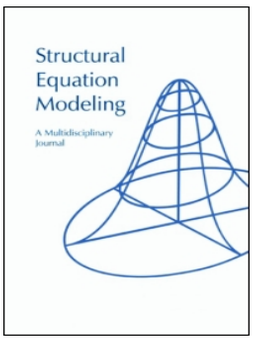

Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: http://www.tandfonline.com/loi/hsem20

Multilevel Mediation With Small Samples: A Cautionary Note on the Multilevel Structural Equation Modeling Framework

Daniel McNeish

To cite this article: Daniel McNeish (2017) Multilevel Mediation With Small Samples: A Cautionary Note on the Multilevel Structural Equation Modeling Framework, Structural Equation Modeling: A Multidisciplinary Journal, 24:4, 609-625, DOI: 10.1080/10705511.2017.1280797

To link to this article: https://doi.org/10.1080/10705511.2017.1280797

Structural Equation Modeling: A Multidisciplinary Journal, 24: 609–625, 2017

ISSN: 1070-5511 print / 1532-8007 online
DOI: 10.1080/10705511.2017.1280797

Routledge
Taylor & Francis Group

TEACHER'S CORNER

Multilevel Mediation With Small Samples: A Cautionary Note on the Multilevel Structural Equation Modeling Framework

Daniel McNeish
University of North Carolina, Chapel Hill

Multilevel structural equation modeling (ML-SEM) for multilevel mediation is noted for its flexibility over a system of multilevel models (MLMs). Sample size requirements are an overlooked limitation of ML-SEM (100 clusters is recommended). We find that 89% of ML-SEM studies have fewer than 100 clusters and the median number is 44. Furthermore, 75% of ML-SEM studies implement 2–1–1 or 1–1–1 models, which can be equivalently fit with MLMs. MLMs theoretically have lower sample size requirements, although studies have yet to assess small sample performance for multilevel mediation. We conduct a simulation to address this pervasive problem. We find that MLMs have more desirable small sample performance and can be trustworthy with 10 clusters. Importantly, many studies lack the sample size and model complexity to necessitate ML-SEM. Although ML-SEM is undeniably more flexible and uniquely positioned for difficult problems, small samples often can be more effectively and simply addressed with MLMs.

Keywords: multilevel mediation, multilevel SEM, small sample

Since the seminal work of Baron and Kenny (1986), mediation has been one of the most prominently featured statistical methods not only within psychology, but across all scientific disciplines. For instance, as of 2014, Baron and Kenny (1986) was the 33rd most cited paper of all-time based on Web of Science citations and the most cited paper published in a psychology journal (van Noorden, Maher, & Nuzzo, 2014). Similarly, Baron and Kenny (1986) is the 18th most cited paper of all-time based on Google Scholar citations and, at the time of this writing, the paper has almost 65,000 Google Scholar citations. Clearly, mediation is a relevant and widely used method.

Despite the popularity of the Baron–Kenny method for mediation, there have been some noted disadvantages of its use in the 30 years since its publication. One notable limitation is that the Baron–Kenny method makes an independence assumption that precludes its use with data that are clustered such as employees within companies commonly found in organizational science or students within schools, ubiquitous in education research (Krull & MacKinnon, 1999; Preacher, Zhang, & Zyphur, 2011; Preacher, Zyphur, & Zhang, 2010; Zhang, Zyphur, & Preacher, 2009). This limitation has been circumvented by adapting mediation models to settings with multiple levels. Initially, most of these developments occurred in the multilevel model (MLM) framework, presumably due to the similarity between multilevel regression models and the single-level regression models used in Baron and Kenny (1986). Krull and MacKinnon (1999) first outlined how mediation effects could be assessed with a system of multilevel models, somewhat reminiscent of the Baron–Kenny method with single-level data. Pituch, Whittaker, and Stapleton (2005) showed how a 1–1–1 mediation model (i.e., a model where the independent variable X, the mediator M, and the dependent variable Y, respectively, are all collected at Level 1 of the hierarchy) could be modeled if all

Correspondence should be addressed to Daniel McNeish, 100 E. Franklin Street Suite 200, Chapel Hill, NC, 27599. E-mail: dmcneish@email.unc.edu

effects are fixed. Bauer, Preacher, and Gil (2006) extended the 1–1–1 model to allow for random slopes of the coefficients and for moderated mediation.

Although a useful extension for correlated data that arise from a hierarchical structure, the Krull–MacKinnon method (hereafter referred to simply as MLMs) relies on a system of MLMs that precludes their use with types of research questions where either *M* or *Y* occur are collected at Level 2 (Preacher et al., 2010; Zhang et al., 2009) meaning that mediation models such as the 1–1–2 model, 1–2–1 model, 1–2–2 model, or 2–1–2 model cannot be assessed via MLMs because a dependent variable in the system (either *M* or *Y*) in MLMs cannot be collected at a level of the hierarchy above a predictor variable (either *X* or *M*). Alternatively, the 2–2–1 model can be modeled with MLMs although it must necessarily be done with a nonsimultaneous process (Krull & MacKinnon, 2001; Pituch, Stapleton, & Kang, 2006).

To combat these limitations and position any multilevel mediation question within a single general framework, Preacher et al. (2010) outlined multilevel structural equation modeling (ML-SEM) via a multivariate path diagram for accommodating multilevel mediation, broadly defined. The advantage of the general framework is that the existing MLM approach is featured as a special case and ML-SEM also extends to types of models that cannot be fit with MLMs (Pituch & Stapleton, 2011).

Despite the undeniable advantages of the more general and more flexible ML-SEM specification, there is a potentially important drawback of such a framework: sample size requirements. Despite the commonly held adage that the SEM and MLM are mathematically equivalent (e.g., Curran, 2003), there are differences in the estimation of these models that can lead to vastly different performance in smaller samples (Bauer, 2003; Curran, 2003).[1] ML-SEM is specified in the more general SEM framework, which employs general estimation methods like full maximum likelihood. Although not generally problematic, full maximum likelihood is an asymptotic method that is known to yield biased estimates (especially downwardly biased variance component and standard error estimates) when the number of clusters is small or modest (Browne & Draper, 2006; Maas & Hox, 2005). Hox and Maas (2001) conducted a study on sample size requirements for ML-SEM and recommend at least 100 clusters under most conditions to obtain trustworthy estimates of cluster-level variance and regression path standard errors (although 50 clusters sufficed with low intraclass correlations and balanced data in their study).

Preacher et al. (2010) explicitly noted this limitation and dedicated half a journal page to discussing potential issues related to having few clusters in multilevel mediation in ML-SEM. Despite the rapid adoption of ML-SEM for multilevel mediation (as evidenced by the 700+ Google Scholar citations of Preacher et al., 2010 since its publication), methodological studies have yet to evaluate how susceptible ML-SEM is to the common analytical situation where there are few clusters present and, if so, whether alternative options such as MLMs are viable in such situations. Furthermore, no reviews of multilevel mediation exist to examine the extent to which these models are fit to data with less-than-enviable sample sizes.

We first provide a literature review to investigate the types of sample sizes and multilevel mediation models that are used in empirical studies. This evidence is used to justify why the issues discussed in this article are worthwhile. Then, we overview the MLM and ML-SEM approaches to multilevel mediation. We continue with a review of the relevant literature pertaining to modeling clustered data with few clusters. A Monte Carlo simulation study is conducted to compare the performance of ML-SEM and MLMs. We conclude with a discussion of our findings, the relevance to empirical researchers, and recommendations for best practice.

LITERATURE REVIEW

When dealing with clustered data in behavioral sciences, large sample sizes (particularly at Level 2) are difficult to amass (Dedrick et al., 2009; Hox, van de Schoot, & Matthijsse, 2012; Maas & Hox, 2005). To provide evidence for this claim in the context of multilevel mediation, we conduct a literature review of empirical multilevel mediation studies using ML-SEM because this method has higher theoretical sample size requirements and is becoming the default method with which to model multilevel mediation. We located studies on Google Scholar by searching papers that had cited the seminal Preacher et al. (2010) paper. At the time of this writing, there are almost 700 such papers total and reviewing each would have been a daunting task. Instead, we took a random sample of 10% of these papers so that the number of reviewed studies is 70. There are a number of dissertations, book chapters, non-English papers, review papers, and methodological papers that cite Preacher et al. (2010) and these were outside of our scope. When one of these papers was sampled (19 papers fell into one of these categories, 21%), we discarded it and drew another paper at random until we accumulated 70 empirical studies that fell within our scope. We reviewed these 70 studies to determine

[1] This equivalence is frequently cited in the context of growth models where there is the most overlap between the modeling frameworks. However, the equivalence can be more tenuous for other types of models (Bauer, 2003). Even within the context of growth models, there are a handful of models that can only be specified in one framework but not the other. For instance, the latent basis model, which estimates loadings from the slope factor to the observed variables to empirically model nonlinear growth, is easily fit as a SEM but cannot be fit as an MLM (Curran, 2003). Thus, even though there can be a large degree of overlap between the frameworks, there is not always a strict one-to-one mapping. A presentation by Muthén and Muthén (2010) further differentiates the MLM and SEM frameworks (Slides 29–38, https://www.statmodel.com/download/Topic3-v.pdf).

