006; Fox & Glas, 2003; Goldstein, Kounali, & Robinson, 2008; Rabe-Hesketh, Skrondal, & Pickles, 2004). In

*Correspondence should be addressed to Sun-Joo Cho, Peabody College of Vanderbilt University, 230 Appleton Place, Nashville, TN 37203-5721, USA (email: sj.cho@vanderbilt.edu).

[1] In this study, we use the term 'measurement error' to refer to random measurement error, not systematic measurement error.

DOI:10.1111/bmsp.12051

detecting an intervention effect controlling pre-test scores, the effects of pre-test scores can be biased in the presence of measurement error in pre-test scores (e.g., Lüdtke, Marsh, Robitzsch, & Trautwein, 2011). However, for the intervention effect, previous research has shown that covariate measurement error is a problem only for non-experimental designs with groups that differ in average covariate value in analysis of covariance (e.g., Culpepper & Aguinis, 2011; Porter & Raudenbush, 1987). When there is no group difference in pre-test scores, bias in the intervention effect estimate may not be of concern in the presence of measurement error in pre-test scores (Cho & Preacher, 2015). The assumption that there is no group difference in pre-test scores must be tested. Item response models can be used to model the relationship between ability and the set of individual items when ability cannot be measured perfectly.

In addition, students' outcomes in the evaluation of intervention studies often involve multiple domains even though the test is supposedly unidimensional. The multiple-domain design provides the possibility of detecting intervention effects for each domain and thus facilitates diagnostic interpretations of the results. To do so, separate unidimensional item response models can be fitted to obtain item response theory (IRT) scale scores and an intervention effect on the scale of each domain. However, this approach can lead to inaccurate results when the number of test items related to each domain is small (e.g., de la Torre, Song, & Hong, 2011).

Multilevel multidimensional item response models (MMIRMs; Muthén & Asparouhov, 2013; Rabe-Hesketh *et al.*, 2004) allow for explicitly modelling measurement error and IRT subscoring for multilevel data. The MMIRM provides the opportunity to model latent variables with multiple observed items to reduce the effects of measurement error. In addition, multiple latent variables for multiple domains are modelled, and the linear relationship between the domain-specific latent variables can be obtained at each level of multilevel data in the MMIRM.

Measurement error adjustment is achieved by applying the MMIRM to response variables and covariates. Up to this point, MMIRMs have been mainly applied to response variables (see Muthén & Asparouhov, 2013, sections 7 and 8). There are examples of researchers correctly accounting for measurement error in covariate(s) using unidimensional item response models (Battauz & Bellio, 2011; Fox & Glas, 2003). There are also a few examples of measurement error adjustment in response variables and covariates. Raudenbush and Sampson (1999) used a multilevel Rasch model to control for measurement error in both response variables and covariates. Rabe-Hesketh *et al.* (2004, equation 18, p. 180) specified the linear predictor in a generalized linear model for measurement error adjustment in response variables and covariates in multilevel data. When a measurement model is specified for both response variables (i.e., post-test scores) and covariates (i.e., pre-test scores), latent variables for the covariates are used to explain latent variables for response variables at each level of the multilevel data. This makes symmetric score mapping possible between post-test scores and pre-test scores.[2] However, to our knowledge, the multidimensional specification of the linear predictor with a logit link or probit link (two-parameter MMIRM) has not been applied to adjust measurement error in a response variable and a covariate.

When an MMIRM as a latent covariate [3] (i.e., a pre-test model) is added to an MMIRM as a response variable (i.e., a post-test model), other manifest covariates including a grouping

[2] The authors thank the reviewer of a previous version of this paper for clearly pointing out this modelling feature.
[3] We define the term *latent covariate* as a covariate measured with measurement error, in contrast to a *manifest covariate* measured without measurement error.

variable for the intervention (i.e., a control group vs. a treatment group) and demographic variables can be added to account for the ability parameters of the post-test model in a structural model. All parameters in the measurement models and the structural model can be estimated simultaneously in explanatory item response modelling (De Boeck & Wilson, 2004) or generalized multilevel structural equation modelling (McDonald, 1993; Muthén, 1994; Rabe-Hesketh *et al.*, 2004).

The purpose of this paper is to present a model specification that includes a two-parameter MMIRM in which measurement error is corrected for a response variable and a covariate at *each* level of the multilevel data structure. The MMIRM in this study is an MMIRM with a *multilevel latent covariate* (MMIRM-MLC). The rest of this paper is organized as follows. First, we specify an MMIRM-MLC and describe parameter estimation. Then we present an empirical study for applications of the MMIRM-MLC, followed by a simulation study to evaluate an MMIRM-MLC and to compare its performance with other approaches using the total scores. We conclude with a summary and discussion.

2. MMIRM with a multilevel latent covariate

In this section an MMIRM-MLC is described, with a measurement model and a structural model, for binary responses. Crossed and nested data structures are possible in multilevel item response data at pre-test and post-test. If every item is offered to all individuals and every individual responds to all items, the item and individual classifications are found at the same level, and they are crossed. In addition to the crossed design, there is a multilevel design in which individuals (e.g., students) are nested with clusters (e.g., teachers). To frame this data structure within the multilevel literature (e.g., Bryk & Raudenbush, 1992), item responses at level 1 are cross-classified with individuals and items at level 2. Individuals are nested within clusters at level 3. The model description is limited to a *between-item design* in which an item is loaded on one dimension or latent variable for subscoring.

A measurement model, an MMIRM, for correct item responses at post-test (denoted by a subscript 2) is as follows, assuming that there is no evidence of measurement bias regarding clusters and groups (e.g., control and treatment groups):

$$P(y_{2jki} = 1|\boldsymbol{\theta}_{2jk}, \boldsymbol{\theta}_{2k}) = \Phi[\boldsymbol{\alpha}_{2i} \cdot (\boldsymbol{\theta}_{2jk} + \boldsymbol{\theta}_{2k}) - \boldsymbol{\beta}_{2i}], \quad (1)$$

where Φ denotes the standard normal cumulative distribution function, j is an index for an individual ($j = 1,\ldots,J$), k is an index for a cluster ($k = 1,\ldots,K$), i is an index for an item ($i = 1,\ldots,I$), d is an index for a dimension (i.e., domain) ($d = 1,\ldots,D$), $y_{2jki} = [y_{2jki1},\ldots,y_{2jkid},\ldots,y_{2jkiD}]'$ are item responses across domains at post-test, $\boldsymbol{\theta}_{2jk(D \times 1)} = [\theta_{2jk1},\ldots,\theta_{2jkd},\ldots,\theta_{2jkD}]'$ are multidimensional latent variables at level 2, $\boldsymbol{\theta}_{2k(D \times 1)} = [\theta_{2k1},\ldots,\theta_{2kd},\ldots,\theta_{2kD}]'$ are multidimensional latent variables at level 3, $\boldsymbol{\alpha}_{2i\,(I \times D)}$ are item slopes or item discrimination parameters at post-test, and $\boldsymbol{\beta}_{2i(I \times D)}$ are item intercepts or item difficulty parameters at post-test. $\boldsymbol{\theta}_{2jk}$ and $\boldsymbol{\theta}_{2k}$ are assumed to follow a multivariate normal distribution, $\boldsymbol{\theta}_{2jk} \sim MN(\mathbf{0}_{(D \times 1)},\Sigma_{1(D \times D)})$ and $\boldsymbol{\theta}_{2k} \sim MN(\mathbf{0}_{(D \times 1)},\Sigma_{2(D \times D)})$, respectively.

A measurement model, an MMIRM, for correct item responses at pre-test (denoted by a subscript 1) is as follows, assuming that there is no evidence of measurement bias regarding clusters and groups:

$$P(y_{1jki}|\boldsymbol{\theta}_{1jk}, \boldsymbol{\theta}_{1k}) = \Phi[\boldsymbol{\alpha}_{1i} \cdot (\boldsymbol{\theta}_{1jk} + \boldsymbol{\theta}_{1k}) - \boldsymbol{\beta}_{1i}], \tag{2}$$

where $y_{1jki} = [y_{1jki1},\ldots,y_{1jkid},\ldots,y_{1jkiD}]'$ are item responses across domains at pre-test, $\boldsymbol{\theta}_{1jk(D \times 1)} = [\theta_{1jk1},\ldots,\theta_{1jkd},\ldots,\theta_{1jkD}]'$ are multidimensional latent variables at level 2, $\boldsymbol{\theta}_{1k(D \times 1)} = [\theta_{1k1},\ldots,\theta_{1kd},\ldots,\theta_{1kD}]'$ are multidimensional latent variables at level 3, $\boldsymbol{\alpha}_{1i(I \times D)}$ are item slopes or item discrimination parameters at pre-test, and $\boldsymbol{\beta}_{1i(I \times D)}$ are item intercepts or item difficulty parameters at pre-test. $\boldsymbol{\theta}_{1jk}$ and $\boldsymbol{\theta}_{1k}$ are assumed to follow a multivariate normal distribution, $\boldsymbol{\theta}_{1jk} \sim MN(\mathbf{0}_{(D \times 1)},\Sigma_{3(D \times D)})$ and $\boldsymbol{\theta}_{1k} \sim MN(\boldsymbol{\mu}_{(D \times 1)},\Sigma_{4(D \times D)})$, respectively, where $\boldsymbol{\mu}_{(D \times 1)} = [\mu_1,\ldots,\mu_d,\ldots,\mu_D]'$ are intercepts of latent variables (i.e., grand mean).

