Article

Applied Psychological Measurement
2015, Vol. 39(8) 627–642

DOI: 10.1177/0146621615591094
apm.sagepub.com
SAGE

Detecting Intervention Effects in a Cluster-Randomized Design Using Multilevel Structural Equation Modeling for Binary Responses

Sun-Joo Cho[1], Kristopher J. Preacher[1], and Brian A. Bottge[2]

Abstract

Multilevel modeling (MLM) is frequently used to detect group differences, such as an intervention effect in a pre-test–post-test cluster-randomized design. Group differences on the post-test scores are detected by controlling for pre-test scores as a proxy variable for unobserved factors that predict future attributes. The pre-test and post-test scores that are most often used in MLM are summed item responses (or total scores). In prior research, there have been concerns regarding measurement error in the use of total scores in using MLM. To correct for measurement error in the covariate and outcome, a theoretical justification for the use of multilevel structural equation modeling (MSEM) has been established. However, MSEM for binary responses has not been widely applied to detect intervention effects (group differences) in intervention studies. In this article, the use of MSEM for intervention studies is demonstrated and the performance of MSEM is evaluated via a simulation study. Furthermore, the consequences of using MLM instead of MSEM are shown in detecting group differences. Results of the simulation study showed that MSEM performed adequately as the number of clusters, cluster size, and intraclass correlation increased and outperformed MLM for the detection of group differences.

Keywords

binary responses, cluster-randomized design, group difference, item response model, multilevel structural equation modeling

Introduction

The cluster-randomized trial design is one of the more common designs in education (Raudenbush, 1997).[1] In the cluster-randomized design, clusters of persons rather than individual persons are assigned at random to treatments. Because many educational and psychological attributes are unobservable as outcomes for an intervention, multiple indicators (or items) are often collected to infer the unobserved attributes. There are many circumstances in which items

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

Corresponding Author:
Sun-Joo Cho, Peabody College of Vanderbilt University, 230 Appleton Place, Nashville, TN 37203-5721, USA.
Email: sj.cho@vanderbilt.edu

measuring these unobserved attributes are binary responses, such as correct–incorrect answers, true–false answers, present–absent symptoms, and endorsed–not endorsed attitudes in educational and psychological testing programs.

The objective of statistical analysis in most cluster-randomized designs is to explore group differences between a control group and a treatment group at the cluster level. Multilevel modeling (MLM) is the general approach used to detect a group difference on post-test outcomes; often, related covariates at different levels of multilevel data are controlled in the model (e.g., Goldstein, 2003). Pre-test scores are important covariates to be controlled because they are proxy variables for unobserved factors that predict future attributes (e.g., Bloom, Richburg-Hayes, & Black, 2007). It has been shown that inconsistent estimates can be found when the initial condition (e.g., pre-test scores) is not considered (e.g., Aitkin & Alfo, 2003). Binary item responses on pre- and post-test measures are frequently summed (i.e., total score) or averaged (i.e., proportion score) when using MLM.

Referring to previous research findings, measurement error in covariates (e.g., Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Shin & Raudenbush, 2010) and outcomes (Raudenbush & Sadoff, 2008) are two major concerns in using total scores in MLM. However, previous studies have not presented the effects of these concerns on detecting group differences using MLM. To mitigate problems due to measurement error, latent variable models can be used for explicit modeling of unobserved attributes and multiple binary responses. Within a structural equation modeling (SEM) framework, there are several studies demonstrating the use of latent variable models to test ANOVA-like mean differences across groups at the latent construct level. For example, structured means models (SMMs; Sörbom, 1974) and multiple-indicator multiple-cause (MIMIC; Jöreskog & Goldberger, 1975) approaches have been presented to detect latent group differences with factor analysis. A relatively novel analytic framework, multilevel SEM (MSEM), has been used to account for multilevel data in the use of SEM (McDonald, 1993; Muthén & Muthén, 1998-2014; Rabe-Hesketh, Skrondal, & Pickles, 2004). Within an item response modeling framework, MSEM for categorical variables is also known as explanatory item response modeling (De Boeck & Wilson, 2004). However, to the authors' knowledge, MSEM for binary responses has not been applied to educational intervention data collected from cluster-randomized designs.

Thus, the purpose of the current study is to illustrate the use of MSEM to correct for measurement error in both covariate (i.e., pre-test) and outcome (i.e., post-test) and to evaluate MSEM via a simulation study. As another purpose of this study, MLM (specifically, the two-level random intercept model) and MSEM are compared to show the consequences of using MLM instead of MSEM. The authors examine the degree of bias in the group difference estimate when MLM is used instead of using MSEM, given a variety of study design conditions (e.g., number of clusters, cluster size, and intraclass correlation [ICC]). A two-level design was chosen because of its prevalence in education research (Schochet, 2008) and to ensure a simpler and clearer explanation of the results than would be possible with a more complex design.

This article is organized as follows. First, MLM is described in assessing group differences. Next, MSEM is specified for multiple binary responses, which are obtained from pre-test and post-test, and MSEM and MLM are applied to data from an educational intervention study having a pre-test–post-test cluster-randomized design. Subsequently, MSEM is compared with MLM in assessing group differences when MSEM is the data-generating model and MSEM is evaluated in terms of parameter recovery in various multilevel designs.

Assessing Group Differences in MLM

For comparison with MSEM, a two-level random intercept MLM is considered to detect group differences on post-test scores, controlling for pre-test scores (Equation 4.6 in Moerbeek, Van

Breukelen, & Berger, 2008). Due to space considerations, MLM is specified in the online supplementary material.

Assessing Group Differences in MSEM

In this section, MSEM is described. To frame this data structure within the multilevel literature (e.g., Bryk & Raudenbush, 1992), item responses are at Level 1, with individuals and items cross-classified at Level 2. Individuals are nested within clusters at Level 3. In the following model, a measurement model for the pre-test is specified as a covariate and a measurement model for the post-test is specified as an outcome (e.g., Battauz, Bellio, & Gori, 2011; Rabe-Hesketh et al., 2004, Equation 18). A treatment condition variable is a Level 3 covariate because the intervention is applied equally to all individuals within a cluster in a cluster-randomized design. In the specification, it is assumed that the same construct is being measured in a control group and a treatment group to detect the group difference (i.e., measurement invariance between the two groups; Bejar, 1980). Furthermore, the authors also assume that each individual answers several items and the same items are administered to all individuals.