(a) how many clusters were present in the data, (b) the estimation method used in the analysis, (c) the type of multilevel mediation model fit, and (d) the content area of the study. A list of the sampled papers and the extracted information is included in the Appendix for interested readers.

In these 70 studies, the median number of clusters was 44 and the mean was 60.41 clusters, with a standard deviation of 56.53 (the sample was positively skewed as there were a few studies with a much larger number of clusters; range = 6–389). Of these 70 studies, 39 studies (56%) had fewer than 50 clusters and 62 studies (89%) had fewer than 100 clusters, demonstrating that relatively few studies achieve the sample sizes that satisfy the recommendations of Hox and Maas (2001) for ML-SEM. Table 1 shows the count and percentage of studies that feature each of the possible multilevel mediation designs. Some studies contained multiple hypotheses so Table 1 contains two separate counts: one for the primary research question and one for the all research questions in the paper. The most common primary model tested was the 2–1–1 mediation model (39%), followed closely by the 1–1–1 mediation model (37%). Three additional studies conducted secondary analyses with 1–1–1 models, pushing the overall percentage to 41%. Although we initially hypothesized that Bayesian methods would be somewhat common, as they are an attractive option for smaller samples, only two reviewed studies (3%) estimated the model with Bayesian methods and both used M*plus*'s default priors for all parameters in the model (Carr & Chung, 2014; Miranda et al., 2016). As more descriptive information of this review, 14% of studies were longitudinal rather than cross-section and the most common data structures involved students or teachers nested within teachers or schools (30%) and employees within teams or managers (40%).

Although ML-SEM is undoubtedly more flexible (i.e., 30% of studies in this sample of studies cannot be fit with MLMs), a majority of studies fail to meet the minimum threshold for estimates to be trustworthy and Type I error rates to be well-behaved in ML-SEM. Furthermore, the review showed that a majority of multilevel mediation models fit in empirical studies are either 2–1–1 or 1–1–1 models (76% of primary models), which can be equivalently fit MLMs. As discussed in more detail in the subsequent sections, MLMs can use estimation methods and corrections that are more suitable for smaller samples that appear to be the rule rather than the exception. Furthermore, not only do these small sample methods not exist in ML-SEM software, but no developments in statistical theory for their implementation exist in ML-SEM. We provide a dedicated section later in this article to discuss the theory and application of these small sample methods in MLMs and why they cannot be directly applied to ML-SEM in their current form. Before this discussion, we overview mediation in each of these respective frameworks.

TABLE 1
Results From Review of Random Sample of Studies Citing Preacher et al. (2010)

Model	*Primary Hypothesis*	*Primary %*	*All Hypotheses*	*All %*
1–1–1	26	37%	29	41%
1–1–2	1	1%	2	3%
1–2–1	3	4%	3	4%
1–2–2	2	3%	2	3%
2–1–1	27	39%	27	39%
2–1–2	4	6%	4	6%
2–2–1	6	9%	9	13%
2–2–2[a]	1	1%	1	1%

Note. The Primary column count sums to 70 and the percentage sums to 100% because each study was counted only once. In the All columns, each study could appear multiple times so the column count sums to 77 and the percentage sums to 110%.

[a]Preacher et al. (2010) did not consider the 2–2–2 model within their general framework. However, one study explicitly referred to their model as a 2–2–2 model, so it is reported as such here.

MULTILEVEL MEDIATION WITH A SYSTEM OF MULTILEVEL MODELS

The Krull–MacKinnon method for using MLMs to assess multilevel mediation is similar in principle to the Baron–Kenny method in that it uses a system of univariate models to estimate mediation paths. To maintain a targeted narrative, the discussion in this section and the subsequent section on ML-SEM focuses on the 2–1–1 model because it can be fit in either framework and was the most common model used in empirical studies in our literature review.

Using a traditionally specified MLM,

$$M_{ij} = \beta_{M0j} + r_{Mij} \tag{1a}$$

$$\beta_{M0j} = \gamma_{M00} + \gamma_{M01}X_j + u_{M0j} \tag{1b}$$

represents the mediation path for the effect of the Level 2 independent variable X_j on the Level 1 mediator variable M_{ij} where i and j are indexes for the ith observation nested within the jth cluster. γ_{M00} is the intercept for the mediation variable, γ_{M01} is the cross-level effect of X on M, and u_{M0j} is the random intercept for M. The remainder of the mediation model that is concerned with the outcome variable is then represented by a separate set of equations,

$$Y_{ij} = \beta_{Y0j} + \beta_{Y1j}M_{ij} + r_{Yij} \tag{2a}$$

$$\beta_{Y0j} = \gamma_{Y00} + \gamma_{Y01}X_j + u_{Y0j} \tag{2b}$$

$$\beta_{Y1j} = \gamma_{Y10}. \tag{2c}$$

The indirect effect of X on Y through M is often calculated as the product of $\gamma_{M01} \times \gamma_{Y10}$. As noted in Preacher et al. (2010), Preacher et al. (2011), and Zhang et al. (2009), this specification is often inadequate because it does not separate the effect that

M_{ij} has on Y_{ij} at Level 1 and at Level 2. That is, unless the contextual effect is exactly 0, γ_{Y10} represents a conflated effect that is a weighted average of the Level 1 and Level 2 components of M_{ij}. If the effect of M_{ij} at Level 1 and Level 2 is not equal, then the resulting indirect effect will be biased.

Fortunately, the issue of conflated effects can be addressed rather simply with a strategic choice regarding the centering of M_{ij} in Equation 2a. If M_{ij} is cluster-mean centered and then the cluster means $M_{.j}$ are added as a predictor in Equation 2b, the model will appropriately disaggregate the effects of M_{ij} into Level 1 and Level 2 effects. Equation 2 would be transformed such that,

$$Y_{ij} = \beta_{Y0j} + \beta_{Y1j}(M_{ij} - M_{.j}) + r_{Yij} \tag{3a}$$

$$\beta_{Y0j} = \gamma_{Y00} + \gamma_{Y01}X_j + \gamma_{Y02}M_{.j} + u_{Y0j} \tag{3b}$$

$$\beta_{Y1j} = \gamma_{Y10} \tag{3c}$$

where $M_{.j}$ is the mean of the jth cluster. In Equation 3b, γ_{Y02} now represents only the effect of M on Y at Level 2, the same level as the independent variable X. Therefore, the appropriate calculation of the indirect effect should be $\gamma_{M01} \times \gamma_{Y02}$ because this appropriately captures only the Level 2 effect of X on Y through M. If the contextual effect is zero, then $\gamma_{M01} \times \gamma_{Y10} = \gamma_{M01} \times \gamma_{Y02}$.

MULTILEVEL SEM FRAMEWORK

Due to computational complexity, ML-SEM historically was fraught with estimation challenges that often mandated that all slopes in the model be fixed (e.g., Bauer, 2003; Metha & Neale, 2005) or that led to poor estimates in the presence of missing data and unbalanced clusters (e.g., Muthén, 1989, 1991, 1994; Muthén & Satorra, 1995). However, recent methodological advances have largely obviated these concerns (Muthén & Asparouhov, 2008; Rabe-Hesketh, Skrondal, & Pickles, 2004), which has led to increased feasibility and usage of these types of models, especially because they can be implemented in the mainstream software.

A general ML-SEM model is specified by three separate components: the Level 1 measurement model, the Level 1 structural model, and the Level 2 structural model. As outlined by Preacher et al. (2011), these three components can be written as follows:

$$\text{Level 1 measurement model: } \mathbf{Y}_{ij} = \boldsymbol{\nu}_j + \boldsymbol{\Lambda}_j\boldsymbol{\eta}_{ij} + \mathbf{K}_j\mathbf{X}_{ij} + \boldsymbol{\varepsilon}_{ij} \tag{4a}$$

$$\text{Level 1 structural model: } \boldsymbol{\eta}_{ij} = \boldsymbol{\alpha}_j + \mathbf{B}_j\boldsymbol{\eta}_{ij} + \boldsymbol{\Gamma}_j\mathbf{X}_{ij} + \boldsymbol{\zeta}_{\mathbf{ij}} \tag{4b}$$

$$\text{Level 2 structural model: } \boldsymbol{\eta}_j = \boldsymbol{\mu} + \boldsymbol{\beta}\boldsymbol{\eta}_j + \gamma\boldsymbol{X}_j + \boldsymbol{\zeta}_j \tag{4c}$$

It might seem confusing that there is a measurement equation at Level 1 even though we are only discussing mediation models in which the variables are all observed. This is due to the way that ML-SEM parameterizes the model: Each observed variable is partitioned into Level 1 and Level 2 components (often alternatively referred to as within and between components, respectively). Each of these components is then treated as a latent variable. For example, a mediator variable M will be partitioned into two latent variables, and these latent variables are used in the model (one at each level) rather than the observed variable, even if the conceptual model only contains observed variables. This is similar to the variance partition that occurs within MLM when assessing Level 1 and Level 2 variance; however, as ML-SEM, this is notationally represented as a measurement model. In the context of multilevel mediation with only observed variables, the "measurement model" merely consists of a latent variable loading on a single item with the loading constrained to one.