A structural model for person parameters at level 2 (e.g., the student level) is as follows:

$$\boldsymbol{\theta}_{2jk} = \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1 \cdot \boldsymbol{\theta}_{1jk} + \sum_{n=1} \boldsymbol{\gamma}_{(n+1)} \cdot \boldsymbol{Z}_{jk.n} + \boldsymbol{\varepsilon}_{2jk}, \tag{3}$$

where $\boldsymbol{Z}_{jk.n(D \times 1)}$ is the nth covariate for an individual j nested with a cluster k at level 2, $\boldsymbol{\gamma}_{0(D \times 1)} = [\gamma_{01},\ldots,\gamma_{0d},\ldots,\gamma_{0D}]'$ are intercepts at level 2 (fixed to 0s to identify the model), $\boldsymbol{\gamma}_{1(D \times D)} = \text{diag}[\gamma_{11},\ldots,\gamma_{1d},\ldots,\gamma_{1D}]'$ are the effects of the pre-test score at level 2, $\boldsymbol{\gamma}_{(n+1)(D \times D)} = \text{diag}[\gamma_{(n+1)1},\ldots,\gamma_{(n+1)d},\ldots,\gamma_{(n+1)D}]'$ are the effects of covariates $\boldsymbol{Z}_{jk.n}$, and $\boldsymbol{\varepsilon}_{2jk(D \times 1)} = [\varepsilon_{2jk1},\ldots,\varepsilon_{2jkd},\ldots,\varepsilon_{2jkD}]'$ are residuals of post-test latent scores at level 2, assumed to follow $MN(\mathbf{0}_{(D \times 1)},\Sigma_{5(D \times D)})$.

A structural model for person parameters at level 3 (e.g., the teacher level) is as follows:

$$\boldsymbol{\theta}_{2k} = \boldsymbol{\delta}_0 + \boldsymbol{\delta}_1 \cdot \boldsymbol{\theta}_{1k} + \boldsymbol{\delta}_2 \cdot \boldsymbol{TRT}_k + \sum_{m=1} \boldsymbol{\delta}_{(m+2)} \cdot \boldsymbol{Z}_{k.m} + \boldsymbol{\varepsilon}_{2k}, \tag{4}$$

where $\boldsymbol{TRT}_{k(D \times 1)}$ is a covariate of an intervention condition with a value of 0 for members of the control group and a value of 1 for members of the treatment group, $\boldsymbol{Z}_{k.m(D \times 1)}$ is the mth covariate for a cluster k at level 3, $\boldsymbol{\delta}_{0(D \times 1)} = [\delta_{01},\ldots,\delta_{0d},\ldots,\delta_{0D}]'$ are intercepts at level 3 (i.e., grand mean), $\boldsymbol{\delta}_{1(D \times D)} = \text{diag}[\delta_{11},\ldots,\delta_{1d},\ldots,\delta_{1D}]'$ are the effects of the pre-test score at level 3, $\boldsymbol{\delta}_{2(D \times D)} = \text{diag}[\delta_{21},\ldots,\delta_{2d},\ldots,\delta_{2D}]'$ are the intervention effects at level 3, $\boldsymbol{\delta}_{(m+2)(D \times 1)} = [\delta_{(m+2)1},\ldots,\delta_{(m+2)d},\ldots,\delta_{(m+2)D}]'$ are the effects of covariates $\boldsymbol{Z}_{k.m}$, and $\boldsymbol{\varepsilon}_{2k(D \times 1)} = [\varepsilon_{2k1},\ldots,\varepsilon_{2kd},\ldots,\varepsilon_{2kD}]'$ are residuals of post-test latent scores at level 3, assumed to follow $\boldsymbol{\varepsilon}_{2k(D \times 1)} \sim MN(\mathbf{0}_{(D \times 1)},\Sigma_{6(D \times D)})$.

Adding the two structural models (equations 3 and 4) to the measurement model for a post-test (equation 1), the model for correct item responses across domains ($y_{2jki} = [y_{2jki1},\ldots,y_{2jkid},\ldots,y_{2jkiD}]'$) leads to the following:

$$\begin{aligned} P(y_{2jki}) = \Phi[\boldsymbol{\alpha}_{2i} \cdot \{&(\boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1 \cdot \boldsymbol{\theta}_{1jk} + \sum_{n=1} \boldsymbol{\gamma}_{(n+1)} \cdot \boldsymbol{Z}_{jk.n} + \boldsymbol{\varepsilon}_{2jk}) \\ &+ (\boldsymbol{\delta}_0 + \boldsymbol{\delta}_1 \cdot \boldsymbol{\theta}_{1k} + \boldsymbol{\delta}_2 \cdot \boldsymbol{TRT}_k + \sum_{m=1} \boldsymbol{\delta}_{(m+2)} \cdot \boldsymbol{Z}_{k.m} + \boldsymbol{\varepsilon}_{2k})\} - \boldsymbol{\beta}_{2i}]. \end{aligned} \tag{5}$$

To identify the model, the $\boldsymbol{\gamma}_0$ are set to 0s, and variances in $\Sigma_{3(D \times D)}$ and $\Sigma_{5(D \times D)}$ (i.e., variances at the student level for the pre-test and residual variances at the student level for the post-test, respectively) are set to 1s. Alternatively, the item discrimination for one of the items (e.g., the first item) in each dimension can be set to 1 instead of setting variances to 1 to identify the scale unit of the parameters. Variances at the teacher level can be estimated for the pre-test and post-test because the same item discriminations are used

over levels (assuming no cluster bias). See the online supporting information (Appendix S1) for a diagram depicting the MMIRM-MLC for person parameters with two domains.

2.1. Comparisons with other approaches to measurement error adjustment

Measurement error adjustment using the MMIRM-MLC is different from measurement error adjustment methods in previous structural modelling approaches in which specific assumptions are made about the distributional structure of the unobserved variables. A description of those differences follows.

First, measurement error in the MMIRM-MLC is adjusted for a response variable and a covariate simultaneously, as in Rabe-Hesketh *et al.* (2004) and Raudenbush and Sampson (1999). Specifically, this simultaneous approach allows us to detect the group difference on the error-free latent variable scale (i.e., the ability parameter in IRT is equal to an (unbiased) estimator minus (random) error) at post-test, by controlling for the possible measurement error in the pre-test scores and by mapping pre-test scores and post-test scores on the latent variable scales. However, in previous studies, measurement error was mainly adjusted for the response variable (e.g., Fox, 2004) or for the covariate (e.g., Battauz & Bellio, 2011; Carroll *et al.*, 2006; Fox & Glas, 2003; Goldstein *et al.*, 2008). That is, in these previous applications, either a measurement model for the response variable (e.g., equation 1 or a classical true score model) or a measurement model for the covariate (e.g., equation 2 or a classical true score model) was used.

Second, a set of multiple items is used to correct for measurement error in the covariate using item response models in the MMIRM-MLC (see equations 1 and 2). That is, the set of multiple items at level 1 in the MMIRM-MLC is used for correcting for measurement error at the individual level and at the cluster level. This approach is different from previous approaches to correcting for measurement error in the covariate, including Carroll *et al.* (2006) and Goldstein *et al.* (2008). These previous studies used a classical true score model for total scores (only at the individual level).

Third, measurement error adjustment in the MMIRM-MLC is done at each level of the multilevel data. Specifically, multiple items for each domain (indicated by d) are modelled for a latent variable at level 2 (θ_{1jkd}) and a latent variable at level 3 (θ_{1kd}) to correct for measurement error in the pre-test scores. Further, multiple items for each domain are used for a latent variable at level 2 (θ_{2jkd}) and a latent variable at level 3 (θ_{2kd}) to correct for measurement error in the post-test scores. The group differences, the intervention effects ($\boldsymbol{\delta}_2$ in equation 5), can be detected on the error-free latent variable scale, θ_{2kd}. Raudenbush and Sampson (1999) used multiple items at level 1 to measure constructs at level 2 within level 3 as in the MMIRM-MLC. However, they did not include item discriminations at level 2 (such as $\boldsymbol{\alpha}_{1i}$ and $\boldsymbol{\alpha}_{2i}$ in the MMIRM-MLC) or regressions among the latent variables (such as γ_1 and $\boldsymbol{\delta}_1$ in the MMIRM-MLC).

2.2. Measurement invariance test

In multiple-measurement (or longitudinal) multilevel data arising from multiple groups, there are at least three sources of measurement invariance to test: across time, across clusters, and across groups (e.g., control and treatment groups). The measurement invariance assumption across time points is not necessary when a pre-test score is used as a proxy variable for unobserved factors that predict or explain future attributes (e.g., Lockwood & McCaffrey, 2014). Further, it is possible that item discrimination(s) can be different for an individual-level latent variable and for a cluster-level latent variable in

multilevel item response models. This possibility, called cluster bias (Jak, Oort, & Dolan, 2013), can be investigated by testing whether item discriminations are equal over levels. Finally, invariance across groups is necessary for comparing group means (Bejar, 1980). Item response models to test cluster bias and group bias are described in the online supporting information (Appendix S2).