The measurement model in MSEM for binary responses can be called a multilevel item response model. In most multilevel item response models, it is assumed that item discrimination over Level 2 and Level 3 are the same and constant (often 1) in Rasch multilevel item response models (e.g., Kamata, 2001) and are the same in two-parameter multilevel item response models (e.g., Fox, 2010; Jeon & Rabe-Hesketh, 2012). The model formulation having the same item discriminations over levels can be considered a variance component factor model (Rabe-Hesketh et al., 2004). Unlike these previous specifications of the multilevel item response model, in this study's specification, the different item discriminations over Level 2 and Level 3 were specified in the measurement model of MSEM as a general two-level factor model. A path diagram is presented to depict an MSEM in the online supplementary material.

Item responses for pre-test (denoted by the subscript 1) are specified as $[y_{1jk1}, \ldots, y_{1jki}, \ldots, y_{1jkI}]'$ and item responses for post-test (denoted by the subscript 2) are specified as $[y_{2jk1}, \ldots, y_{2jki}, \ldots, y_{2jkI}]'$ for an individual j $(j = 1, ..., J)$, a cluster k $(k = 1, ..., K)$, and an item i $(i = 1, ..., I)$. At pre-test, dependency in item responses is explained by two latent variables, θ_{1jk} and θ_{1k}, for individual level and cluster level, respectively. The measurement model for the pre-test is as follows:

$$P\left(y_{1jki} = 1 | \theta_{1jk}, \theta_{1k}\right) = \Phi\left(a_{1i,W} \cdot \theta_{1jk} + a_{1i,B} \cdot \theta_{1k} - b_{1i}\right), \quad (1)$$

where Φ denotes the normal cumulative distribution function, y_{1jki} is an item response at pre-test, $a_{1i,W}$ is an item discrimination parameter at Level 2 for pre-test, $a_{1i,B}$ is an item discrimination parameter at Level 3 for pre-test, b_{1i} is an item location parameter for pre-test, θ_{1jk} is a latent variable at Level 2, and θ_{1k} is a latent variable at Level 3.

The measurement model for the post-test is as follows:

$$P\left(y_{2jki} = 1 | \theta_{2jk}, \theta_{2k}\right) = \Phi\left(a_{2i,W} \cdot \theta_{2jk} + a_{2i,B} \cdot \theta_{2k} - b_{2i}\right), \quad (2)$$

where $a_{2i,W}$ is an item discrimination parameter at Level 2 for post-test, $a_{2i,B}$ is an item discrimination parameter at Level 3 for post-test, b_{2i} is an item location parameter for post-test, θ_{2jk} is a latent variable at Level 2, and θ_{2k} is a latent variable at Level 3.

Covariates are introduced for the latent variables, $\boldsymbol{\theta}$. A structural model for the latent variable at Level 2 (e.g., the student level) is as follows:

$$\theta_{2jk} = \vartheta_{00} + \vartheta_{10} \cdot \theta_{1jk} + \varepsilon_{2jk}, \tag{3}$$

where ϑ_{00} is an intercept at Level 2, ϑ_{10} is the effect of the pre-test latent score at Level 2, and ε_{2jk} is the residual of post-test latent scores at Level 2.

The structural model for the latent variable at Level 3 (e.g., the teacher level) is as follows:

$$\theta_{2k} = \delta_{00} + \delta_{10} \cdot \theta_{1k} + \delta_{20} \cdot TRT_k + \varepsilon_{2k}, \tag{4}$$

where TRT_k is a covariate for treatment condition with a value of 0 for members of the control group and a value of 1 for members of the treatment group, δ_{00} is the intercept at Level 3, δ_{10} is the effect of the pre-test latent score at Level 3, δ_{20} is the effect of TRT_k at Level 3, and ε_{2k} is the residual of post-test latent score at Level 3.

In the structural models (Equations 3 and 4), pre-test covariates (θ_{1jk} and θ_{1k}) are latent covariates. As explained earlier, a pre-test covariate is a proxy variable for unobserved factors that is used to predict post-test scores. Adding the two structural models (Equations 3 and 4) to the measurement model for the post-test (Equation 2), the combined model leads to the following:

$$\begin{aligned} P\left(y_{2jki} = 1 \middle| \varepsilon_{2jk}, \varepsilon_{2k}, \theta_{1jk}, \theta_{1k}\right) = \Phi[a_{2i,W} \cdot \left(\vartheta_{00} + \vartheta_{10} \cdot \theta_{1jk} + \varepsilon_{2jk}\right) \\ + a_{2i,B} \cdot (\delta_{00} + \delta_{10} \cdot \theta_{1k} + \delta_{20} \cdot TRT_k + \varepsilon_{2k}) - b_{2i}]. \end{aligned} \tag{5}$$

To identify the model, the following constraints are set as indicated by asterisks: $\vartheta_{00} = 0^*$, $\theta_{1jk} \sim N(0^*, 1^*)$, $\theta_{1k} \sim N(\mu, 1^*)$, $\varepsilon_{2jk} \sim N(0^*, 1^*)$, and $\varepsilon_{2k} \sim N(0^*, 1^*)$. Alternatively, the item discrimination for one of the items (e.g., the first item) can be set to 1 instead of setting variances to 1: $a_{21,W} = 1$, $a_{21,B} = 1$, $a_{11,W} = 1$, and $a_{11,B} = 1$, to identify the scale unit of parameters.