The equivalent 2–1–1 model that corresponds to Equations 1 and 3 occurs when $\boldsymbol{\nu}_j = \boldsymbol{\nu} = \mathbf{0}$ (equivalent to cluster-mean centering because the mean within each of the j clusters is 0), $\boldsymbol{\Lambda}_j = \boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_{\mathbf{L1}} | \boldsymbol{\Lambda}_{\mathbf{L2}}]$ (the measurement model is constant across all j clusters because all variables are observed), $\mathbf{K}_j = \mathbf{K} = \mathbf{0}$ (there are no predictors of the observed variables), $\mathbf{B}_j = \mathbf{B} = \mathbf{0}$ (there are no paths between Level 2 variables), and $\boldsymbol{\Gamma}_j = \boldsymbol{\Gamma} = \mathbf{0}$ (there are no additional Level 1 predictors) such that Equations 4a, 4b, and 4c reduce to

$$\mathbf{Y}_{ij} = \boldsymbol{\Lambda}\boldsymbol{\eta}_{ij} = \begin{bmatrix} X_{ij} \\ M_{ij} \\ Y_{ij} \end{bmatrix} = \left[\begin{array}{cc|ccc} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{array}\right] \begin{bmatrix} \eta_{M_{ij}} \\ \eta_{Y_{ij}} \\ -- \\ \eta_{X_j} \\ \eta_{M_j} \\ \eta_{Y_j} \end{bmatrix} \tag{5a}$$

$$\boldsymbol{\eta}_{ij} = \boldsymbol{\alpha}_j + \mathbf{B}\boldsymbol{\eta}_{\mathbf{ij}} + \boldsymbol{\zeta}_{ij} = \begin{bmatrix} \eta_{M_{ij}} \\ \eta_{Y_{ij}} \\ -- \\ \eta_{X_j} \\ \eta_{M_j} \\ \eta_{Y_j} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -- \\ \alpha_{\eta_{X_j}} \\ \alpha_{\eta_{M_j}} \\ \alpha_{\eta_{Y_j}} \end{bmatrix} + \left[\begin{array}{cc|ccc} 0 & 0 & 0 & 0 & 0 \\ B_{YM} & 0 & 0 & 0 & 0 \\ -- & -- & -- & -- & -- \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] \begin{bmatrix} \eta_{M_{ij}} \\ \eta_{Y_{ij}} \\ -- \\ \eta_{X_j} \\ \eta_{M_j} \\ \eta_{Y_j} \end{bmatrix} + \begin{bmatrix} \zeta_{M_{ij}} \\ \zeta_{Y_{ij}} \\ -- \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{5b}$$

$$\boldsymbol{\eta}_j = \boldsymbol{\mu} + \boldsymbol{\beta}\boldsymbol{\eta}_j + \boldsymbol{\zeta}_j$$
$$= \begin{bmatrix} \alpha_{\eta_{X_j}} \\ \alpha_{\eta_{M_j}} \\ \alpha_{\eta_{Y_j}} \end{bmatrix} = \begin{bmatrix} \mu_{\alpha_{\eta_{X_j}}} \\ \mu_{\alpha_{\eta_{M_j}}} \\ \mu_{\alpha_{\eta_{Y_j}}} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ \boldsymbol{\beta}_{MX} & 0 & 0 \\ \boldsymbol{\beta}_{YX} & \boldsymbol{\beta}_{YM} & 0 \end{bmatrix} \begin{bmatrix} \alpha_{\eta_{X_j}} \\ \alpha_{\eta_{M_j}} \\ \alpha_{\eta_{Y_j}} \end{bmatrix} + \begin{bmatrix} \zeta_{B_{YM_j}} \\ \zeta_{\alpha_{\eta_{X_j}}} \\ \zeta_{\alpha_{\eta_{M_j}}} \end{bmatrix}. \quad (5c)$$

Figure 1 depicts the 2–1–1 model from Equation 5 as a multilevel path diagram. Equation 5a shows the measurement model. Because X is a Level 2 variable in this context, it has all zero values left of the vertical dashed line in $\boldsymbol{\Lambda}$ (which corresponds to the lower rectangle in Figure 1) because it does not load on any Level 1 latent variables (e.g., there is zero Level 1 variance for X). Both M and Y are Level 1 variables so they are partitioned into Level 1 and Level 2 latent variables and have nonzero entries on either side of the vertical dashed line in Λ (i.e., there are latent variables for M and Y in both rectangles in Figure 1). Again, because all variables are observed, the loadings are not estimated and are instead constrained to either 1 or 0 because each latent variable loads on only one observed variable. There are five nonzero values in $\boldsymbol{\Lambda}$ and five paths going into X, Y, and M in Figure 1.

The α_j vector represents estimates of the cluster means, similar to the advantage of including $M_{.j}$ as a Level 2 predictor in Equation 3b. Because $\boldsymbol{\nu}_j = \boldsymbol{\nu} = \mathbf{0}$ in Equation 5, estimating α_j similarly decomposes effects in Level 1 and Level 2 components. The $\mathbf{B}$ matrix contains one nonzero entry because there is only one relation completely within Level 1 in the model, namely from M to Y. This corresponds to Equation 3a within MLMs where there is only one nonintercept predictor in Equation 3a. There are only two error terms in Equation 5b because X does not have any variability at Level 1. Equation 5c shows the paths between the Level 2 latent variables (i.e., the between effects). Because X is located at Level 2, these are the paths that are used to calculate the indirect effect—namely, the effect of X on Y through M is equal to $\beta_{MX} \times \beta_{YM}$.

MULTILEVEL DATA WITH FEW CLUSTERS

As alluded to in the introduction, estimation methods for multilevel data tend to be greatly affected when the data either have few clusters (e.g., Bell, Morgan, Schoeneberger, Kromrey, & Ferron, 2014; Browne & Draper, 2006; Hox et al., 2012; Maas & Hox, 2005; McNeish & Stapleton, 2016) or if there are few observations per cluster (e.g., Clarke, 2008). The general issue with small sample sizes (beyond the obvious that power will be reduced) is that the parameter estimates can be highly biased with standard approaches to estimation, making it difficult to trust and interpret the model. Specifically, with few clusters, the variance components tend to be highly downwardly biased (they are too small) with traditional maximum likelihood estimation, which

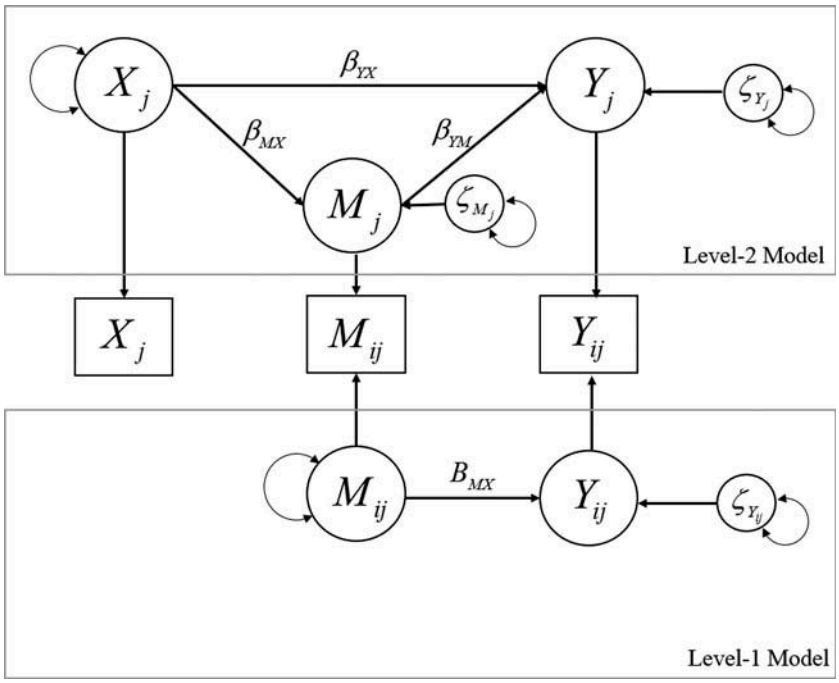

FIGURE 1 Path diagram for a three-variable mediation model in the multilevel structural equation modeling framework where X is at Level 2 and M and Y are at Level 1.

further leads the standard error estimates of the coefficients to be too small (because the variance components are prominent in the standard error formula). This, coupled with inaccurate Fisher information due to poor approximations of the curvature of the likelihood function with small samples, leads to highly inflated operating Type I error rates (Maas & Hox, 2005).