Two models are compared to test cluster bias: (1) a cluster invariance model, in which item discriminations over levels 2 and 3 are the same; and (2) a cluster bias model, in which item discriminations over levels 2 and 3 are different. Three invariance models are compared to investigate the measurement invariance across groups (e.g., Vandenberg & Lance, 2000; Widaman & Reise, 1997): (1) a configural invariance model, in which all item parameters are estimated simultaneously in each group under the same factor structures; (2) a weak invariance model, in which only discrimination parameters are constrained to be equal across groups; and (3) a strong invariance model, in which all item parameters are constrained to be equal across groups.

3. Parameter estimation and model evaluation

Bayesian analysis was chosen to fit MMIRM-MLCs and the (multigroup) multilevel longitudinal item response model to test measurement invariance assumptions. In (hierarchical) Bayesian analysis, it is possible to sample complex and high-dimensional posterior densities with Markov chain Monte Carlo (MCMC) methods through sampling from the conditional distributions of parameters without numerical integration. In this study, WinBUGS 1.4.3 (Spiegelhalter, Thomas, Best, & Lunn, 2003) was used to implement MCMC.

For the MMIRM-MLC, joint posterior distributions for parameters $\vartheta = \{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\gamma_0, \gamma_1,\gamma_{(n+1)d},\boldsymbol{\delta}_0,\boldsymbol{\delta}_1,\boldsymbol{\delta}_{(m+2)},\boldsymbol{\mu},\boldsymbol{\theta}_{1jk},\boldsymbol{\theta}_{1k},\boldsymbol{\varepsilon}_{2jk},\boldsymbol{\varepsilon}_{2k},\Sigma_3,\Sigma_4,\Sigma_5,\Sigma_6\}$ can be rewritten as

$$
\begin{aligned}
P(\vartheta|\boldsymbol{y}_{1jki},\boldsymbol{y}_{2jki}) \propto & P(\boldsymbol{y}_{1jki}|\vartheta)P(\boldsymbol{y}_{2jki}|\vartheta) \\
& \cdot \{P(\boldsymbol{\alpha}_1)P(\boldsymbol{\alpha}_2)P(\boldsymbol{\beta}_1)P(\boldsymbol{\beta}_2)P(\gamma_0)P(\gamma_1)P(\gamma_{(n+1)})P(\boldsymbol{\delta}_0)P(\boldsymbol{\delta}_1)P(\boldsymbol{\delta}_{(m+2)})P(\boldsymbol{\mu})\} \\
& \cdot \{P(\boldsymbol{\theta}_{1jk}|\mathbf{0},\Sigma_3)P(\boldsymbol{\theta}_{1k}|\mathbf{0},\Sigma_4)P(\boldsymbol{\varepsilon}_{2jk}|\mathbf{0},\Sigma_5)P(\boldsymbol{\varepsilon}_{2k}|\mathbf{0},\Sigma_6)\} \\
& \cdot \{P(\Sigma_3)P(\Sigma_4)P(\Sigma_5)P(\Sigma_6)\},
\end{aligned}
\tag{6}
$$

where $P(\boldsymbol{y}_{1jki}|\vartheta)$ is a likelihood function of item responses across domains for pre-test, $P(\boldsymbol{y}_{2jki}|\vartheta)$ is a likelihood function of item responses across domains for post-test, the probabilities in the first braces indicate prior distributions of fixed parameters, the probabilities in the second braces indicate prior distributions of latent variables, and the probabilities in the third braces indicate hyperprior distributions of population parameters of the latent variables. A similar specification was also applied to the (multigroup) multilevel longitudinal item response model to test measurement invariance assumptions across clusters and groups.

Priors for all fixed effects in a structural model for person parameters (except γ_0 to identify the model) and item difficulty parameters were set as $N(0,0.1)$ in WinBUGS. Item discrimination parameters were set to $N(0,1)$ truncated at 0 [4] for $\boldsymbol{\alpha}_{1i}$ and $\boldsymbol{\alpha}_{2i}$, respectively,

[4] The specification in WinBUGS is $N(0,1)I(0,)$ where 1 is a variance.

to have stable item discrimination parameter estimates (e.g., Béguin & Glas, 2001, for the normal ogive multidimensional item response model). To match the priors on $\Sigma_{3(D \times D)}$ and $\Sigma_{5(D \times D)}$ with the model identification constraints, variances in the variance–covariance matrix were set to 1 and the priors on correlation coefficient parameters were set to Uniform(−1,1). Prior and hyperprior distributions for other parameters were specified in WinBUGS as follows:

$$\begin{aligned}
\boldsymbol{\theta}_{1k} &\sim MN(\mathbf{0}_{(D\times 1)}, \Sigma_{4(D\times D)}),\\
\boldsymbol{\varepsilon}_{2k} &\sim MN(\mathbf{0}_{(D\times 1)}, \Sigma_{6(D\times D)}),\\
\Sigma_{4(D\times D)} &\sim \text{Wishart}(R, \nu), R = I_D, \nu = D, \text{and}\\
\Sigma_{6(D\times D)} &\sim \text{Wishart}(R, \nu), R = I_D, \nu = D.
\end{aligned}$$

I_D denotes the unit matrix of size D, and the degrees of freedom ν in the Wishart distribution are set to D as the rank of $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$ to represent vague prior knowledge (the mean and variance in the prior distribution on elements in $\Sigma_{4(D \times D)}$ and $\Sigma_{4(D \times D)}$ are 2 with R=I_D). Similar priors and hyperpriors for fixed parameters and random effects were chosen for item and person parameters of the (multigroup) multilevel longitudinal item response model to test measurement invariance assumptions across clusters and groups.

In order to ensure that stable parameter estimates are obtained, Gelman and Rubin's (1992) method was chosen as implemented in WinBUGS. Using the results of the convergence checking, initial samples are discarded ('burn-in') and posterior means or medians and standard deviations (i.e., Bayesian standard errors) calculated from subsequent iterations.

3.1. Bayesian model fit

Competing models (i.e., measurement invariance models, unidimensional vs. multidimensional model) were compared using a relative fit criterion, the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002; see also Verhagen & Fox, 2013). A smaller DIC represents a better fit of the model, and a difference of <5 or 10 units between models does not provide sufficient evidence for favouring one model over another (Spiegelhalter *et al.*, 2003). The DIC can be calculated easily by specifying the log-likelihood along with a model specification in WinBUGS.

In addition, the adequacy of the fit of the MMIRM is evaluated by comparing observed and posterior predictive score frequencies (e.g., Béguin & Glas, 2001) with posterior predictive model checking (Rubin, 1984). In addition to the overall model evaluation using the posterior predictive frequencies, item fit and person fit were considered individual checks. Standardized residuals (Spiegelhapter, Thomas, Best, & Gilks, 1996) were considered as a discrepancy measure. Item fit was calculated as the mean of the standardized residuals over persons, and person fit was calculated as the mean of the standardized residuals over items. Posterior predictive p-values (*ppp*-values; Meng, 1994) for the person fit (Glas & Meijer, 2003) and item fit (Sinharay, 2005) were calculated. Values around .5 indicate that a person or an item fits well to the data while values close to zero or 1 indicate misfit (Gelman & Meng, 1996). We consider *ppp*-values smaller than .025 or larger than .975 as extreme values indicative of misfit at the 5% level (e.g., Sinharay, 2005).

4. Empirical illustration

Data for the current study were gathered as part of a larger efficacy trial for enhanced anchored instruction (EAI). The experimental instruction was designed to improve the mathematics skills of middle and high school students, especially those with learning difficulties in maths (MD). The design of the efficacy trial was a pre-test–post-test cluster randomized trial. Schools, rather than classes or students, were randomly assigned to EAI and business as usual (BAU) because the research team did not have control over the students' class assignment. In this illustration, we evaluate an instructional intervention called EAI and its impact on students by showing the effect of intervention using an MMIRM-MLC. Our purpose for conducting the analysis was to answer the following question: If the intervention effect is detected, is it possible to interpret it across cognitive skill areas?

4.1. Teacher and student samples

Twenty-four urban and rural middle schools in the southeastern United States participated in the study. Half were randomly assigned to EAI and BAU. Each school had one participating inclusive maths classroom, although one school had two participating classrooms. Teachers in both conditions were comparable in terms of gender (mostly female), ethnicity (mostly white), education level (well educated), and years of experience (Bottge, Ma, Gassaway, Toland, Butler, & Cho, 2014). In our study, one inclusive maths class from each school was sampled, with the exception of one school that had two inclusive maths classes. Therefore, a two-level data structure (students nested within 25 teachers) was used because there was only one school for which we needed to be concerned about clustering at the school level. The smallest number of students analysed for a teacher was 7, and the largest was 28. The average cluster size was 17.84.

Roughly equal numbers of students in each condition had an identified MD: 62 (28%) of 223 in EAI and 72 (29%) of 248 in BAU. Of the initial sample, 25 students did not respond to all items in the pre-test or post-test. As a result, 232 BAU (29% MD) and 214 EAI (26% MD) remained in the final sample. Based on chi-square tests of equal proportions, students were comparable across instructional conditions in gender, ethnicity, subsidized lunch, and disability area, and teachers were comparable in both conditions in terms of gender (mostly female), ethnicity (mostly white), education level (well educated), and years of experience (Bottge *et al.*, 2014). Bottge *et al.* (2014) found that there was no EAI and BAU group difference on the pre-test total score scales.