Parameter Estimation

The authors chose Bayesian estimation to deal with the high-dimensional structure inherent in MSEMs for binary responses. Maximum likelihood estimation of models for binary data is challenging because the marginal likelihood does not have a closed form so that maximum likelihood estimation requires numerical or Monte Carlo integration. In Bayesian analysis, it is possible to sample complex and high-dimensional posterior densities by Markov chain Monte Carlo (MCMC) methods through sampling from the conditional distributions of parameters without numerical integration. Mplus version 7.11 (Muthén & Muthén, 1998-2014) was chosen to model the MSEMs. The algorithm used for MCMC in Mplus is GIBBS(PX1). In addition to the constraints needed to identify MSEMs, residual variances in the Mplus item response theory (IRT) model specification (see Asparouhov & Muthén, 2013, for IRT specification in Mplus) were set to 1 to identify the model and also to provide parameters specified in Equation 5.

For the specified MSEM, the joint posterior distribution for the parameters, $\Theta = \{a_{1i,W}, a_{1i,B}, a_{2i,W}, a_{2i,B}, b_{1i}, b_{2i}, \vartheta_{10}, \delta_{00}, \delta_{10}, \delta_{20}, \mu, \theta_{1jk}, \theta_{1k}, \varepsilon_{2jk}, \varepsilon_{2k}\}$, can be rewritten as

$$\begin{aligned} P\left(\Theta | y_{1jki}, y_{2jki}\right) \propto P\left(y_{1jki} | \Theta\right) \cdot P\left(y_{2jki} | \Theta\right) \cdot \{P(a_{1i,W}) \cdot P(a_{1i,B}) \cdot P(a_{2i,W}) \cdot P(a_{2i,B}) \cdot P(b_{1i}) \cdot P(b_{2i})\} \cdot \\ \{P(\vartheta_{10}) \cdot P(\delta_{00}) \cdot P(\delta_{10}) \cdot P(\delta_{20})\} \cdot \left\{P\left(\theta_{1jk}\right) \cdot P(\theta_{1k} | \mu) \cdot P(\mu) \cdot P\left(\varepsilon_{2jk}\right) \cdot P(\varepsilon_{2k})\right\}, \end{aligned}$$

where $P(y_{1jki}|\Theta)$ and $P(y_{2jki}|\Theta)$ are likelihood functions, probabilities within the first braces indicate prior distributions of item parameters, probabilities within the second braces indicate prior distributions of regression coefficients, and probabilities within the third braces indicate

prior and hyper-prior distributions of latent variables or residuals. In the current study, item discrimination parameters were given a mildly informative prior, $N(0, 1)$, as used in many IRT applications to have stable item discrimination parameter estimates (e.g., Béguin & Glas, 2001; Patz & Junker, 1999) and item location parameter and regression coefficients were given noninformative priors, $N(0, 5)$. The prior distributions of latent variables or residuals were set as constraints (noted by asterisks) as follows: $\theta_{1jk} \sim N(0^*, 1^*)$, $\theta_{1k} \sim N(\mu, 1^*)$, $\varepsilon_{2jk} \sim N(0^*, 1^*)$, $\varepsilon_{2k} \sim N(0^*, 1^*)$. The hyper-prior distribution on the mean was set as $\mu \sim N(0, 5)$. The posterior median and standard deviation of sample values were reported as posterior moments. To ensure that stable parameter estimates are obtained, Gelman and Rubin's (1992) method was implemented in Mplus.

Characterizing Multilevel Structure: ICC

With the measurement models (Equations 1 and 2), an ICC can be specified for each item to indicate the proportion of variance that is attributable to clusters. ICC is the correlation coefficient (Corr) among probabilities of item responses on the probit scale[2] ($P'(y_{jki})$) for the same cluster k, but different persons j and j', and can be defined as follows[3]:

$$\mathrm{ICC}_i = \mathrm{Corr}\left(P'\left(y_{jki}\right), P'\left(y_{j'ki}\right)\right) = \frac{\mathrm{Cov}\left(P'\left(y_{jki}\right), P'\left(y_{j'ki}\right)\right)}{\sqrt{\mathrm{Var}\left(P'\left(y_{jki}\right)\right)} \cdot \sqrt{\mathrm{Var}\left(P'\left(y_{j'ki}\right)\right)}} \tag{6}$$

$$= \frac{a_{i,B}^2}{\sqrt{a_{i,W}^2 + a_{i,B}^2} \cdot \sqrt{a_{i,W}^2 + a_{i,B}^2}}. \tag{7}$$

Characterizing Measurement Error in Scores: Reliability

In this section, reliability is presented using measurement models for pre-test and post-test results. One of the advantages of using MSEM is the ability to cope with measurement error for each individual and each cluster. In Bayesian analysis, standard errors, $SE_{\theta_{jk}}$ and SE_{θ_k}, are produced automatically from the MCMC simulations as standard deviations of sampled values.

It is also possible to present a test-level reliability, within-reliability, and between-reliability in MSEM. Within-reliability represents the ratio of within-cluster true score variance to total within-cluster variance, whereas between-reliability represents the ratio of the between-cluster true score variance to total between-cluster variance (Geldhof, Preacher, & Zyphur, 2014; Muthén, 1991). Applying these definitions of reliability for the specified MSEM leads to a within-reliability (ρ_W) estimate as follows[4]:

$$\rho_W = \frac{\mathrm{Var}\left(\theta_{jk}\right)}{\mathrm{Var}\left(\widehat{\theta}_{jk}\right)} = \frac{\mathrm{Var}\left(\widehat{\theta}_{jk}\right) - \left(\frac{1}{J}\right) \sum_{j=1}^{J} \left(SE_{\theta_{jk}}\right)^2}{\left\{\frac{1}{J-1}\right\} \cdot \sum_{j=1}^{J} \left(\widehat{\theta}_{jk} - \overline{\widehat{\theta}}_1\right)^2}, \tag{8}$$

where $\widehat{\theta}_{jk}$ is the estimate of θ_{jk} and $\overline{\widehat{\theta}}_1$ is the mean of $\widehat{\theta}_{jk}$ across individuals. Between-reliability (ρ_B) can be defined as follows:

$$\rho_B = \frac{\mathrm{Var}(\theta_k)}{\mathrm{Var}\left(\widehat{\theta}_k\right)} = \frac{\mathrm{Var}\left(\widehat{\theta}_k\right) - \left(\frac{1}{K}\right)\sum_{k=1}^{K}(SE_{\theta_k})^2}{\left\{\frac{1}{K-1}\right\}\cdot\sum_{k=1}^{K}\left(\widehat{\theta}_k - \overline{\widehat{\theta}}_2\right)^2}, \quad (9)$$

where $\widehat{\theta}_k$ is the estimate of θ_k and $\overline{\widehat{\theta}}_2$ is the mean of $\widehat{\theta}_k$ across clusters.