With a traditional MLM, conventional wisdom suggests a Level 2 sample size of 30 clusters with each cluster containing 30 observations (Kreft, 1996). Recent simulations along with innovations in small sample methods have shown that fewer clusters might be necessary if proper precautions are taken. For instance, McNeish and Stapleton (2016) suggested that 20 clusters with five or more observations per cluster might be sufficient if the model is estimated with restricted maximum likelihood (REML) instead of full maximum likelihood. Furthermore, a Kenward–Roger correction (which adjusts the standard error estimates and more accurately approximates degrees of freedom for inferential tests) can allow MLMs to perform well with Level 2 sample sizes into the single digits for small or moderately sized models.

Specific to ML-SEM, Hox and Maas (2001) and Hox, Maas, and Brinkhuis (2010) found that the number of clusters needs to be larger than 50 with more ideal data (low intraclass correlation and equal Level 1 sample sizes in each cluster) to produce reasonable estimates in the Level 2 portion of the model (where all the mediation paths are in an ML-SEM multilevel mediation model) with 100 being recommended as a desirable number for more common scenarios (unbalanced clusters, higher intraclass correlations). Small sample methodological studies have not been conducted in the explicit context of multilevel mediation and such small sample properties will be assessed in our forthcoming simulation.

Although the focus of this article is not explicitly on Bayesian methods, some reviewed studies applied Bayesian methods to accommodate small sample issues, so this topic deserves some mention. Bayesian methods are often recommended with smaller samples because sampling variability is straightforward to assess and does not require asymptotics, Fisher information, or the central limit theorem as is necessary with maximum likelihood (e.g., Muthén & Asparouhov, 2012; van de Schoot et al., 2014). Although Bayesian methods possess advantages with smaller samples, these advantages are not automatic and are highly dependent on the specification of informative prior distributions (Depaoli, 2014; van de Schoot, Broere, Perryck, Zondervan-Zwijnenburg, & van Loey, 2015). Using highly diffuse priors as are the default in user-friendly software (e.g., M*plus*) does not trigger these small sample advantages and researchers must set informative priors to benefit from the advantages provided by Bayesian methods (Depaoli & Clifton, 2015; Depaoli & van de Schoot, 2015; van de Schoot et al., 2015; van de Schoot et al., 2014). Colloquially, Bayesian methods at their root are about combining information. Small sample analyses have limited information by definition, so additional information can be included to assist the analysis through informative prior distributions (because the posterior is a combination of the information from the data and the information from the prior). Thus, an advantage of Bayesian methods with smaller samples is that the resulting estimates can be based on more information than a comparable frequentist analysis (which relies on only the information provided by the data). However, diffuse prior distributions negate this advantage and can result in poorer estimates because the additional information that is added through the priors can detract from the information provided from the data. This will be demonstrated for the multilevel mediation context in the forthcoming simulation.

SMALL SAMPLE METHODS FOR MULTILEVEL MEDIATION

Although REML and the Kenward–Roger correction have been shown to yield more appropriate estimates of variance components and standard errors, respectively, these methods are not uniformly available in all frameworks for multilevel mediation. That is, REML and the Kenward–Roger correction were developed specifically for MLMs and the portability of these methods to SEM is tenuous because the necessary components for these corrections are not generalizable enough to accommodate the heightened flexibility of SEM. Although some work has advanced rudimentary forms of REML for the specific types of models that can be fit as SEMs (mostly growth models; Cheung, 2013), applications of REML for SEM are still extremely limited generally and no statistical theory exists for REML estimation of ML-SEM (for difficulties associated with extending REML beyond MLMs, see McNeish, 2016a).[2] Cheung (2013) noted this explicitly: "Further research should address how REML estimation can be implemented in general MLSEM" (p. 163).

Specifically, Cheung (2013) noted that the REML is straightforward to calculate for SEMs that can be equivalently fit as an MLM. In matrix notation, the MLM from Equations 1, 2, and 3 can be written as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i. \quad (6)$$

Cheung (2013) explained how some SEMs can be transformed into the form of Equation 6 such that existing MLM REML equations can be applied. However, Cheung did not derive a

[2] Additionally, there are no attempts to extend the Kenward–Roger correction to SEM, presumably because a primary component of the correction is the degrees of freedom for inferential tests. SEM software tends to use asymptotic Z tests for inference instead of finite sample test statistics that require degrees of freedom.

general REML estimator for SEM broadly—he instead took advantage of the overlap that exists between the two modeling frameworks.[3] As such, REML for SEM exists only for models that have fixed design matrices. That is, all elements of the loading matrix from latent variables to observed variables in SEM must be constrained to prespecified values, meaning that outcomes and predictor variables must necessarily be observed and cannot be latent for REML to exist for SEM. As noted in Cheung (2013), these parameters are indeed known a priori for many (but not all) growth models (e.g., latent basis models and second-order growth models are notable counterexamples). However, matrices of fixed loadings are not the norm for ML-SEM models generally, nor for confirmatory factor analyses or mixture models.

The issue of the overlap between MLMs and SEMs is discussed by Muthén and Muthén (2010). Figure 2 visually depicts arguments in Muthén and Muthén (2010) and shows the overlap between MLMs and SEM. Although the overlap is considerable, many types of SEMs cannot be reparameterized into the form of Equation 6. As a result, there is no REML available for the model types in the white portion of the SEM circle. As directly referenced by Cheung (2013), ML-SEM is one such model that falls in the nonoverlapping section of Figure 2. This is particularly problematic considering the frequency with which small samples occur for multilevel mediation models where REML is desirable. Therefore, for multilevel mediation in the ML-SEM framework, full maximum likelihood (along with its known small sample biases) remains the primary frequentist estimation method.

CONFIDENCE INTERVAL METHODS

Inference about the indirect effect of X on Y through M is typically the primary focus of a mediation model; however, the sampling variability of this parameter is often difficult to estimate in the frequentist framework because the product of two normally distributed variables is not necessarily normal or symmetric (Aroian, 1947; Craig, 1936). As a result, there have been many recent developments in the literature to devise methods that appropriately reflect this asymmetry in the sampling distribution of the indirect effect. Although bootstrapping is a popular approach for single-level data, it can be problematic when data are clustered because the manner in which to resample the data is not necessarily straightforward (Preacher & Selig, 2012; van der Leeden, Meijer, & Busing, 2008; Wang, Carpenter, & Kepler, 2006). This difficulty is reflected in M*plus*'s lack of support for bootstrapping with multilevel data (Muthén & Muthén, 2012). Bootstrapping multilevel data is still possible—Pituch et al. (2006) recommended residual bootstrapping with multilevel data to assuage concerns about how to sample units and Cameron, Gelbach, and Miller (2008) recommended a wild cluster bootstrap for similar reasons because their simulation showed that most other cluster bootstrap methods have difficulty maintaining reasonable Type I error rates with smaller samples. Moreover, computational complexity also becomes more prevalent in multilevel data and can present a larger barrier to implementation because bootstrapping requires estimating the model parameters after each random draw, and models for multilevel data do not necessarily converge quickly (Preacher & Selig, 2012). Specific to data with small samples, bootstrapping can also be problematic because variance can be attenuated due to ineffective resampling or even randomly selecting only a single cluster so that there is no between-cluster variance (Fritz, Taylor, & MacKinnon, 2012; Polansky, 1999).

In the following subsections and the subsequent simulation, we focus on three common frequentist methods for estimating confidence intervals of indirect effects that are useful with clustered data: the delta method (often referred to as Sobel's test; Sobel, 1982), the distribution of the product method (sometimes referred to by the program used to calculate it, PRODCLIN; MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002), and the Monte Carlo method (Preacher & Selig, 2012). The performance of these methods has not been systemically investigated in the context of multilevel data with few clusters.