4.2. Measure: Fraction computation test

The researcher-developed test, the fraction computation test, administered at the pre-test and post-test, was used in the current study to illustrate MMIRM-MLC. The test comprised 20 items assessing students' ability to manually add and subtract fractions. Item features differed in several ways: (1) *addition* or *subtraction*; (2) *like* denominators ($\frac{1}{4} + \frac{3}{4}$) or *unlike* denominators ($8\frac{2}{9} + 2\frac{1}{2}$); (3) *simple* fractions ($\frac{3}{8} + \frac{3}{4}$) or *mixed* numbers ($4\frac{1}{16} + \frac{1}{8} + \frac{1}{2}$); and (4) *two stacks* ($\frac{3/4}{1/2}$) or *three stacks* (i.e., one more stack in the two-stack example). There were a total of 42 points on the test. For 18 of the 20 items, students could earn 0, 1 or 2 points. On two items, students could earn 3 points if they simplified the answer (i.e., revised the fraction to simple terms). Inter-rater agreement was 99% on the pre-test and 97% on the post-test.

Less than 1% of students in the sample received partial scores (i.e., score 1 for 18 of the items and scores 1 or 2 for two of the items) on any of the items on the test. Thus, in this paper, binary responses were considered, 1 for correct responses and 0 for incorrect responses. Partial scores were also considered as incorrect responses. There were no missing item responses in the final sample of 446 students for analysis.

4.3. Analysis and results

To answer the research question using the MMIRM-MLC, our analysis proceeded as follows. All codes, including the model specification in WinBUGS used in the current analyses, are available from the first author upon request.

4.3.1. Step 1: Determining distinct domains

As shown in Table 1, each item had four item attributes. In order to find the most distinct item feature for domain scoring, we compared a set of exploratory factor analyses using polychoric correlations with Bayes estimator (GIBBS(PX1) option) at each time point

Table 1. Fit Indices from Exploratory Factor Analyses Extracting 1 and 2 Factors and (GEOMIN Rotated) Factor Loadings for a 2-Factor Solution.

					Model fit			
					Pre-test		Post-test	
					1-Factor	2-Factor	1-Factor	2-Factor
ppp-value					0.399	0.443	0.401	0.453
					Factor loadings			
	Attributes				Pre-test		Post-test	
Item	Operation	Denominator	Type	Stacks	Factor 1	Factor 2	Factor 1	Factor 2
1	Addition	Like	Simple	2	**0.640**	0.354	**0.651**	0.080
2	Addition	Like	Simple	2	**0.610**	**0.216**	**0.678**	0.201
3	Addition	Unlike	Simple	2	−0.135	**1.048**	**−0.265**	**1.075**
4	Addition	Unlike	Simple	2	−0.186	**1.067**	**−0.231**	**1.029**
5	Addition	Unlike	Simple	2	0.000	**0.965**	**−0.189**	**1.024**
6	Addition	Unlike	Simple	2	−0.026	**0.985**	**−0.214**	**1.054**
7	Addition	Unlike	Mixed	2	**−0.132**	**1.034**	0.007	**0.940**
8	Addition	Unlike	Mixed	2	0.041	**0.912**	0.087	**0.900**
9	Addition	Unlike	Mixed	2	0.007	**0.946**	0.062	**0.930**
10	Addition	Unlike	Mixed	2	0.032	**0.939**	0.007	**0.924**
11	Addition	Unlike	Simple	3	0.089	**0.902**	0.063	**0.886**
12	Addition	Unlike	Simple	3	0.092	**0.902**	0.052	**0.921**
13	Addition	Unlike	Mixed	3	0.065	**0.901**	0.084	**0.916**
14	Addition	Unlike	Mixed	3	0.004	**0.924**	0.005	**0.924**
15	Subtraction	Like	Simple	2	**0.888**	0.032	**0.883**	**0.087**
16	Subtraction	Unlike	Simple	2	0.137	**0.871**	**0.160**	**0.851**
17	Subtraction	Like	Mixed	2	**0.804**	0.097	**0.731**	0.085
18	Subtraction	Unlike	Mixed	2	**0.456**	**0.698**	0.002	**0.752**
19	Subtraction	Unlike	Mixed	2	0.235	**0.772**	**0.212**	**0.804**
20	Subtraction	Unlike	Mixed	2	**0.397**	**0.557**	0.096	**0.737**

Note. Bold factor loadings are significant at 5% level.

using Mplus version 7.11 (Muthén & Muthén, 1998–2014). The model fit of the exploratory factor analyses was evaluated based on posterior predictive model checking with a summary measure of fit, the likelihood-ratio chi-square statistic. Corresponding *ppp*-values were calculated for each factor solution; *ppp*-values around .5 indicate that the observed pattern would likely be seen in replications of the data if the model were true.

Table 1 shows the model fit results at each time point for a one-factor and two-factor solution and (GEOMIN rotated) factor loadings for a two-factor solution. According to the *ppp*-values, the one-factor model provided a good fit to the data according to the criteria at each time point. However, there was an improvement in model fit with the two-factor model at each time point. Shifting from the one-factor to the two-factor model produced noteworthy decreases in residual variances for four *like* items that loaded on the first factor, especially at the post-test. Factor loadings were clearly clustered regarding *like* items versus *unlike* items. As reported in Table 1, the *like* items (items 1, 2, 15 and 17) were highly loaded on factor 1 while the *unlike* items were highly loaded on factor 2 at the pre-test and post-test. Moderate (GEOMIN) factor correlations of .585 and .497 for the pre-test and post-test, respectively, indicated that two factors can provide two scores with distinct meaning. Based on these results, we chose the two-factor model with a between-item design where an item loaded on the *like* factor or *unlike* factor for domain scoring. When there is evidence of a second dimension on a specific skill domain, having a two-factor model yields diagnostic interpretations as compared to a one-factor model.

4.3.2. Step 2: Selecting the measurement model

Intraclass correlations (ICCs) were calculated to investigate the multilevel structure of the data using the data at each time point. The ICC for the observed outcomes for each item (e.g., Muthén & Asparouhov, 2013) ranged from .058 to .297 for the pre-test and from .071 to .347 for the post-test, based on results of the two-parameter multilevel unidimensional normal ogive model at each time point. A common rule of thumb is that ICCs over .05 indicate the necessity of multilevel analysis (e.g., Jak *et al.*, 2013). According to the rule of thumb, there is non-ignorable dependency due to clusters (teachers). Accordingly, the MMIRM was chosen as a (multilevel) measurement model for the pre-test and post-test.

Table 2 reports summary information about the standard deviation (i.e., Bayesian standard error) of θ estimates from MMIRM and within and between reliability of the total scores (Geldhof, Preacher, & Zyphur, 2014) for each domain at pre-test and post-test. This information presents evidence that there was non-ignorable measurement error on both the latent variable scale and the total score scale.

4.3.3. Step 3: Checking measurement invariance

From step 1, a two-factor model was chosen to provide diagnostic interpretations based on the specific skill domain, even though there is evidence that the one-factor model fitted relatively well compared to the two-factor model. For measurement invariance checking, the one-factor model was estimated to check the measurement invariance over the clusters (i.e., teachers) and groups (i.e., BAU vs. EAI, non-MD vs. MD).

Table S1 in the online supporting information presents the measurement models, their constraints, and DIC values for three invariance models for clusters and groups. The 'burn-in' period ranged from 4,000 to 6,000 for invariance models in the MCMC analyses. Posterior means were used for calculating the DIC. Differences in the DIC values between cluster bias and cluster invariance models were <5, so that the cluster invariance model

Table 2. Measurement error information for IRT scale scores of MMIRM-MLC and for total scores

	Pre-test		Post-test	
	Like	*Unlike*	*Like*	*Unlike*
IRT-based: Descriptive information for standard deviation of θ estimates				
Student level				
Mean	0.80	0.49	0.71	0.45
SD	0.12	0.26	0.18	0.24
Min.	0.44	0.33	0.43	0.35
Max.	0.99	0.75	1.09	0.76
Teacher level				
Mean	0.77	0.45	0.58	0.39
SD	0.06	0.07	0.05	0.07
Min.	0.61	0.62	0.48	0.31
Max.	0.83	0.78	0.63	0.52
Total score-based				
Within reliability	0.43	0.69	0.50	0.70
Between reliability	0.56	0.69	0.51	0.74

was chosen as the simpler model. Given the cluster invariance model, group invariance tests were investigated. A weak invariance model was chosen for BAU versus EAI and non-MD versus MD.

Whether BAU and EAI or non-MD and MD can be scored and compared on the same scale in the presence of weak invariance violation was checked by comparing the correlations between the scores from the two MMIRM-MLC models (without any manifest covariates) with weak invariance and strong invariance assumptions. The correlation coefficients of the scores from the two MMIRM-MLC models for BAU and EAI and for non-MD and MD were highly correlated (>.927). This indicates that the relative ordering of persons' scores did not change much when measurement weak invariance was ignored for BAU and EAI or non-MD and MD. In addition, the results in the group mean differences (i.e., BAU and EAI or non-MD and MD) were similar between the two MMIRM-MLC models with weak invariance and strong invariance assumptions. Thus, in the following analysis, a strong invariance model for BAU and EAI or non-MD and MD was assumed.

4.3.4. Step 4: Adding covariates to the measurement model and model evaluation

Now the MMIRM-MLC was fitted to answer the research question by adding an intervention condition covariate to the measurement model. The model is called MMIRM-MLC model 1. MMIRM-MLC model 2 is MMIRM-MLC model 1 plus student-level and teacher-level demographic information.