Empirical Study

The data used in this study were collected in an efficacy trial of Enhanced Anchored Instruction (EAI). EAI aims to raise math achievement in middle and high school students. Next, a brief description of EAI is given; interested readers may find additional details in Bottge, Ma, Gassaway, Butler, and Toland (2014). The design of the efficacy trial was a pre-test–post-test cluster-randomized design. 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 students' class assignment. The research question was as follows:

Do group (EAI vs. BAU) differences emerge for computation?

Teacher and Student Samples

Twenty-four urban and rural middle schools in the Southeast United States participated in the study. Half were randomly assigned to EAI and BAU. Each school had one participating inclusive math classroom, although one school had two participating classrooms. Of the initial sample, 25 students did not respond to all items in the pre-test or post-test. As a result, 232 BAU and 214 EAI students remained in the final sample. The cluster size ranged from 7 to 28 students, with an average of 17.84. Students were comparable across instructional conditions in terms of gender, ethnicity, subsidized lunch, and disability area, based on chi-square tests of equal proportions. In the study, one inclusive math class from each school was sampled, with the exception of one school that had two inclusive math classes. Therefore, clustering due to schools was ignored and a two-level structure (i.e., 446 students nested in 25 teachers) was considered.

Measures: Fraction Computation Test

The researcher-developed *Fraction Computation Test*, administered at pre-test and post-test, was used in the current study to illustrate MSEM. The test comprised 20 items assessing students' ability to manually add and subtract fractions. Item features were described in the online supplementary material. There were no missing item responses in the final sample of 446 students for analyses.

Analysis and Results

The authors chose a Bayesian analysis using Mplus 7.11 to fit the MLM specified for comparability of results with MSEM. Mplus code to fit MLM is available in the online supplementary material. The same precision in prior distributions as used in MLM was used for fixed effects, $N(0, 5)$. An inverse-gamma distribution was used for variances of random effects in MLM, $IG(-1, 0)$ in the Mplus specification. In MLM, the group difference effect between BAU and EAI groups is coded with a value of 0 for members of the BAU group and a value of 1 for

members of the EAI group. Based on a Gelman-Rubin statistic with 2 chains, a conservative burn-in of 1,000 iterations was used followed by 8,000 post-burn-in iterations. The ICC obtained from the unconditional MLM was .23 for pre-test and .24 for post-test, indicating 23% and 24% of the total variance of total scores was explained at the teacher level for pre-test and post-test, respectively. Analysis steps and results of MSEM are described below.

Step 1. Determining the number of dimensions at the individual level. To determine the number of dimensions at the individual level for MSEM, the authors first compared a set of exploratory factor analyses (polychoric correlations with limited information robust weighted least squares estimation with Oblimin rotation) at each time point. Mplus version 7.11 was used for exploratory factor analysis. The 1-factor model at the individual level provides a good fit to the data according to comparative fit index (CFI) and Tucker-Lewis index (TLI) ($>$.994 at both pre-test and post-test) and root mean square error of approximation (RMSEA; 0.052, CI = [0.050, 0.052] at pre-test; 0.055, CI = [0.054, 0.056] at post-test). Based on this result, a latent variable was modeled at the individual level in MSEM in the subsequent analysis.

Step 2. Fitting a multilevel measurement model. The multilevel measurement model of MSEM (Equations 1 and 2) was fit to check if the multilevel measurement model is required instead of the single-level measurement model. ICCs for items obtained from the multilevel measurement model of MSEM ranged from .058 to .297 for pre-test and ranged from .016 to .318 for post-test, indicating a nonignorable dependency due to clusters (i.e., teachers).

Step 3. Testing measurement invariance. In multiple measurement (or longitudinal) multilevel data arising from multiple groups, measurement invariance assumptions can be tested across time points (i.e., pre-test and post-test) and across groups (i.e., BAU vs. EAI). Measurement invariance across groups is necessary for comparing group means (Bejar, 1980). However, 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). Thus, measurement invariance for BAU versus EAI was tested at each time point.

Using the multigroup multilevel measurement model (i.e., multigroup extension of Equations 1 and 2), three invariance models were compared to investigate measurement invariance across groups at pre-test and post-test (e.g., Widaman & Reise, 1997): (a) a configural invariance model, in which all item parameters are estimated simultaneously in each group under the same factor structures; (b) a weak invariance model, in which only discrimination parameters are constrained to be equal across groups; and (c) a strong invariance model, in which all item parameters are constrained to be equal across groups.

Competing models to test measurement invariance assumptions were compared using a relative fit criterion, deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002) (Verhagen & Fox, 2013). A smaller DIC represents a better fit of the model, and a difference of less than 5 or 10 units between models does not provide sufficient evidence for favoring one model over another (Spiegelhalter, Thomas, Best, & Lunn, 2003). MSEM was fit at each time point using Mplus 7.11 to calculate DIC. Based on DIC, the weak invariance model was chosen at pre-test (DIC = 8,226 for the configural invariance model, DIC = 8,168 for the weak invariance model, and DIC = 8,212 for the strong invariance model) and at post-test (DIC = 8,276 for the configural invariance model, DIC = 8,148 for the weak invariance model, and DIC = 8,280 for the strong invariance model).