Delta Method

The delta method to compute the standard error and the confidence interval of an indirect effect is widely implemented because of its simplicity due its closed-form solution. The ease of computation makes the delta method the default method for assessing indirect effects in software (Preacher & Selig, 2012). Once the two mediation paths (M on X and Y on M) and their standard errors have been estimated, then the standard error of the indirect effect is calculated with a straightforward formula,

$$SE(\hat{a}\hat{b}) = \left[\hat{a}Var(\hat{b}) + \hat{b}Var(\hat{a}) + Var(\hat{a})Var(\hat{b})\right]^{1/2} \quad (7)$$

where $\hat{a}$ is the estimated coefficient for M on X, $\hat{b}$ is the estimated coefficient for Y on M, $Var(\hat{a})$is the square of the estimated standard error for $\hat{a}$, and $Var(\hat{b})$ is the square of the estimated standard error for $\hat{b}$. The indirect effect can then be tested with a Z test such that

$$Z = \frac{\hat{a}\hat{b}}{SE(\hat{a}\hat{b})} \dot{\sim} N(0,1) \quad (8)$$

[3] Although such manipulations are justifiable, the oft-cited paper by Curran (2003) questions the utility of such a transformation. As Curran (2003) noted, "Of most importance, I believe that if no other elements of the SEM are incorporated in a MLM then the SEM approach has nothing unique to offer over the standard MLM. … My recommendation is that if a particular research hypothesis can be fully evaluated using a standard MLM, by all means use the MLM approach" (pp. 564–565). Essentially, if SEMs necessarily convert models to MLMs to implement REML, researchers might as well cut out the intermediate SEM step and resort to MLMs from the beginning.

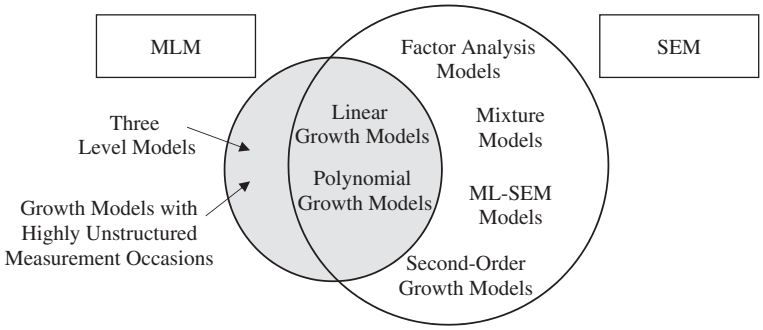

FIGURE 2 Venn dsiagram showing the degree of overlap between multilevel models (MLMs) and structural equation models (SEM). The gray area represents models that can be fit with restricted maximum likelihood (REML) via a multilevel regression framework or the REML method advanced by Cheung (2013). This figure is only concerned with models that are linear in the parameters and does not address models with discrete outcomes or models with random effects that enter the model nonlinearly (e.g., Gompertz or exponential curve models). The MLM circle is smaller because the MLM framework is less flexible than the SEM framework. Preacher (2011) discussed how three-level models can be fit in the multilevel structural equation modeling (ML-SEM) framework, although Preacher noted that proposed methods are not fully general (p. 725). Three-level MLMs are a general extension of two-level MLMs.

Alternatively, the asymptotic confidence interval for the indirect effect can be calculated by $\hat{a}\hat{b} \pm 1.96 \times SE(\hat{a}\hat{b})$. This is the method employed in M*plus* when the indirect effect is calculated by multiplying the two mediation paths together in a MODEL CONSTRAINT statement.

Although easy to calculate, the approximation of the product of the mediation paths (as noted by the dot above the tilde in Equation 8) as a normal distribution is not always appropriate (MacKinnon et al., 2002; Springer & Thompson, 1966). This approximation has also been noted to be less serviceable when sample sizes decrease in single-level models (Fritz & MacKinnon, 2007) and tends to have lower power to detect nonnull effects (MacKinnon et al., 2002), although its performance has not been explicitly tested in multilevel data with small samples.

Distribution of the Product Method

The delta method assumes that the product of the normally distributed random variables (the mediation paths) is normally distributed, but this is not necessarily true, mathematically speaking. MacKinnon, Fritz, Williams, and Lockwood (2007) noted that the distribution of the product of normal variables does not necessarily follow a familiar distribution (although the gamma distribution can sometimes be a suitable approximation). We do not delve into the mathematical details here because they are intricate and are beyond the scope of this article, but full details can be found in MacKinnon et al. (2007) or Tofighi and MacKinnon (2011), among others. The general idea, however, is to analytically derive the approximate distribution and then take the 2.5 and 97.5 quantiles of this distribution to form the 95% confidence interval (or other relevant quantiles for different confidence intervals).

We included this method because it is an analytic method that does not require resampling from the raw data to assess the sampling variability of the indirect effect (nor are the raw data even needed). Therefore, the difficulties that are evoked by bootstrapping with multilevel data are not a concern. That is, when using the PRODCLIN program to estimate the confidence interval of the indirect effect, researchers only need to provide the estimates of the mediation paths ($\hat{a}$ and $\hat{b}$) and the standard errors of these paths—the raw data are not necessary. This also makes the distribution of the product method equally feasible regardless of whether the model is fit with MLMs or ML-SEM.

Monte Carlo Method

The Monte Carlo method is highly related to bootstrapping in that it makes use of resampling to calculate the sampling distribution of the indirect effect. However, instead of resampling the data and reestimating the model, the Monte Carlo method resamples from sample statistics, which alleviates the need to sample from the data and therefore circumvents the difficulty evoked with resampling data with a multilevel structure (as is done with bootstrapping).

To implement the Monte Carlo method, like the distribution of the product method, researchers need only provide $\hat{a}$ and $\hat{b}$ along with their standard errors. These estimated values (whether obtained from MLMs or ML-SEM) are then used to define a multivariate normal distribution. Then, many draws (e.g., a few thousand) are taken from this distribution to form an empirical sampling distribution for the indirect effect. The 95% confidence interval is calculated from the 2.5 percentile and the 97.5 percentile of this distribution. This process can be represented statistically as

$$\begin{bmatrix} a^* \\ b^* \end{bmatrix} \sim MVN\left(\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix}, \begin{bmatrix} Var(\hat{a}) & Cov(\hat{a},\hat{b}) \\ Cov(\hat{a},\hat{b}) & Var(\hat{b}) \end{bmatrix} \right) \quad (9)$$

although $Cov(\hat{a}, \hat{b})$ is typically set to 0. Thousands of replications of a^* and b^* are generated from the multivariate normal distribution and the sampling distribution of $\hat{a} \times \hat{b}$ is then estimated by the empirical distribution of $a^* \times b^*$.

Similar to bootstrapping, the Monte Carlo method makes no assumptions about the sampling distribution of the indirect effect because the multiple repeated samples create an empirical sampling distribution—even though the draws are taken from a joint normal distribution, there is no parametric form specified on $a^* \times b^*$(Preacher & Selig, 2012). The advantage of the Monte Carlo method is that the raw data are not required because, unlike bootstrapping methods, the resampling is done from a multivariate normal distribution based on the path estimates rather than from a model estimated from resampled data. This gives the Monte Carlo method similar flexibility to bootstrapping because there is no assumption imparted on the shape of the sampling distribution. The benefit is that researchers can avoid the challenge of how to resample data with clustered observations concomitant with bootstrapping.

SIMULATION DESIGN

To assess the performance of MLMs and ML-SEM for multilevel mediation when there are few clusters in the data, we conduct a Monte Carlo simulation study to explore properties related interval coverage of the indirect effect (related to Type I error rate in the frequentist context) and nonnull detection rates (also called power in the frequentist context). We manipulate the number of clusters in the study to represent reasonably small samples ranging from 10 to 100 clusters (10, 15, 25, 50, 100). This range was selected because 10 is generally seen as the minimum number of clusters needed to conduct a multilevel analysis (Snijders & Bosker, 1993), 50 clusters tends to be the sample size at which small sample issues associated with full maximum likelihood begin to dissipate, and 100 is the current recommendation for ML-SEM (e.g., Hox & Maas, 2001). Although 100 might not necessarily seem to fall under the umbrella term of "few clusters" because the total sample size could be well into the thousands, with respect to clustered data, small sample bias could exist even at seemingly large samples for complex multivariate models such as those tested in the multilevel mediation framework, especially because the Level 2 sample size is the most important to consider (McNeish & Stapleton, 2016).

The sample size within clusters had two conditions, both of which were unbalanced to more closely mirror data that are seen in empirical studies (i.e., each cluster could have a potentially different number of observations). The first condition was 7 to 14 observations per cluster to represent a real-life context of students in classrooms or employees within teams. The second condition was 80 to 160 to represent the real-life context where clustering is due to geographical units such as countries or U.S. states where Level 1 sample sizes tend to be rather large but the Level 2 sample sizes tend to be modest.

We generate data from and fit 2–1–1 mediation models because this was found to be the most commonly used model for primary research questions and because this has been the model of interest in the foundational work on multilevel mediation (e.g., Preacher et al., 2011). Each model featured a three-variable mediation model and did not feature additional covariates. Intercepts were allowed to vary randomly but slopes were fixed across clusters. We tested two sets of parameter values to represent a small effect size for the indirect effect (explaining 4% of the total variance in *Y*) and medium effect size for the indirect effect (explaining 15% of the total variance in *Y*). These values were selected to roughly correspond to values used in the numerous simulations by MacKinnon and colleagues. The intraclass correlation was not manipulated and was constant at 0.20 for all conditions. This value was selected based on findings from Hedges and Hedberg (2007), which found intraclass correlation values in behavioral science research to be approximately 0.20, on average.