A burn-in of 4,000 iterations was used for all parameters of MMIRM-MLC models 1 and 2, based on Gelman and Rubin's (1992) statistic with three chains. The 10,000 post-burn-in iterations were obtained to calculate posterior moments. Monte Carlo errors for all parameters in all analyses were less than about 5% of the sample standard deviation. All 95% posterior intervals included the observed data, which indicates that MMIRM-MLC models 1 and 2 was appropriate for the data. The *ppp*-values for all items at pre-test and post-test were between .025 and .075, indicating that the items were a good fit to the data. In MMIRM-MLC model 1, there were 8% and 7% of persons with *ppp*-values >.075 for the pre-test and post-test, respectively, indicating a misfit in this model. In MMIRM-MLC model

2, there were 8% and 6% of persons with *ppp*-values >.075 for the pre-test and post-test, respectively. They all were at the lower end of the score distribution.

Table 3 shows the item parameter estimates of MMIRM-MLC model 1.[5] Item parameter estimates of MMIRM-MLC model 2 were similar to those of MMIRM-MLC model 1. At the pre-test and post-test, items vary in terms of item discriminations and difficulties, and items that have a *like* denominator were less discriminating and less difficult than items that have an *unlike* denominator.

4.4. Answers to research question

Table 4 presents the results of MMIRM-MLC models 1 and 2 for person parameters. The change in significance and magnitude of the intervention effect was small when other demographic information for the students and teachers was added to MMIRM-MLC models 1 and 2. For illustration purposes, the results of MMIRM-MLC model 2 with two student-level covariates (MD and gender) and one teacher-level covariate (years of teaching special education) are shown in Table 4. Covariates were coded as follows: BAU (coded as 0) and EAI (coded as 1) groups, non-MD students (coded as 0) and MD students (coded as 1), female students (coded as 0) and males students (coded as 1), and mean-centred years of teaching general education (M = 11.1, SD = 7.7).

In Table 4, 'L.TRT' and 'U.TRT' represent the estimated difference between the means of the EAI and BAU post-test scores for a *like* domain and an *unlike* domain, respectively, adjusted for the pre-test scores on the post-test scores. In MMIRM-MLC model 1, the pre-test score effects on the *like* domain and the *unlike* domain were statistically significant at the student level and at the teacher level. Significant intervention effects were found for *like* and *unlike* domains (i.e., 0.890, credible interval [CI] [0.506, 1.499], for the *like* domain and 1.039, CI [0.601, 1.471], for the *unlike* domain). Specifically, the EAI group performed 0.890 higher than the BAU group for the *like* domain and the EAI group performed 1.039 higher than the BAU group for the *unlike* domain. In MMIRM-MLC model 2, the effects of student-level and teacher-level covariates were not significant and the effects of pre-test scores and the group difference between the EAI and BAU groups were similar to those of MMIRM-MLC model 1.

4.5. Result comparisons across different approaches for measurement error treatment

An MMIRM with multilevel manifest covariate (MMC) and a multilevel model (MM, specifically a multilevel multivariate random intercept model) with MLC were fitted to the same empirical data to show the consequences of ignoring measurement error in a covariate or a response variable. Measurement error in pre-test scores (covariate) is ignored in the MMIRM-MLC, whereas measurement error in post-test scores (response variable) is ignored in the MM-MLC. In the MMC of the MMIRM-MMC, pre-test total scores ($\sum_{i=1}^{I} y_{1jkid} = y_{1jk.d}$) were decomposed into within pre-test total scores ($y_{1jk.d} - y_{1.k.d}$, where $y_{1.k.d}$ is a cluster mean) and between pre-test total scores ($y_{1.k.d}$) for each domain. For the multilevel (linear) model, post-test total scores ($\sum_{i=1}^{I} y_{2jkid} = y_{2jk.d}$) were used for each domain. In these two models, pre-test scores and an intervention condition were considered covariates as in MMIRM-MLC model 1.

[5] As shown in equations 1 and 2, the item parameters have an $I \times D$ vector. With a between-item design, all items have one set of item parameters for the dimension d. In the table, the item parameter estimates are presented in one column for simplicity.

Table 3. Item attributes and item parameter estimates (95% CIs) of MMIRM-MLC model 1

		Attributes			Pre-test		Post-test	
Item	Operation	Denominator	Type	Stacks	α_1	β_1	α_2	β_2
1	Addition	Like	Simple	2	1.11 (0.87, 1.38)	−1.27 (−2.14, −0.35)	0.80 (0.57, 1.09)	−2.04 (−2.63, −1.47)
2	Addition	Like	Simple	2	1.28 (0.98, 1.57)	−1.62 (−2.55, −0.66)	1.21 (0.89, 1.62)	−2.59 (−3.33, −1.92)
3	Addition	Unlike	Simple	2	2.05 (1.72, 2.41)	−0.10 (−1.19, 0.89)	1.80 (1.49, 2.13)	−0.84 (−1.63, −0.09)
4	Addition	Unlike	Simple	2	1.83 (1.56, 2.17)	−0.13 (−1.15, 0.81)	1.38 (1.13, 1.70)	−0.48 (−1.13, 0.13)
5	Addition	Unlike	Simple	2	1.98 (1.63, 2.38)	0.19 (−0.92, 1.20)	1.57 (1.31, 1.84)	0.05 (−0.67, 0.75)
6	Addition	Unlike	Simple	2	2.02 (1.66, 2.36)	−0.12 (−1.20, 0.87)	1.91 (1.57, 2.33)	−0.43 (−1.27, 0.38)
7	Addition	Unlike	Mixed	2	1.87 (1.55, 2.23)	0.18 (−0.86, 1.14)	1.69 (1.38, 2.02)	−0.02 (−0.76, 0.65)
8	Addition	Unlike	Mixed	2	1.73 (1.43, 2.09)	0.45 (−0.55, 1.38)	1.58 (1.29, 1.98)	0.25 (−0.40, 0.88)
9	Addition	Unlike	Mixed	2	1.83 (1.52, 2.17)	0.33 (−0.72, 1.29)	1.70 (1.39, 2.02)	0.30 (−0.54, 1.12)
10	Addition	Unlike	Mixed	2	1.87 (1.53, 2.27)	0.23 (−0.88, 1.23)	1.49 (1.19, 1.80)	0.29 (−0.38, 0.92)
11	Addition	Unlike	Simple	3	1.91 (1.59, 2.27)	−0.22 (−1.25, 0.71)	1.42 (1.17, 1.73)	−0.28 (−0.94, 0.32)
12	Addition	Unlike	Simple	3	1.99 (1.67, 2.35)	0.15 (−0.92, 1.13)	1.52 (1.24, 1.81)	0.08 (−0.72, 0.83)
13	Addition	Unlike	Mixed	3	1.70 (1.39, 2.03)	0.32 (−0.73, 1.28)	1.71 (1.40, 2.03)	0.73 (−0.03, 1.49)
14	Addition	Unlike	Mixed	3	1.52 (1.23, 1.84)	0.47 (−0.54, 1.39)	1.51 (1.24, 1.86)	0.85 (0.16, 1.54)
15	Subtraction	Like	Simple	2	1.01 (0.77, 1.27)	−1.30 (−2.15, −0.44)	0.56 (0.37, 0.84)	−1.75 (−2.31, −1.22)
16	Subtraction	Unlike	Simple	2	1.91 (1.60, 2.29)	0.30 (−0.77, 1.28)	1.38 (1.15, 1.68)	0.10 (−0.54, 0.70)
17	Subtraction	Like	Mixed	2	0.85 (0.68, 1.05)	−0.47 (−1.30, 0.36)	0.41 (0.28, 0.57)	−0.78 (−1.25, −0.32)
18	Subtraction	Unlike	Mixed	2	0.96 (0.74, 1.21)	0.83 (−0.11, 1.70)	0.68 (0.51, 0.85)	0.83 (0.23, 1.38)
19	Subtraction	Unlike	Mixed	2	1.56 (1.28, 1.85)	0.51 (−0.49, 1.43)	1.22 (0.99, 1.53)	0.57 (−0.08, 1.18)
20	Subtraction	Unlike	Mixed	2	1.05 (0.77, 1.37)	1.28 (0.27, 2.21)	0.77 (0.55, 1.04)	1.53 (0.88, 2.13)