In the presence of weak invariance, the relative ordering of students' scores can be different because of the different difficulty levels between the groups. Whether BAU versus EAI can be scored and compared on the same scale under violation of weak invariance was checked with

the correlation between scores ($\widehat{\theta}_{1jk} + \widehat{\theta}_{1.k}$ for pre-test and $\widehat{\theta}_{2jk} + \widehat{\theta}_{2.k}$ for post-test) using a calibration by BAU versus EAI and a calibration with all students. The correlation coefficients were high: .975 at pre-test and .950 at post-test. This indicates that the relative ordering of students' scores does not change in the presence of weak invariance for BAU versus EAI. In addition, the group mean difference between BAU and EAI was compared between the weak invariance model and the strong invariance model in MSEM, by fitting multigroup MSEM[5] with the weak invariance model and MSEM with the strong invariance model, separately. The weak invariance model does not distort the scale score much at the group level at pre-test or post-test.[6] Based on both individual score and group mean comparisons, strong measurement invariance was assumed for BAU versus EAI in Step 4.

Step 4. Fitting MSEM. Mplus code to fit MSEMs is available in the online supplementary material.[7] The group difference effect between BAU and EAI groups is coded with a value of 0 for members of the BAU group and a value of 1 for members of the EAI group. Based on a Gelman-Rubin statistic with 2 chains, a conservative burn-in of 7,000 iterations was used, followed by 8,000 post-burn-in iterations. Thinning was set at 40, meaning that 32,000 iterations were required after burn-in to obtain the 8,000 iterations.

Table 1 presents person parameter estimates and 95% credibility intervals (CrIs) for MLM and MSEM. The pre-test score effect at the cluster level is smaller than the pre-test score effect at the individual level in both models. The same pattern was found in MSEM. However, the effect of the pre-test latent score at the teacher level was not significant in MSEM. The group difference, or EAI intervention effect, was statistically significant in both models. From the MLM results, the standardized effect size associated with the intervention effect 3.970 was about 0.571, which indicates that the EAI group performed 0.571 units higher than the BAU group on a standardized total score scale. However, in MSEM results, the EAI group performed 1.139 units higher than the BAU group on a latent variable scale. Item parameter estimates are shown in the online supplementary material. At pre-test and post-test, items vary in terms of item discriminations and difficulties, and items having a *like* denominator were less discriminating and less difficult at both within and between levels than items having an *unlike* denominator. The within-level reliability (ρ_W) obtained from the results of MSEM (Equation 8) was .693 for pre-test and .805 for post-test. The between-level reliability (ρ_B) obtained from the results of MSEM (Equation 9) was .560 for pre-test and .767 for post-test. These results indicate that there is nonignorable within- and between-level measurement error.

Demographic variables for students and teachers to MSEM were also added to see if the effect of the intervention was changed with the inclusion of those variables. In fact, the change in significance and magnitude of the intervention effect was small when other demographic information for the students and teachers was added to MSEM.

Simulation Study

A simulation study was designed to investigate the performance of MSEM for binary responses in various multilevel designs including a condition similar to the empirical study and to compare the performance of MSEM with MLM in detecting group differences. The population data-generating model was MSEM as specified in Equation 5 and data sets were generated using R (R Core Team, 2014). Simulation conditions that may affect the group difference result were selected from previous research (e.g., Lüdtke et al., 2011; Preacher, Zhang, & Zyphur, 2011). These include the number of clusters (K = 24, 50, or 100), the number of individuals per cluster (n_k = 5, 20, or 50), and the ICC at post-test (ICC = .05, .10, or .20). These three factors were

Table 1. Parameter Estimates and 95% CrI From MLM and MSEM.

	MLM						MSEM					
	Pre-test			Post-test			Pre-test			Post-test		
	Est	*SD*	CrI	Est	*SD*	CrI	Est	*SD*	CrI	Est	*SD*	CrI
Fixed effect												
Intercept	—			**6.667** (1.902)	1.049	[4.402, 8.692]	—			0.091	0.518	[0.924, 1.118]
Student level												
Pre-test	—			**0.619** (0.567)	0.043	[0.529, 0.698]	—			**0.829**	0.071	[0.698, 0.975]
Teacher level												
Pre-test	—			**0.585** (0.512)	0.148	[0.263, 0.837]	—			0.363	0.234	[0.075, 0.844]
TRT	—			**3.970** (0.571)	1.043	[1.834, 5.893]	—			**1.139**	0.445	[0.275, 2.014]
Random effect												
Student level												
Variance	—			**24.567**	1.740	[21.421, 28.250]	1[a]			1[a]		
Teacher level												
Mean	—			—			**−1.885**	0.344	[−2.600, 1.246]	0[a]		
Variance	—			**4.905**	2.380	[2.555, 11.854]	1[a]			1[a]		

Note. Values in parentheses present standardized values; STDYX Standardization in Mplus. Bold effects are significant at the 5% level. MLM = multilevel modeling; MSEM = multilevel structural equation modeling; EST = estimates (posterior median); *SD* = standard deviation; CrI = credibility interval; "—" = not modeled; TRT = treatment condition.
[a]Constraint.

fully crossed, yielding 27 (=3×3×3) conditions. One thousand replications were simulated for each of the 27 conditions. Each generated data set was analyzed using both MSEM and MLM.

Simulation Conditions

Number of clusters. The numbers of clusters were set to $K =$ 24, 50, or 100. A sample of 24 and 50 clusters is common in educational experimental intervention research (e.g., Bottge et al., 2014). Examples of large numbers of clusters include national or international educational assessments such as the National Assessment of Educational Progress (NAEP) and the Trends in International Mathematics and Science Study (TIMSS). Accordingly, 100 clusters were chosen.

Cluster size. Balanced cluster sizes were selected as n_k = 5, 20, or 50 as used in MSEM 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 cluster size, the total numbers of individuals result in 9 different sample sizes: $J =$ 120, 250, 480, 500, 1,000, 1,200, 2,000, 2,500, or 5,000.

ICC at post-test. The ICC at post-test was set at ICC = .05, .10, or .20, given the fixed pre-test ICC of .3. ICC values are rarely greater than .30 in educational and organizational studies (e.g., Fox, 2010). As considered in Preacher et al. (2011), values of .05, .10, and .20 are considered small, medium, and large, respectively.