Three estimation methods are addressed. The first is REML with a Kenward–Roger correction with MLMs because this combination has repeatedly been found to yield estimates with desirable properties in previous studies on multilevel regression with few clusters (Bell et al., 2014; Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009). The second estimation method is full maximum likelihood with ML-SEM, which has become popular since its generalization by Preacher et al. (2010). The third is Bayesian Markov chain Monte Carlo (MCMC) conducted within ML-SEM (e.g., Yuan & MacKinnon, 2009).[4] Although not the primary focus of this article, a small number of studies in our review used Bayesian methods and Bayesian methods in general as growing increasingly popular in psychological studies (van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017). MLMs are estimated in SAS 9.3 with Proc Mixed and ML-SEM are estimated in M*plus* 7.3 (both full maximum likelihood and MCMC).

For the MCMC models, we retained the M*plus* default prior distributions on all parameters. This strategy was employed in the only two Bayesian papers found in our review, although recent research has noted that highly diffuse priors are not a wise decision with few clusters (e.g., McNeish, 2016b). We retain the M*plus* defaults to further demonstrate some of the issues associated with this practice and to hopefully expand the scope of recent work that has shown that Bayesian analysis with diffuse priors is not a

[4] Our focus here is multilevel mediation broadly, so we do not provide extensive background detail on MCMC estimation and we assume that readers have a general understanding of Bayesian methods. For readers unfamiliar with Bayesian methods, readable introductions can be found in Kruschke, Aguinis, and Joo (2012), van de Schoot et al. (2014), and Zyphur and Oswald (2015).

simple fix for small sample problems (much of this work pertains to growth models and might not have much traction for researchers working with mediation models). As other important MCMC options, we used two chains, set the minimum number of iterations to 10,000 per chain with a maximum of 100,000 iterations, discarded the first 10,000 iterations per chain, did not thin the chains, and reduced the proportional scale reduction convergence criteria from .10 to .03 to ensure that chains mixed properly.

Inferences for the indirect effect in the frequentist models were assessed with the delta method, the distribution of the product method, and the Monte Carlo method. The delta method formula was manually programmed, the distribution of the product method was obtained via the PRODCLIN SAS macro available from MacKinnon et al. (2007), and the Monte Carlo method with 5,000 random draws was programmed in SAS Proc IML and was adapted from R code provided on Preacher's Web site (http://www.quantpsy.org/medmc/medmc.htm). The indirect effect in the MCMC models was assessed by inspecting the 95% credible interval because the indirect effect is resampled with MCMC, which produces a credible interval that is directly calculable and not dependent on asymptotic normality.

Software Details

Despite the different programs used to estimate the models, the entire simulation was run from SAS with 1,000 replications in each cell of the simulation design. Because some utilities for multilevel mediation are only available in one software program (e.g., the Kenward–Roger correction is unique to SAS or Stata, ML-SEM is not available in SAS, the MONTECARLO utility in M*plus* is not available for 2–1–1 mediation models and MCMC estimation), the coding of the simulation presented some difficulties to work around all software limitations. We used the X command to call M*plus* as a DOS prompt from SAS when necessary. This was advantageous because (a) it allowed us to estimate ML-SEM models in M*plus*, which is arguably what most researchers would use if approaching multilevel mediation from the ML-SEM or a Bayesian approach (especially given the wealth of M*plus* code provided by Preacher et al., 2010), and (b) it allowed us to automate alternative methods for computing confidence intervals for indirect effects from M*plus* output.

Outcome measures

Three outcome measures are of interest in this study. The first is descriptive and was concerned with the number of replications that converged (definitions for what is considered a "converged" replication are described in the next section).

The second measure is the coverage of the confidence or credible interval (we use CI to refer to these quantities collectively across estimation type because they address the same concept). Because the simulation mixed frequentist and Bayesian estimation methods, the conventional Type I error rate simulation outcome is not available because it does not fundamentally apply to the Bayesian condition, as there are no null hypotheses to reject. Instead, we assess the information captured by Type I error rates with the CI coverage to keep the reported measures consistent across all estimation conditions. To give some background on how CI coverage is calculated, for each replication of the simulation study, a CI is estimated for the indirect effect. For each replication, we inspect whether the population value for that condition is located within the estimated CI. The percentage of replications where the population value is within the CI is then calculated. For a 95% CI, it is expected that 95% of replications will contain the population value. Based on criteria in Bradley (1978), if less than 92.5% percent of replications contain the population value, this indicates that the variability of the estimate is likely deficient (or that the point estimate is highly biased, skewing the interval left or right). Bradley also suggested that CI coverage rates that exceed 97.5% are similarly problematic. The operating Type I error rate for frequentist estimation conditions can be approximated by 1 – (CI coverage); for example, a coverage rate of 90% implies a Type I error rate of 10%).

The third outcome is the percentage of replications in which the estimated CI for the indirect effect did not contain 0. In the frequentist models, this is akin to statistical power because it represents the percentage of replications in which the models were able to detect that the indirect effect was (truly) nonnull. Statistical power does not apply to Bayesian methods by definition although the general idea of inferring whether a parameter is equal to 0 is still relevant. This measure is not intended to provide a recommendation for necessary sample sizes for designing studies because such determinations inherently vary based on the size of the model, the strength of the effects, and so forth. Instead, this outcome measure is intended to compare the relative performance of the different frameworks, estimation methods, and frequentist CI methods.

SIMULATION RESULTS

Nonconvergence

Whenever dealing with clustered data with few clusters, convergence to a solution could be an issue. Although the convergence in this study was rather good because the analysis model was fairly simple, convergence still must be assessed. To maintain a fair comparison, we only included replications in which all three methods (MLM, ML-SEM, and MCMC) converged so that

methods that did not converge were not artificially rewarded because other methods that converged might have produced poor estimates. Nonconvergence was defined by inadmissible estimates (e.g., negative variance estimate), nonpositive definite covariance matrices, or any replication where the software failed to produce estimates due to any error. Table 2 shows the number of replications that converged for all three methods that were used in subsequent analyses.

CI Coverage

Coverage of the 95% CI across conditions for each of the estimation methods is provided in Table 3. Immediately on inspecting Table 3, the most salient finding is that the CI coverage for the indirect effect from ML-SEM with full maximum likelihood is extremely poor when the number of clusters is small, especially when the cluster size is also small. For instance, in the 10-cluster, 7 to 14 observation per cluster, small effect size condition, all three CI methods produced a CI coverage rate in the mid-60% range, which is quite distant from the nominal rate of 95%. This represents an operating Type I error rate nearly seven times the nominal rate. With a small indirect effect size in the small cluster size condition, all ML-SEM CI methods failed to produce coverage inside Bradley's range with 50 clusters. CI coverage improved with larger cluster sizes. Otherwise, the estimates are so poor that they are likely unusable.

On the contrary, MLMs with REML and a Kenward–Roger correction did not encounter issues when faced with data that have few clusters. None of the three CI methods deviated from Bradley's range for any conditions in the simulation and no issues were present with as few as 10 clusters.

Bayesian MCMC with the M*plus* default priors shows the opposite pattern to maximum likelihood. With smaller cluster sizes, the CI coverage was too large with 25 or fewer clusters. This finding mirrors results from previous simulations on using diffuse priors with small samples: Diffuse prior distributions dominate the likelihood with smaller samples and, because the support of the prior is vast, the variance estimates are highly overestimated. This overcoverage has a dramatic adverse effect on the ability of the model to detect nonnull findings, which is discussed in the next section.

TABLE 2
Number of Replications in Each Condition That Converged for All Three Estimation Methods

	7–14 Cluster Size		*80–160 Cluster Size*	
Number of Clusters	*Small Effect*	*Medium Effect*	*Small Effect*	*Medium Effect*
10	792	865	897	944
15	864	921	938	964
25	921	973	978	994
50	984	998	996	1,000
100	1,000	1,000	1,000	1,000

TABLE 3
95% Confidence Interval (CI) Coverage Rates for Each Estimation and CI Method Across Simulation Conditions

			MLM			*ML-SEM*			*MCMC*
Cluster Size	*Effect Size*	*Number of Clusters*	*Delta*	*DP*	*MC*	*Delta*	*DP*	*MC*	*Credible Interval*
7–14	Small	10	94	93	94	**63**	**64**	**66**	**100**
		15	94	94	93	**78**	**78**	**79**	**100**
		25	94	94	94	**86**	**87**	**89**	**98**
		50	94	94	94	**90**	**91**	**92**	95
		100	95	95	94	93	93	94	95
	Medium	10	93	93	93	**76**	**76**	**77**	**100**
		15	94	93	94	**82**	**82**	**84**	**99**
		25	95	94	95	**90**	**90**	**92**	**98**
		50	96	96	93	96	96	96	**98**
		100	96	96	93	97	97	97	97
80–160	Small	10	93	93	93	**80**	**79**	**81**	**100**
		15	93	93	93	**86**	**86**	**88**	**98**
		25	93	93	93	**92**	**91**	93	96
		50	96	96	96	95	95	96	97
		100	97	97	97	96	96	96	96
	Medium	10	93	93	93	**87**	**88**	**89**	**99**
		15	94	94	93	**91**	**91**	**92**	**98**
		25	93	93	93	93	94	94	97
		50	97	97	97	95	95	95	96
		100	97	97	96	95	94	94	95

Note. MLM = multilevel model; ML-SEM = multilevel structural equation model; MCMC = Markov chain Monte Carlo; Delta = delta method; DP = distribution of the product method; MC = Monte Carlo method. Values shown in bold indicate coverage rates that are outside Bradley's range of (92.5, 97.5) for being reasonably close to the 95% nominal rate.