Table 4. Results of empirical study: Parameter estimates and 95% CIs

	MMIRM-MLC model 1				MMIRM-MLC model 2				MMIRM-MMC		MM-MLC			
	Pre-test		Post-test		Pre-test		Post-test		Post-test		Pre-test		Post-test	
	Est	CI	Est	CI	Est	CI	Est	CI	Est	CI	Est	CI	Est	CI
Fixed effects														
Student level														
L.Inter[γ_{01}]	–		0[a]		–		0[a]		0[a]		–		0[a]	
U.Inter[γ_{02}]	–		0[a]		–		0[a]		0[a]		–		0[a]	
L.Pre[γ_{11}]	–		**1.074**	0.682, 1.422	–		**1.195**	0.815, 1.405	**0.487**	0.368, 0.585	–		**0.305**	0.223, 0.383
U.Pre[γ_{12}]	–		**0.865**	0.724, 1.013	–		**0.859**	0.730, 1.018	**0.158**	0.141, 0.177	–		**2.149**	1.805, 2.484
L.MD[γ_{21}]	–		–		–		0.245	−0.094, 0.625	–		–		–	
U.MD[γ_{22}]	–		–		–		−0.204	−0.445, 0.041	–		–		–	
L.Gender[γ_{31}]	–		–		–		0.067	−0.240, 0.407	–		–		–	
U.Gender[γ_{32}]	–		–		–		0.015	−0.209, 0.229	–		–		–	
Teacher level														
L.Inter[δ_{01}]	–		2.662	−0.640, 4.711	–		2.372	−0.830, 4.202	−0.294	−2.837, 2.054	–		**3.217**	2.384, 3.979
U.Inter[δ_{02}]	–		0.009	−0.795, 0.912	–		−0.040	−0.982, 1.070	**−1.511**	−2.328, −0.679	–		**7.701**	6.104, 9.318
L.Pre[δ_{11}]	–		**0.947**	0.466, 1.467	–		**0.972**	0.452, 1.487	**0.727**	0.306, 1.176	–		**0.358**	0.075, 0.781
U.Pre[δ_{12}]	–		**0.677**	0.339, 1.038	–		**0.640**	0.279, 1.027	**0.189**	0.089, 0.291	–		**1.106**	0.534, 1.694
L.TRT[δ_{21}]	–		**0.890**	0.506, 1.499	–		**0.885**	0.378, 1.406	**0.862**	0.380, 1.375	–		**0.303**	0.027, 0.627
U.TRT[δ_{22}]	–		**1.039**	0.601, 1.471	–		**0.989**	0.441, 1.376	**1.060**	0.654, 1.451	–		**1.204**	0.579, 1.822
L.Yrs[δ_{31}]	–		–		–		0.061	−0.002, 0.129	–		–		–	
U.Yrs[δ_{32}]	–		–		–		−0.008	−0.054, 0.041	–		–		–	
Random effects														
Student level														
L.Mean	0[a]		0[a]		0[a]		0[a]		0[a]		0[a]		0[a]	
U.Mean	0[a]		0[a]		0[a]		0[a]		0[a]		0[a]		0[a]	
L.Var	1[a]		1[a]		1[a]		1[a]		1[a]		1[a]		**0.881**	0.767, 1.011
U.Var	1[a]		1[a]		1[a]		1[a]		1[a]		1[a]		**21.781**	18.864, 25.232
Cov	**0.793**	0.785, 0.795	**0.691**	0.684, 0.694	**0.791**	0.785, 0.796	**0.688**	0.685, 0.690	**0.593**	0.585, 0.595	**0.786**	0.785, 0.795	**0.986**	0.985, 0.995

Continued

Table 4. (*Continued*)

	MMIRM-MLC model 1				MMIRM-MLC model 2				MMIRM-MMC		MM-MLC			
	Pre-test		Post-test		Pre-test		Post-test		Post-test		Pre-test		Post-test	
	Est	CI	Est	CI	Est	CI	Est	CI	Est	CI	Est	CI	Est	CI
Teacher level														
L.Mean[μ_1]	0.417	−1.477, 2.235	0[a]		0.490	−1.372, 2.435	0[a]		0[a]		0[a]		0[a]	
U.Mean[μ_2]	**−1.284**	−2.005, −0.606	0[a]		**−1.346**	−1.977, −0.684	0[a]		0[a]		0[a]		0[a]	
L.Var	**0.374**	0.161, 0.854	**0.268**	0.096, 0.817	**0.373**	0.157, 0.882	**0.210**	0.129, 0.901	**0.271**	0.094, 0.725	**0.303**	0.117, 0.734	**0.065**	0.014, 0.190
U.Var	**0.677**	0.368, 1.321	**0.460**	0.225, 0.990	**0.660**	0.369, 1.001	**0.390**	0.291, 0.889	**0.389**	0.195, 0.801	**0.758**	0.416, 1.482	**7.643**	3.677, 16.998
Cov	**0.338**	0.102, 0.791	0.099	−0.157, 0.473	**0.335**	0.099, 0.806	0.057	−0.201, 0.392	0.112	−0.073, 0.491	**0.305**	0.072, 0.734	0.225	−0.353, 0.960

L, *like* domain; U, *unlike* domain; –, Not modelled.

Effects in bold are significant at the 5% level.

[a]Constraint.

The MMIRM-MLC and MM-MLC are specified in the online supporting information (Appendix S3). WinBUGS was used to fit the MMIRM-MLC and MM-MLC with priors and hyperpriors comparable to those used in MMIRM-MLCs.

Table 4 presents the results for the MMIRM-MLC and MM-MLC. Pre-test effects and intervention effects were standardized on their relevant scale for comparison among the three models, MMIRM-MLC model 1, MMIRM-MLC model 2 and MM-MLC. Table S2 in the online supporting information shows the standardized estimates of the three models, based on results presented in Table 4. Compared to the standardized effects of pre-test scores in MMIRM-MLC model 1, pre-test effects were underestimated in the MMIRM-MMC and MM-MLC. The effects of standardized intervention conditions were similar between the MMIRM-MMC and MMIRM-MLC model 1. However, intervention effects were underestimated in the MM-MLC.

5. Simulation study

A simulation study was designed to examine parameter recovery of the MMIRM-MLC under Bayesian estimation using WinBUGS in various multilevel designs when the population data-generating model is an MMIRM-MLC. In addition, the results of pre-test effects and intervention condition effects were compared across the MMIRM-MLC, MMIRM-MMC and MM-MLC to show the consequences of using total scores when the population data-generating model is MMIRM-MLC model 1. WinBUGS was used to fit the MMIRM-MLC and MMIRM-MMC (see the online supporting information [Appendix S3] for a description of the MMIRM-MLC and MMIRM-MMC).

5.1. Simulation design

We selected simulation conditions that may affect the results of person parameters at the cluster level, as has been found in the empirical research question (e.g., the effect of the intervention effect) in previous research (e.g., Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Preacher, Zhang, & Zyphur, 2011). The design includes the number of clusters and the number of individuals per cluster. The number of clusters was set to $K = 24$, 50, or 100. A sample of 24 and 50 clusters is common in educational experimental intervention research, as in our empirical illustration. Examples of large numbers of clusters include national or international educational assessments such as the National Assessment of Educational Progress and the Trends in International Mathematics and Science Study. Accordingly, 100 clusters were chosen. Unlike in our empirical study, balanced cluster sizes were considered to investigate the effect of cluster sizes, including $n_k = 5$, 20, or 50, as used in other multilevel studies (e.g., Preacher *et al.*, 2011). A cluster size of 5 is found in small group designs (e.g., Kenny, Mannetti, Pierro, Livi, & Kashy, 2002). Given a selected number of clusters and number of individuals per cluster, the total number of individuals results in nine different sample sizes, $J = 120$, 250, 480, 500, 1,000, 1,200, 2,000, 2,500, or 5,000. One hundred replications were simulated for each of the nine different multilevel designs.

The same number of clusters were assigned to be either a control group or a treatment group for a balanced design. As in the empirical study, a 20-item test with a between-item design was considered: Four items for domain 1 and 16 items for domain 2. The item parameter estimates and person parameter estimates, including an intervention effect that we obtained in the empirical study, were considered as the true parameters of MMIRM-

MLC model 1, as reported in Table 4. The ICC varied across items between .058 and .297 at the pre-test and between .016 and .347 at the post-test.

5.2. Result hypotheses

The MMIRM-MLC was expected to yield the least bias because the population data-generating model was the MMIRM-MLC. The effects of pre-test scores (γ_1 and $\boldsymbol{\delta}_1$) and the effect of intervention ($\boldsymbol{\delta}_2$) are expected to be different, depending on the treatment of measurement error in pre-test scores and post-test scores. In the presence of measurement error in pre-test scores as in the MMIRM-MMC, the effects of pre-test scores are expected to be biased (e.g., Lüdtke *et al.*, 2011). However, the effects of intervention conditions are not expected to be biased in the presence of measurement error in pre-test scores when there is no intervention effect at pre-test (Cho & Preacher, 2015). In the presence of measurement error in response variables as in the MM-MLC, both the effect of pre-test scores and the effects of intervention can be biased (e.g., Fox, 2004).

5.3. Analysis

The same priors specified earlier were used in the MMIRM-MLC and comparable priors and hyperpriors used in the MMIRM-MLC were used for the MMIRM-MLC and MM-MLC. Gelman and Rubin's (1992) statistic was used to evaluate convergence with three chains. One replication of each condition was used for convergence checking. No convergence problems were encountered in any replications for the MMIRM-MLC, except the sample size condition $n_k = 5$ and $K = 24$ (total sample size = 120). This non-convergence problem may be because the sample size is too small to estimate 98 parameters (80 item parameters (20 items $\times$ 4 kinds), 10 structural parameters, and 8 variance or covariance parameters). Only the converged results are reported below. There were no convergence problems in the MMIRM-MMC or the MM-MLC. A burn-in of 5,000 iterations was used for all parameters in the MMIRM-MLC, and a burn-in of 4,000 iterations was used for all parameters in the MMIRM-MMC and MM-MLC. The same burn-in was set for the other replications in each condition. An additional 6,000 iterations were obtained to estimate the posterior moments in the MMIRM-MLC, MMIRM-MMC and MM-MLC. Monte Carlo errors for all parameters were less than about 5% of the sample standard deviation in all three models.

Percentage relative bias was calculated to show the accuracy of the parameter estimates from the MMIRM-MLC, MMIRM-MMC and MM-MLC. It is given by $100 \times [(\widehat{\delta} - \delta)/\delta]$ as an example. Before calculating percentage relative bias, estimates of the three models were standardized (see Table S2 in the online supporting information for the calculation of standardized estimates of the three models based on the empirical results in Table 4), as an example.