As shown in Equation 7, each item can have a different degree of ICC in MSEM. To control for the effect of ICC in these simulation studies, however, the same item discriminations were set across items and scale constraints were used with respect to latent variables. Because post-test scores are explained by both a pre-test and an intervention effect at the cluster level, smaller ICCs for post-test than for pre-test were considered and a smaller $a_{2i,W}$ for post-test than $a_{1i,W}$ for pre-test was set for all items ($i = 1, \ldots, I$). Given fixed $a_{2i,W} = .8$ and $a_{1i,W} = 1$, other population item discriminations were derived using Equation 7 for all items ($i = 1, \ldots, I$). For the level of ICC = .05, the following item discriminations were set: $a_{1i,W} = 1.000, a_{1i,B} = .655, a_{2i,W} = .800,$ and $a_{2i,B} = .184$. For the level of ICC = .10, the following item discriminations were set: $a_{1i,W} = 1.000, a_{1i,B} = .655, a_{2i,W} = .800,$ and $a_{2i,B} = .267$. For the level of ICC = .20, the following item discriminations were set: $a_{1i,W} = 1.000, a_{1i,B} = .655, a_{2i,W} = .800,$ and $a_{2i,B} = .400$.

The mean and regression coefficients are specified as follows: $\mu = 0$ (overall mean of pre-test latent variable), $\vartheta_{01} = 0.4$ (effect of pre-test at the individual level), $\delta_{00} = -1$ (mean of post-test latent variable), $\delta_{01} = 0.7$ (effect of pre-test at the cluster level), and $\delta_{02} = 1.0$ (intervention effect with dummy coding). Item difficulty was generated with a standard normal distribution, $b_{1i} \sim N(0, 1)$ and $b_{2i} \sim N(0, 1)$. Random variables (θ_{2jk} and θ_{2k}) and residuals (ε_{2jk} and ε_{2k}) were generated from a normal distribution with a unit variance to match with model identification constraints. Twenty items were set as used in this empirical study. The same number of clusters were assigned to control and treatment groups for a balanced design.

Hypotheses for Group Differences

Three outcome measures were considered: relative percentage bias, root mean square error (RMSE), and the observed coverage of the 95% CrI.

Little bias and variability for MSEM were expected because the population data-generating model was MSEM. For MLM, bias will be inversely related to ICC (e.g., Preacher et al., 2011).

The relative percentage bias is given by $100\times[(\widehat{\delta}_{20}-\delta_{20})/\delta_{20}]$ and implies the accuracy of parameter estimates. The RMSE was computed using $\sqrt{\sum_{r=1}^{R}(\delta_{20}-\delta_{20})^2/R}$, where r indicates the rth replication from a converged solution ($r=1, \ldots, R$). When a parameter estimate is unbiased, the RMSE quantifies the precision. For biased parameter estimates, the RMSE combines the parameter bias and precision into an overall measure of accuracy.

It is expected that the number of clusters and ICC will influence coverage for MLM. Given n_k, increasing the number of clusters increases the total sample size. As the total sample size increases, the precision is increased and the CrI width is reduced. This results in reducing coverage for biased estimates. Because larger ICCs correspond to greater between-reliability, coverage should improve with increasing ICC. MSEM is expected to yield the least bias, which results in CrIs closely centered around the population parameter when the standard error of the estimate is precise. The precision of the standard error of the group effect was analyzed by determining the observed coverage of the 95% CrI. Coverage was determined by noting the proportion of trials in which the 95% CrI for the group difference effect included the population value of the group difference. Coverage close to .95 indicates that the precision is well estimated.

Results

No convergence problems were encountered in any replications for MLM and MSEM, except the sample size condition, $n_k=5$ and $K=24$ (total sample size = 120) in MSEM. This nonconvergence problem may be due to the fact that the sample size is too small to estimate 125 parameters.[8] In the next section, results from only the converged solutions are reported.

Group difference effect in MLM and MSEM. The group difference effects are shown in Table 2.

Relative percentage bias. The following overall patterns in relative percentage bias were observed as reported in Table 2. First, the relative percentage bias was much lower for MSEM than for MLM in all conditions. Relative percentage bias ranged in magnitude from −204.400 to −5.440 in MLM, whereas it ranged in magnitude from −16.780 to 0.050 in MSEM. On setting an acceptable bias criterion of 10%, unacceptable bias was found in all conditions in MLM, whereas unacceptable bias was found only in the condition with $n_k=5$ and $K=50$ in MSEM. Second, the degree of bias decreased with increasing cluster size (n_k), number of clusters (K), and ICC in both MLM and MSEM. Third, relative percentage bias is negative in all conditions for MLM, indicating that the group difference effect was underestimated. However, a negative relative percentage bias was mainly found in MSEM with a small cluster size ($n_k=5$ or 20), a small number of clusters ($K=24$ or 50), and a small ICC (ICC = .05 or .10).

RMSE. Given biased parameter estimates, the RMSE presents the combined effect of parameter bias and parameter variance. RMSE has no accepted cutoff value for deciding whether an estimate is acceptable or not, but it represents a trade-off between bias and variability. In all conditions, MSEM outperformed MLM. RMSE ranged in magnitude from 0.132 to 1.025 in MLM, whereas it ranged in magnitude from 0.027 to 0.243 in MSEM. As was found with relative percentage bias, RMSE decreased with increasing cluster size (n_k), number of clusters (K), and ICC in both MLM and MSEM.

Coverage. As predicted, MLM coverages were less than .95 when ICC is small and both K and n_k are large. For large ICC, small k, and small n_k, coverage was 1.00 in MLM. MSEM coverages were 1.00 in all conditions except one condition ($n_k=5$, $K=50$, and ICC = .05), which indicates that standard errors can be overestimated in MSEM.

Table 2. Comparisons of the Group Difference Effect in MLM (γ_{20}) and MSEM (δ_{20}).