Detection of Nonnull Effects

Table 4 shows the percentage of replications in which the CI of the indirect effect did not contain 0 across all simulation conditions. Again, Table 4 conceptually shows power rates (although power is not strictly defined in a Bayesian context) and is intended to provide a relative comparison between methods. It is not meant to be interpreted for sample size determination. The most salient results in Table 4 pertain to the MCMC conditions. Specifically, the percentage of replications without 0 in the CI is vastly smaller compared to either the MLM or ML-SEM conditions at comparable sample sizes. As a prime example, with 25 clusters, small

TABLE 4
Nonnull Detection Rates for Each Method Across Simulation Conditions

			MLM			*ML-SEM*			*MCMC*
Cluster Size	*Effect Size*	*Number of Clusters*	*Delta*	*DP*	*MC*	*Delta*	*DP*	*MC*	*Credible Interval*
7–14	Small	10	17	23	23	—	—	—	0
		15	23	30	30	—	—	—	1
		25	38	41	41	—	—	—	7
		50	63	65	65	—	—	—	34
		100	78	81	81	62	78	79	79
	Medium	10	49	57	58	—	—	—	1
		15	67	72	72	—	—	—	6
		25	89	91	91	—	—	—	30
		50	99	99	99	65	72	71	67
		100	100	100	100	95	95	96	94
80–160	Small	10	19	26	26	—	—	—	1
		15	26	35	36	—	—	—	5
		25	42	47	47	—	—	36	24
		50	79	82	82	63	80	78	72
		100	98	98	98	98	99	98	98
	Medium	10	63	64	64	—	—	—	3
		15	80	80	80	—	—	—	21
		25	95	94	95	59	74	72	61
		50	99	99	99	94	95	95	94
		100	100	100	100	100	100	100	100

Note. MLM = multilevel model; ML-SEM = multilevel structural equation model; MCMC = Markov chain Monte Carlo; Delta = delta method; DP = distribution of the product method; MC = Monte Carlo method. Dashes indicate that coverage rates were too small and that nonnull detection rates are not trustworthy.

cluster sizes, and a medium effect size, 30% of MCMC CIs did not contain 0 compared to about 90% of CIs from the MLM condition.[5] With 15 or fewer clusters with small effects, MCMC was essentially useless for determining that any indirect effects were nonnull. As the number of clusters became larger, the percentage of CIs that did not contain 0 in the MCMC condition more closely matched percentages observed in the MLM and ML-SEM conditions because the prior was far less overbearing and the likelihood began to carry more relative weight in the posterior distribution. However, to reiterate findings from previous studies, widely diffuse priors with small samples are clearly a poor option.

The MLM conditions had a much higher probability to detect nonnull indirect effects than the MCMC conditions. Many of the ML-SEM rates were untrustworthy due to poor standard error estimates (which directly mischaracterizes rejection rates). The delta method tended to have the lowest power of the competing frequentist CI methods, although the distribution of the product method and the Monte Carlo method were comparable. As the number of clusters increased, differences between CI methods and between estimation methods were less noticeable.

LIMITATIONS AND CONCLUSIONS

Limitations

One limitation of this study is that we did not allow the mediation paths to randomly vary across clusters. We restricted the model to random intercepts to serve as a baseline assessment because including random slopes with few clusters will increase convergence problems and also decrease the quality of the estimates. Should readers be interested in small sample issues with multilevel mediation with random slopes, the results in this study are likely too optimistic.

We also investigated the context of mediation among three variables. Although the three-variable model is commonplace in statistical simulations to gauge the basic performance of various methods, in empirical studies, this type of model is rarely applied and additional variables are almost certainly included in the model (e.g., control variables, predictor variables). For empirical studies with many

[5] Although these power values might seem high for data with few clusters, keep in mind that the path from *M* to *Y* occurs at Level 1. In the smallest sample size condition in the simulation design, the overall Level 1 sample size is about 100 and the 25-cluster, small cluster size condition contains about 250 individuals. The values obtained in this study are approximately equal to values obtained by MacKinnon et al. (2002) if the effective sample sizes in this study are compared to their single-level sample sizes. The effective sample size approximates the amount of information present (in terms of sample size) if there were no clustering where $N_{eff} = N/[1 + (m - 1)\rho]$ where m is the average cluster size, N is the total sample size, N_{eff} is the effective sample size, and ρ is the intraclass correlation (Kish, 1965).

additional variables, the effect of few clusters is likely to be even worse. Additionally, researchers might wish to test several mediation models simultaneously (i.e., multiple *X*, *M*, and *Y* variables in a single model). This type of question is best assessed through ML-SEM, although sample size requirements increase as a function of model complexity.

We only investigated the performance of methods when the number of clusters (the Level 2 sample size) was small, but we did not address the situation when either the cluster size (Level 1 sample size) is small or sample sizes at both levels are small. Although McNeish (2014) recommended design-based methods for small Level 1 sample sizes, a similar issue emerges as in this study if sample sizes are small at both levels. That is, design-based methods are known to have poor statistical properties when the number of clusters is low (below about 50; Li & Redden, 2015; Morel, Bokossa, & Neerchal, 2003). SEM software can implement design-based methods (e.g., Type = Complex in M*plus*), but these programs do not feature any small sample corrections for these methods. Corrections are included in regression procedures within general statistical software, however (e.g., Proc Glimmix in SAS). This study does not address small sample issues at Level 1 and further studies would be needed to investigate these issues.

We also used CIs to assess whether the indirect effects were nonnull. Although highly related, in the case of indirect effects, CIs and null hypothesis testing are not necessarily equivalent in the frequentist framework (Biesanz, Falk, & Savalei, 2010; Preacher & Selig, 2012).

Conclusions

Although methodological studies on multilevel regression with few clusters have flourished recently, multilevel mediation with few clusters has been largely unaddressed in the literature despite its vast prevalence in empirical studies. Our literature review found that nearly 90% of empirical studies apply ML-SEM to data that fail to satisfy accepted sample size requirements. Our simulation showed that in some favorable contexts such as very large within-cluster samples and larger indirect effect sizes, required sample size was closer to 50 rather than 100 clusters for multilevel mediation (as opposed to previous studies that focused on ML-SEM broadly). This value is much higher than recommendations for multilevel regression models, which tend to be in the 15 to 30 range with continuous outcomes. The increased requirement for multilevel mediation is likely due to the complexity concomitant with multilevel mediation models. Nonetheless, a majority of studies in our review failed to gather samples large enough to meet suggested minimums for ML-SEM. Furthermore, most of these studies are either 1–1–1 or 2–1–1 models that can be directly fit with MLMs and do not require the added flexibility of the ML-SEM framework. Although studies have not explicitly assessed the issue of small sample size in multilevel mediation, our results largely reflected theoretical expectations—namely that MLMs (equipped with a small sample method unique to the framework) are much less demanding than ML-SEM in terms of sample size requirements.

Plainly, we recommend that researchers initially consider MLMs and move to more flexible ML-SEM only if necessary. Our review found that most studies are interested in 1–1–1 or 2–1–1 models for which ML-SEM offers few advantages for straightforward models compared to MLMs and for which ML-SEM encounters more difficulty with the sample sizes typically seen in empirical studies. Based on our review, the number of studies failing to meet the sample size requirements of ML-SEM is sizable; however, MLMs with REML and a Kenward–Roger correction are well equipped to handle these smaller samples, provided that the model can be fit in this less flexible framework. This is highly relevant for empirical researchers because 27 of the 70 reviewed studies (39%) used either a 1–1–1 or 2–1–1 mediation model with 50 or fewer clusters. If a more conservative sample size cutoff of 100 clusters is applied, this number jumps to 47 studies (67%).