5.4. Simulation results

For the analysis of the MMIRM-MMC, intervention effects were first tested on pre-test total scores by adding a covariate of an intervention condition to the MMC of the MMIRM-MMC. No significant intervention effects were found in any conditions in the MMIRM-MMC, based on a 95% CI test.

Table 5 presents the percentage relative bias for pre-test effects (γ_1 and $\boldsymbol{\delta}_1$) and intervention effects ($\boldsymbol{\delta}_2$) for the MMIRM-MLC, MMIRM-MMC and MM-MLC. The following

overall patterns in percentage relative bias were observed, as reported in Table 5. First, the percentage relative bias for pre-test effects was much lower for the MMIRM-MLC than for the MMIRM-MMC and MM-MLC in all conditions, and it was lower for the MMIRM-MMC than for the MM-MLC. For the pre-test effect estimate at the individual-level ($\widehat{\gamma}_1$), percentage relative bias ranged in magnitude from 0.3 to 18.6 in the MMIRM-MLC, from −85.1 to −29.9 in the MMIRM-MMC and from −116.0 to −52.7 in the MM-MLC. For the pre-test effect estimate at the cluster level ($\widehat{\boldsymbol{\delta}}_1$), the percentage relative bias ranged in magnitude from 0.3 to 15.8 in the MMIRM-MLC, from −70.8 to −19.1 in the MMIRM-MMC and from −338.4 to −44.3 in the MM-MLC.

Second, the percentage relative bias for intervention effects was similar between the MMIRM-MLC and MMIRM-MMC, except for the condition with $K = 24$ and $n_k = 5$. The percentage relative bias ranged in magnitude from −12.0 to 0.2 in the MMIRM-MLC and from −14.6 to 0.9 in the MMIRM-MMC. However, the percentage relative bias for intervention effects in the MM-MLC was much larger than in the MMIRM-MLC and MMIRM-MMC. It ranged in magnitude from −107.4 to −49.2.

Third, overall, the percentage relative bias decreased with increasing cluster size (n_k) and number of clusters (K) for pre-test effects (γ_1 and $\boldsymbol{\delta}_1$) and intervention effects ($\boldsymbol{\delta}_2$) in all three models, although there were three conditions that did not have that pattern: $n_k = 50$ in the MMIRM-MMC for $\widehat{\delta}_{11}$, $n_k = 50$ in the MM-MLC for $\widehat{\delta}_{11}$, and $n_k = 50$ in the MM-MLC for $\widehat{\delta}_{22}$.

Table 6 reports the percentage relative bias for fixed parameter estimates (for fixed parameter estimates not reported in Table 5) and population parameter estimates of random (residual) effects in the MMIRM-MLC. The degree of bias decreased as the cluster size (n_k) and number of clusters (K) increased for all parameter estimates. Unlike the item parameters and population parameters of random (residual) effects, the $\widehat{\delta}_0$ in the MMIRM-MLC tended to be underestimated when K and n_k decreased.

6. Summary and discussion

This paper has specified the model for detecting the intervention effect when MMIRMs were used for explicit measurement error modelling in the use of pre-test and post-test scores. The main application of the MMIRM-MLC presented in this paper was to detect a more diagnostic intervention effect by estimating the intervention effect for each domain. In the empirical illustration, a four-step analysis was implemented for applying the MMIRM-MLC to an instructional intervention study in a pre-test–post-test cluster randomized trial.

In the simulation study, the accuracy of parameter estimates for the MMIRM-MLC was investigated in various multilevel designs including a design similar to the empirical study. Parameter accuracy for the all parameters was acceptable (with an acceptable bias criterion set to 15%) in all conditions considered in this study for the MMIRM-MLC, except for conditions with a small cluster size ($n_k = 5$). When using the total scores as a covariate (i.e., pre-test scores) as in the MMIRM-MLC, unacceptable bias was found in pre-test effects, whereas acceptable bias was found in intervention effects, except for a condition with a small cluster size ($n_k = 5$) and number of clusters (K=24). This finding indicates that measurement error in a covariate may not be problematic in detecting intervention effects on post-test when there is no intervention effect on pre-test, which is often the case in cluster randomized trials. On the other hand, in the presence of measurement error in response variables (i.e., post-test scores), unacceptable bias can be found in both pre-test effects and intervention effects.

Table 5. Results of simulation study: Comparisons of pre-test effects and treatment effects of MMIRM-MLC, MMIRM-MMC and MM-MLC

			γ_1			δ_1			δ_2		
	K	n_k	MMIRM-MLC	MMIRM-MMC	MM-MLC	MMIRM-MLC	MMIRM-MMC	MM-MLC	MMIRM-MLC	MMIRM-MMC	MM-MLC
Domain 1	24	5	–	−85.1	−116.0	–	−70.8	−338.4	–	−22.3	−107.4
	50	5	5.9	−85.0	−91.9	−14.5	−69.6	−206.4	−12.0	−14.6	−99.5
	100	5	2.3	−83.2	−55.6	−11.2	−66.2	−88.7	−4.4	−5.9	−51.7
	24	20	3.2	−85.0	−114.0	−12.2	−68.8	−98.4	−6.1	−5.8	−102.1
	50	20	1.5	−84.1	−81.5	−7.2	−63.6	−95.7	3.6	4.2	−95.9
	100	20	0.6	−80.0	−62.6	−1.2	−61.6	−91.7	0.8	0.9	−70.3
	24	50	1.0	−80.9	−95.4	−3.9	−63.3	−92.5	2.1	2.5	−68.1
	50	50	0.4	−79.3	−87.0	−1.1	−59.9	−67.3	1.0	1.9	−58.9
	100	50	0.1	−71.3	−79.2	−0.9	−60.7	−58.2	0.9	1.8	−49.2
Domain 2	24	5	–	−63.5	−73.8	–	−46.1	−97.2	–	−17.3	−76.4
	50	5	18.6	−55.8	−70.8	−15.8	−30.5	−47.8	−10.5	−9.8	−53.0
	100	5	8.8	−51.8	−66.6	−6.2	−29.4	−45.1	−7.0	−7.5	−51.0
	24	20	11.1	−60.8	−71.4	−4.5	−31.9	−49.0	−3.1	−2.9	−66.7
	50	20	3.2	−43.6	−68.9	−2.1	−24.8	−47.9	2.4	3.8	−66.0
	100	20	0.9	−35.8	−67.5	0.9	−23.3	−47.6	0.3	2.6	−60.1
	24	50	3.0	−50.9	−67.1	−1.5	−24.9	−46.4	1.2	3.7	−54.0
	50	50	0.4	−41.8	−64.1	0.5	−22.5	−46.1	0.5	1.1	−56.4
	100	50	0.3	−29.9	−52.7	0.3	−19.1	−44.3	0.2	0.9	−54.0

Table 6. Results of simulation study: Percentage relative bias of MMIRM-MLC

			Pre-test			Post-test		
Fixed effect	K	n_k	α_1	β_1	μ	α_2	β_2	δ_0
Domain 1	24	5	–	–	–	–	–	–
	50	5	26.4	13.3	22.5	23.7	18.0	−13.2
	100	5	17.0	9.9	13.0	10.3	9.8	−5.9
	24	20	10.3	10.4	13.8	10.4	10.1	−6.7
	50	20	9.9	8.0	10.6	9.8	7.0	−3.1
	100	20	7.2	2.3	8.2	6.0	3.2	−0.5
	24	50	9.9	7.3	10.4	8.1	6.6	−2.4
	50	50	5.4	1.3	8.0	9.3	2.9	1.2
	100	50	4.2	0.2	6.1	6.4	0.7	0.2
Domain 2	24	5	–	–		–	–	–
	50	5	17.9	12.4	23.7	13.5	15.2	−12.1
	100	5	10.6	9.1	12.8	4.3	6.8	−5.0
	24	20	15.3	9.9	13.0	8.4	6.7	−5.4
	50	20	6.6	9.9	11.4	3.0	5.2	−2.9
	100	20	2.5	5.0	8.4	2.1	4.0	0.2
	24	50	5.5	7.2	11.1	2.9	4.9	−2.0
	50	50	1.5	4.2	6.1	1.9	2.8	0.5
	100	50	0.8	0.4	3.9	0.1	0.1	0.3

			Pre-test				Post-test			
Random effect	K	n_j	Corr. in Σ_3	Var1 in Σ_4	Var2 in Σ_4	Cov. in Σ_4	Corr. in Σ_5	Var1 in Σ_6	Var2 in Σ_6	Cov. in Σ_6
	24	5	–	–	–	–	–	–	–	–
	50	5	1.3	160.7	140.2	62.7	4.4	40.4	44.2	21.4
	100	5	1.3	10.8	10.4	27.4	3.1	15.3	17.7	12.9
	24	20	1.2	11.2	10.1	13.2	3.2	10.2	14.3	13.7
	50	20	1.1	−15.9	−13.1	−8.4	2.1	−16.9	14.2	10.9
	100	20	0.9	−3.4	−2.9	−4.3	1.4	−14.3	−12.3	11.9
	24	50	1.0	−14.4	−11.5	−6.4	1.0	−14.9	−10.6	9.7
	50	50	0.9	−3.2	−2.5	−3.1	1.0	−8.6	−7.9	−15.8
	100	50	0.9	−2.2	−0.9	−2.3	0.9	−7.7	−1.8	−9.0

We now discuss the limitations of the current study and future work. First, the unique application in the current study was to detect the intervention effect for each domain for its diagnostic value. Sinharay (2010) showed that subscores should meet strict standards of reliability, and that weak correlation between domain scores has added value in terms of mean square error in estimating the true subscore. Reliable subscores should be obtained to make valid inferences about scores in the subtest domains. Lack of sufficient reliability is a concern when there are a small number of items for a domain. For the current application, a moderate correlation coefficient between two domain scores was found at the student level (.691), and a small correlation coefficient was found at the teacher level (.099). However, there were four items for the *like* domain and 16 items for the *unlike* domain. Thus, the reliability of the subscore for the *like* domain can be questioned from a value-added perspective.