		MLM			MSEM		
K	n_k	ICC = .05	ICC = .10	ICC = .20	ICC = .05	ICC = .10	ICC = .20
Relative percentage bias							
24	5	−204.400	−132.400	−129.800	—	—	—
50	5	−129.800	−124.400	−110.000	−16.780	−13.240	−11.000
100	5	−124.400	−116.800	−100.400	−7.440	−5.900	−5.340
24	20	−146.000	−143.000	−140.600	−7.540	−6.720	−5.400
50	20	−95.600	−85.200	−77.800	−4.080	−3.280	−3.060
100	20	−60.100	−44.600	−31.700	−1.300	−0.500	0.200
24	50	−71.200	−66.800	−69.500	−2.820	−2.760	−2.040
50	50	−69.300	−31.800	−8.600	1.430	1.220	0.500
100	50	−20.800	−13.540	−5.440	0.930	0.190	0.050
RMSE							
24	5	1.025	0.834	0.739	—	—	—
50	5	0.665	0.625	0.557	0.243	0.200	0.165
100	5	0.655	0.449	0.520	0.142	0.139	0.098
24	20	0.717	0.655	0.608	0.151	0.143	0.139
50	20	0.712	0.617	0.560	0.148	0.099	0.054
100	20	0.454	0.347	0.332	0.127	0.041	0.019
24	50	0.699	0.543	0.449	0.144	0.127	0.092
50	50	0.334	0.302	0.253	0.121	0.068	0.044
100	50	0.255	0.214	0.132	0.093	0.034	0.027
Coverage							
24	5	1.000	1.000	1.000	—	—	—
50	5	1.000	1.000	1.000	0.960	1.000	1.000
100	5	0.980	1.000	1.000	1.000	1.000	1.000
24	20	1.000	1.000	1.000	1.000	1.000	1.000
50	20	1.000	1.000	1.000	1.000	1.000	1.000
100	20	0.030	0.920	1.000	1.000	1.000	1.000
24	50	0.990	1.000	1.000	1.000	1.000	1.000
50	50	0.000	0.050	0.060	1.000	1.000	1.000
100	50	0.000	0.010	0.210	1.000	1.000	1.000

Note. MLM = multilevel modeling; MSEM = multilevel structural equation modeling; ICC = intraclass correlation; RMSE = root mean square error.

Other parameters in MSEM. The overall measure of accuracy (RMSE) of MSEM parameter estimates (except the group difference estimate) is reported in the online supplementary material. For all item parameters (except $a_{2i,B}$ and δ_{00}), the behavior of RMSE is similar across different ICC conditions, given the same cluster size and number of clusters (with the exception of condition $n_k = 50$, $K = 100$, and ICC = .05), and RMSE decreased as cluster size increased holding number of clusters (K) constant. The item discrimination parameter, $a_{2i,B}$, differs depending on ICC because the degree of ICC in outcomes was manipulated by increasing the magnitudes of $a_{2i,B}$. Thus, patterns of results were interpreted within the same ICC. Given the same ICC, RMSE decreased by increasing cluster size (n_k) and number of clusters (K). For δ_{01} and μ, RMSE decreased with increasing cluster size (n_k), number of clusters (K), and ICC.

Discussion

In summary, the first goal of this study was to illustrate the use of MSEM for binary responses to correct for the measurement error in the covariate and outcome using educational

intervention data and to investigate the behavior of MSEM in various multilevel designs. The authors also sought to describe how model parameters can be estimated using Bayesian analysis in Mplus 7.11 and present calculations of reliability coefficients and ICC for each item. MSEM performed well in terms of bias and RMSE in estimating the group difference in all conditions except conditions with a small cluster size (n_k= 5 or 20), a small number of clusters (K= 24 or 50), and a small ICC (ICC = .05 or .10). In addition, overall accuracy was not problematic[9] for other parameter estimates in MSEM. Overall parameter accuracy for all parameters, including the group difference, was acceptable for the simulation condition (n_k = 20, K = 24, ICC = .20), similar to the empirical study (average n = 17.84, k = 24, average ICC across items at post-test = .202).

The second goal was to answer questions about the consequences of using MLM instead of MSEM. MSEM was chosen as a population data-generating model. Unacceptable bias and lower CI coverage was found in most conditions that were considered in using MLM. However, MSEM performed well in terms of parameter accuracy regarding bias and RMSE in estimating the group difference in all conditions except conditions with a small cluster size (n_k= 5 or 20), a small number of clusters (K= 24 or 50), and a small ICC (ICC = .05 or .10). Given this study's simulation conditions, standard errors of the group difference estimate were overestimated in MSEM.

There are methodological limitations to the present study. First, the authors described MSEM for binary item responses, which are common in educational and psychological research. The description of MSEM, its parameter estimation description, and other quantities (reliability coefficients and ICC) for binary variables in this study cannot be directly applied to MSEM for outcomes with more than two categories (i.e., polytomous item responses). More parameters for item locations (or thresholds) must be modeled for MSEM with polytomous item responses.

Second, one may think that a comparison between MLM and MSEM approaches is unfair when the population data-generating model is MSEM. The authors chose MSEM as the population data-generating model for two main reasons. First, the main interest for the comparison was to investigate the extent to which group difference detection on total scores in MLM may produce misleading inferences about the group difference on an *error-free* latent construct. In addition, the authors were interested in the degree to which MSEM would outperform MLM even though it may be expected that MSEM would perform better than MLM overall in this situation. However, there is no guarantee MSEM will recover its own parameters well even when MSEM is the population data-generating model. Indeed, MSEM would not converge whereas MLM would when the sample size is small (K = 24 and n_k = 5).

Third, the present research as a simulation study shares the same limitations that are present in other simulation studies, that is, the conditions employed in the study design are limited, including the same 20-item test over time points, a balanced design, 1.0 group difference, measurement invariance between a control group and a treatment group, and the same item discriminations within a level to control for the effect of ICC. More extensive simulations that vary these limited conditions should be conducted to make solid generalizations. Furthermore, a future study should investigate sample size design and power issues with various magnitudes of the group difference.

Fourth, it was found that the standard error of the group difference estimate can be overestimated, based on the coverage observed in most MSEM conditions that were considered. The ratio of the mean posterior standard deviation (i.e., Bayesian standard error) to the standard deviation across replications ranged from 1.08 to 1.30 across the simulation conditions in MSEM, which implies that the standard errors were not overestimated dramatically. However, more systematic investigation is required to evaluate Bayesian standard errors, using different degrees of precision on the prior distribution and Monte Carlo error.

MLM is frequently used to estimate group differences on the scale of total scores as evidence of an intervention effect. The main goal of this study was to present and evaluate MSEM as an alternative model for detecting group differences when measurement error exists in the covariate and outcome. When clusters are the unit of analysis as is typically the case in data sets from cluster-randomized designs, evaluations of treatment can be expensive. Researchers should be aware that MSEM can perform adequately only when group size, the number of clusters, and ICC are large enough.

Authors' Note

Any opinions, findings, or conclusions are those of the first author and do not necessarily reflect the views of the supporting agencies.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: 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 article were collected with the following support: the U.S. Department of Education, Institute of Education Sciences, PR Number H324A090179.

Notes

1. According to the authors' survey from the Institute of Educational Sciences (IES) website, two branches of the IES, the National Center for Education Research (NCER) and the National Center for Special Education Research (NCSER), funded 209 Goal 3 (efficacy and replication goal) projects between 2004 and 2013. A common research design in IES Goal 3 projects is a cluster-randomized design where clusters such as teachers or schools are the unit of analysis. As a data analytic strategy for multilevel data, 59% of the IES Goal 3 projects between 2004 and 2013 used a multilevel modeling (MLM) framework based on the sum scores of the outcome measures.
2. $\text{probit}[P(y_{jki}=1|\theta_{jk},\theta_k)]=P'(y_{jki}=1)=a_{i,W}\cdot\theta_{jk}+a_{i,B}\cdot\theta_k-b_i$.
3. Subscripts 1 and 2 in Equation 7 were dropped because Equation 7 can be applied for pre-test and post-test.
4. Subscripts 1 and 2 in θ_{jk} and θ_k were dropped because ρ_W and ρ_B can be applied for pre-test and post-test.
5. Multigroup models are not available with the BAYES estimator in Mplus 7.11. The KNOWNCLASS option for TYPE = MIXTURE can be used to fit multigroup models in Mplus 7.11. However, the Bayes estimator is not allowed with TYPE = TWOLEVEL MIXTURE. Thus, WinBUGS was used to implement multigroup multilevel structural equation modeling (MSEM).
6. At pre-test, the mean score of the Enhanced Anchored Instruction (EAI) group (coded as 1) was −0.120 (credibility interval, CrI = [−0.890, 0.559]) points higher than the mean of the business as usual (BAU) group (coded as 0) on the standardized cluster-level latent trait continuum in the weak invariance model, whereas the mean score of the EAI group was −0.160 (CrI = [−0.980, 0.673]) points higher than the mean of the BAU group on the scale in the strong measurement model. At post-test, the mean score of the EAI group (coded as 1) was 1.100 (CrI = [0.255, 1.994]) points higher than the mean of the BAU group (coded as 0) on the standardized cluster-level latent trait continuum in the weak invariance model, whereas the mean score of the EAI group was 1.139 (CrI = [0.275, 2.014]) points higher than the mean of the BAU group on the scale in the strong measurement model.

7. With the same specification of priors and hyper-priors, results of MSEM are comparable in using WinBUGS and in using Mplus 7.11 with Bayes estimation (ESTIMATOR = BAYES). WinBUGS code is available from the first author upon request.
8. There are 120 item parameters (20 items × 6 kinds of item parameters) and 5 structural parameters.
9. It is not problematic when RMSE from MSEM is compared with that of model parameter estimates in IRT model parameter recovery studies.

Supplemental Material

The online supplementary material is available at http://apm.sagepub.com/supplemental

References

Aitkin, M., & Alfo, M. (2003). Longitudinal analysis of repeated binary data using autoregressive and random effect modeling. *Statistical Modelling*, *3*, 291-303.

Asparouhov, T., & Muthén, B. (2013). *IRT in Mplus* (Technical report). Los Angeles, CA: Muthén & Muthén.

Battauz, M., Bellio, R., & Gori, E. (2011). Covariate measurement error adjustment for multilevel models with application to educational data. *Journal of Educational and Behavioral Statistics*, *36*, 283-306.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation of multidimensional IRT models. *Psychometrika*, *66*, 541-562.

Bejar, I. I. (1980). Biased assessment of program impact due to psychometric artifacts. *Psychological Bulletin*, *87*, 513-524.

Bloom, H., Richburg-Hayes, L., & Black, A. (2007). Using covariates to improve precision. *Educational Evaluation and Policy Analysis*, *29*, 30-59.

Bottge, B. A., Ma, X., Gassaway, L., Butler, M., & Toland, M. D. (2014). Detecting and correcting fractions computation error patterns. *Exceptional Children*, *80*, 236-254.

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

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

Fox, J.-P. (2010). *Bayesian item response modeling*. New York, NY: Springer.

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., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457-472.

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

Jeon, M., & Rabe-Hesketh, S. (2012). Profile-likelihood approach for estimating generalized linear mixed models with factor structures. *Journal of Educational and Behavioral Statistics*, *37*, 518-542.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*, 631-639.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79-93.

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.

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.

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.

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

Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2008). Optimal designs for multilevel studies. In J. de Leeuw & E. Meijer (Ed.), *Handbook of Multilevel Analysis* (pp. 177-206). New York: Springer.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*, 338-354.

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

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146-178.

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.

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

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*, 173-185.

Raudenbush, S. W., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness*, *1*, 138-154.

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

Schochet, P. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, *33*, 62-87.

Shin, Y., & Raudenbush, S. W. (2010). A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics*, *35*, 26-53.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229-239.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, *64*, 583-616.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS user manual*. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

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*, 2988-3005.

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.