Second, there were two main sources of measurement bias in the empirical study: clusters and groups. We first tested cluster invariance and then tested the possibilities of measurement invariance across the groups based on the results of the cluster invariance test. It is important to note that this is not the only step for testing invariance. For example, measurement invariance across groups can be tested first, and then cluster bias can be investigated. Jak *et al.* (2013) stated that there is no universally optimal procedure in most situations, and different procedures generally identify the same items as being biased, but the power to detect bias may vary. A comparison study with alternative procedures is needed to determine the Type I error and power to detect measurement invariance in the use of the DIC.

Third, the simulation study has the same limitations as other simulation studies, that is, the conditions we considered are limited because the simulation study was mainly designed to check the accuracy of parameter estimates in various multilevel designs. The limited conditions include the true parameters and the ICC found from the empirical study, the number of items for each domain, and the balanced design. More extensive simulations that vary the limited conditions should be conducted to make solid generalizations.

Fourth, one may think that a comparison among the MMIRM-MLC, MMIRM-MMC and MM-MLC approaches is unfair when the population data-generating model is the MMIRM-MLC. However, we chose the MMIRM-MLC as the population data-generating model for two main reasons. First, our main interest in comparing the three models was to investigate the extent to which pre-test effects on total scores or intervention effects on total scores may produce misleading inferences on an error-free latent construct. In addition, we were interested in the degree to which the MMIRM-MLC would outperform the MMIRM-MMC and MM-MLC even though it may be obvious that the MMIRM-MLC would perform better than the MMIRM-MMC and MM-MLC overall in this situation. Still, there was no guarantee that the MMIRM-MLC would recover its own parameters well even when the MMIRM-MLC was the population data-generating model. Indeed, the MMIRM-MLC would not converge while the MMIRM-MMC would when the sample size was small ($K = 24$ and $n_k = 5$).

To conclude, the present study focused on the empirical illustration of the MMIRM-MLC and its evaluations in using Bayesian analysis. When measurement error is a concern in using a response variable and a covariate, the MMIRM-MLC can be an analytic tool for detecting an intervention effect in pre-test–post-test cluster randomized trials. However, given the results of the simulation study, the MMIRM-MLC can be used when both the number of clusters and the cluster size are large enough.

Acknowledgements

The first author received the following financial support for the research, authorship, and/or publication of this article: 2013 National Academy of Education/Spencer Postdoctoral Fellowship. The data used in the paper were collected with the following support: US Department of Education, Institute of Education Sciences, PR Number H324A090179. Any opinions, findings, or conclusions are those of the first author and do not necessarily reflect the views of the supporting agencies.

References

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies (with discussion). *Journal of the Royal Statistical Society, Series A*, *149*, 1–42.

Battauz, M., & Bellio, R. (2011). Structural modeling of measurement error in generalized linear models with Rasch measures as covariate. *Psychometrika*, *76*, 40–56. doi:10.1007/s11336-010-9195-z

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation of multidimensional IRT models. *Psychometrika*, *66*, 541–562. doi:10.1007/BF02296195

Bejar, I. I. (1980). Biased assessment of program impact due to psychometric artifacts. *Psychological Bulletin*, *87*, 513–524. doi:10.1037/0033-2909.87.3.513

Bottge, B. A. Ma, X., Gassaway, L., Toland, M. D., Butler, M., & Cho, S.-J. (2014). Effects of blended instructional models on math performance. *Exceptional Children*, Measurement error in nonlinear models, *80*, 423–437. doi:10.1177/0014402914527240

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models in social and behavioral research: Applications and data analysis methods*. Newbury Park, CA: Sage.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective*. Boca Raton, FL: Chapman & Hall/CRC.

Cho, S.-J., & Preacher, K. J. (2015). *Measurement error correction formula for group differences in a cluster-randomized design*. Unpublished manuscript, Psychological Sciences, Vanderbilt University.

Culpepper, S. A., & Aguinis, H. (2011). Using analysis of covariance (ANCOVA) with fallible covariates. *Psychological Methods*, *16*, 166–178. doi:10.1037/a0023355

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

de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscoring. *Applied Psychological Measurement*, *35*, 296–316. doi:10.1177/0146621610378653

Fox, J-.P. (2004). Modelling response error in school effectiveness research. *Statistica Neerlandica*, *58*, 138–160. doi:10.1046/j.0039-0402.2003.00253.x

Fox, J-.P. & Glas, G. A. W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika*, *68*, 169–191. doi:10.1007/BF02294796

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*, 72–91.

Gelman, A. & Meng, X.-L. (1996). Model checking and model improvement. In W. R. Gilks, S. R. Richardson, and D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 189–201). London, UK: Chapman & Hall.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472. doi:10.1214/ss%2F1177011136

Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory model. *Applied Psychological Measurement*, *27*, 217–233. doi:10.1177/0146621603027003003

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

Goldstein, H., Kounali, D., & Robinson, A. (2008). Modelling measurement errors and category misclassification in multilevel models. *Statistical Modelling*, *8*, 243–261. doi:10.1177/1471082X0800800302

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling*, *20*, 265–282. doi:10.1080/10705511.2013.769392

Kenny, D., Mannetti, L., Pierro, A., Livi, S., & Kashy, D. (2002). The statistical analysis of data from small groups. *Journal of Personality and Social Psychology*, *83*, 126–137. doi:10.1037/0022-3514.83.1.126

Lockwood, J. R., & McCaffrey, D. F. (2014). Correcting for test score measurement error in ANOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics*, *39*, 22–52. doi:10.3102/1076998613509405

Lüdtke, O., Marsh, H. W., Robitzsch, A, & Trautwein U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods*, *16*, 444–467. doi:10.1037/a0024376

McDonald, R. P. (1993). A general model for two-level data with responses missing at random. *Psychometrika*, *58*, 575–585. doi:10.1007/BF02294828

Meng, X.-L. (1994). Posterior predictive p-values. *Annals of Statistics*, *22*, 1142–1160. doi:10.1214/aos/1176325622

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, *22*, 376–398. doi:10.1177/0049124194022003006

Muthén, B. O., & Asparouhov, T. (2013). Item response modeling in Mplus: A multi-dimensional, multi-level, and multi-time point example. In W. J. van der Linden & R. K. Hambleton (Eds), *Handbook of item response theory, models, statistical tools, and applications*. Boca Raton, FL: Chapman & Hall/CRC Press.

Muthén, L. K. & Muthén, B. O. (1998–2014). *Mplus [Computer program]*. Los Angeles, CA: Author.

Porter, A. C., & Raudenbush, S. W. (1987). Analysis of covariance: Its model and use in psychological research. *Journal of Counseling Psychology*, *34*, 383–392. doi:10.1037/0022-0167.34.4.383

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling*, *18*, 161–182. doi:10.1080/10705511.2011.557329

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167–190. doi:10.1007/BF02295939

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*, 173–185. doi:10.1037/1082-989X.2.2.173

Raudenbush, S. W., & Sampson, R. (1999). Assessing direct and indirect effects in multilevel designs with latent variables. *Sociology Methods and Research*, *28*, 123–153. doi:10.1177/0049124199028002001

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, *12*, 1151–1172. doi:10.1214/aos/1176346785

Sinharay, S. (2005). Assessing fit of unidimensional item response models using a Bayesian approach. *Journal of Educational Measurement*, *42*, 375–394. doi:10.1111/j.1745-3984.2005.00021.x

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*, 150–174. doi:10.1111/j.1745-3984.2010.00106.x

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1996). *BUGS: Bayesian inference using Gibbs sampling, version 0.5 (version ii)*. Cambridge, UK: MRC Biostatistics Unit.

Spiegelhalter, D. J., Best, N. G., Carlin, B. R., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, *64*, 583–616. doi:10.1111/1467-9868.00353

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS user manual*. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health. Retrieved from http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–69. doi:10.1177/109442810031002

Verhagen, J., & Fox, J.-P. (2013). Longitudinal measurement in health-related surveys. A Bayesian joint growth model for multivariate ordinal responses. *Statistics in Medicine*, *32*, 2988–3006. doi:10.1002/sim.5692

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Received 25 June 2014; revised version received 3 October 2014

Supporting Information

The following supporting information may be found in the online edition of the article:

Appendix S1. Diagram of person parameters of MMIRM-MLC.
Appendix S2. Item response models for invariance tests.
Appendix S3. Model description of MMIRM-MMC and MM-MLC.
Table S1. Results of measurement invariance tests using DIC.
Table S2. Model comparisons: Standardized estimates of pre-test effects and intervention effects among MMIRM-MMC, MM-MLC, and MMIRM-MLC model 1.