To be clear, we are not disparaging Preacher et al.'s (2010) general ML-SEM framework for multilevel mediation—it is conceptually elegant, clearly of wide utility, and uniquely capable of fitting certain types of models that appear in empirical studies. It is certainly true that there are many types of models with complexity that exceeds the capacity of MLMs for which researchers require the added flexibility afforded by ML-SEM (i.e., the percentages cited in the previous paragraph are far from 100%). Rather, we are noting that empirical studies frequently test models that do not require this added flexibility because the complexity of the model does not exceed the bounds of MLMs. Often, the data fall short of the requirements necessary to apply ML-SEM and many studies would be better candidates for MLMs because MLMs can, at times, equivalently model researchers' data and can yield more trustworthy estimates due to advances in small sample statistical theory that are not available in ML-SEM.

Colloquially, researchers pay a tax for the added flexibility of ML-SEM, namely the loss of advantageous small sample estimation methods and corrective procedures that exist in MLMs. In some instances (e.g., 1–2–2 mediation, testing multiple outcomes simultaneously, positing measurement models for variables), this tax is necessary, as some models can only be properly fit via ML-SEM. With large sample sizes and moderate or simple multilevel mediation models, this tax is negligible because the MLMs and ML-SEM converge asymptotically. However, based on our review, many researchers are paying the tax for no reason because their models are not complex enough to necessitate ML-SEM, their data do not have large sample samples at Level 2, and the models can therefore be equivalently fit with MLMs. When needlessly resorting to ML-SEM, estimates can be unnecessarily deficient in the common scenario where the Level 2 sample size is small, due to full rather than restricted maximum likelihood estimation—a problem that is sometimes simple to avoid.

Methods for indirect effect CIs in the frequentist framework have been widely studied in the context of single-level models, but the appropriate method for multilevel mediation has received less attention (Preacher et al., 2010). This is especially salient in multilevel mediation because bootstrap methods that have grown popular in single-level mediation models are not as straightforward to apply in multilevel settings and methods that do exist tend to produce poor results with smaller Level 2 sample sizes (Cameron et al., 2008). In this study, we found only small differences between the delta, distribution of the product, and Monte Carlo methods. With MLMs, all methods did a reasonable job with respect to CI coverage, although the delta method was less effective at detecting nonnull effects for the MLMs and ML-SEM across conditions as has been shown previously (e.g., MacKinnon et al., 2002). CI coverage was quite poor in conditions with fewer clusters with ML-SEM, although this is attributable to the poor estimates that serve as the input for these methods and the methods themselves are not necessarily weak as applied to ML-SEM. The strong performance of these methods suggests that researchers might be better off avoiding bootstrapping in multilevel settings and the added complication that the nested data structure entails.

Although the MCMC condition performed poorly in this study, we reiterate that Bayesian methods certainly represent a viable alternative in modeling small sample data if one requires ML-SEM (e.g., the desired model is a 1–2–1 that cannot be fit with MLMs), especially given the paucity of small sample frequentist options in the ML-SEM framework. However, researchers must do their due diligence with respect to prior distributions and cannot expect the issue to be rectified merely by switching the estimator in M*plus* from ML to BAYES, for example. In fact, as shown here, such a strategy exacerbates the small sample problem and vastly reduces the ability to detect meaningful effects (which is reduced already in small sample data). A pressing question for future studies lies in how informative prior distributions must be for Bayesian methods to be a superior option. Although research in this area is still nascent, readers interested in further detail on Bayesian methods for mediation are referred to Yuan and MacKinnon (2009) or Enders, Fairchild, and MacKinnon (2013).

REFERENCES

Aroian, L. A. (1947). The probability function of the product of two normally distributed variables. *The Annals of Mathematical Statistics*, *18*, 265–271. doi:10.1214/aoms/1177730442

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1982. doi:10.1037/0022-3514.51.6.1173

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, *28*, 135–167. doi:10.3102/10769986028002135

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods*, *11*, 142–163. doi:10.1037/1082-989X.11.2.142

Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Kromrey, J. D., & Ferron, J. M. (2014). How low can you go? An investigation of the influence of sample size and model complexity on point and interval estimates in two-level linear models. *Methodology*, *10*, 1–11. doi:10.1027/1614-2241/a000062

Biesanz, J. C., Falk, C. F., & Savalei, V. (2010). Assessing mediational models: Testing and interval estimation for indirect effects. *Multivariate Behavioral Research*, *45*, 661–701. doi:10.1080/00273171.2010.498292

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144–152. doi:10.1111/bmsp.1978.31.issue-2

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*, 473–514. doi:10.1214/06-BA117

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, *90*, 414–427. doi:10.1162/rest.90.3.414

Carr, E., & Chung, H. (2014). Employment insecurity and life satisfaction: The moderating influence of labour market policies across Europe. *Journal of European Social Policy*, *24*, 383–399. doi:10.1177/0958928714538219

Cheung, M. W. L. (2013). Implementing restricted maximum likelihood estimation in structural equation models. *Structural Equation Modeling*, *20*, 157–167. doi:10.1080/10705511.2013.742404

Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *Journal of Epidemiology & Community Health*, *62*, 752–758. doi:10.1136/jech.2007.060798

Craig, C. C. (1936). On the frequency function of xy. *The Annals of Mathematical Statistics*, *7*, 1–15. doi:10.1214/aoms/1177732541

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, *38*, 529–569. doi:10.1207/s15327906mbr3804_5

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., … Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, *79*, 69–102. doi:10.3102/0034654308325581

Depaoli, S. (2014). The impact of inaccurate "informative" priors for growth parameters in Bayesian growth mixture modeling. *Structural Equation Modeling*, *21*, 239–252. doi:10.1080/10705511.2014.882686

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling*, *22*, 327–351. doi:10.1080/10705511.2014.937849

Depaoli, S., & van de Schoot, R. (2015). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods*. Advance online publication. doi:10.1037/met0000065

Enders, C. K., Fairchild, A. J., & MacKinnon, D. P. (2013). A Bayesian approach for estimating mediation effects with missing data. *Multivariate Behavioral Research*, *48*, 340–369. doi:10.1080/00273171.2013.784862

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods*, *41*, 372–384. doi:10.3758/BRM.41.2.372

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, *18*, 233–239. doi:10.1111/j.1467-9280.2007.01882.x

Fritz, M. S., Taylor, A. B., & MacKinnon, D. P. (2012). Explanation of two anomalous results in statistical mediation analysis. *Multivariate Behavioral Research*, *47*, 61–87. doi:10.1080/00273171.2012.640596

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*, 60–87. doi:10.3102/0162373707299706

Hox, J. J., & Maas, C. J. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples.

Structural Equation Modeling, *8*, 157–174. doi:10.1207/S15328007SEM0802_1

Hox, J. J., Maas, C. J., & Brinkhuis, M. J. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, *64*, 157–170. doi:10.1111/j.1467-9574.2009.00445.x

Hox, J. J., van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods*, *6*, 87–93.

Kish, L. (1965). *Survey sampling*. New York, NY: Wiley.

Kreft, I. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies* (Working paper). Los Angeles, CA: California State University.

Krull, J. L., & MacKinnon, D. P. (1999). Multilevel mediation modeling in group-based intervention studies. *Evaluation Review*, *23*, 418–444. doi:10.1177/0193841X9902300404

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research*, *36*, 249–277. doi:10.1207/S15327906MBR3602_06

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, *15*, 722–752. doi:10.1177/1094428112457829

Li, P., & Redden, D. T. (2015). Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Statistics in Medicine*, *34*, 281–296. doi:10.1002/sim.v34.2

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*, 86–92. doi:10.1027/1614-2241.1.3.86

MacKinnon, D. P., Fritz, M. S., Williams, J., & Lockwood, C. M. (2007). Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. *Behavior Research Methods*, *39*, 384–389. doi:10.3758/BF03193007

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*, 83–104. doi:10.1037/1082-989X.7.1.83

McNeish, D. (2014). Modeling sparsely clustered data: Design-based, model-based, and single-level methods. *Psychological Methods*, *19*, 552–563. doi:10.1037/met0000024

McNeish, D. (2016a). Using data-dependent priors to mitigate small sample bias in latent growth models. *Journal of Educational and Behavioral Statistics*, *41*, 27–56. doi:10.3102/1076998615621299

McNeish, D. (2016b). On using Bayesian methods to address small sample problems. *Structural Equation Modeling*, *23*, 750–773. doi:10.1080/10705511.2016.1186549

McNeish, D., & Stapleton, L. M. (2016). The effect of small sample size on two level model estimates: A review and illustration. *Educational Psychology Review*, *28*, 295–314. doi:10.1007/s10648-014-9287-x

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*, 259–284. doi:10.1037/1082-989X.10.3.259

Miranda, R., MacKillop, J., Treloar, H., Blanchard, A., Tidey, J. W., Swift, R. M., … Monti, P. M. (2016). Biobehavioral mechanisms of topiramate's effects on alcohol use: An investigation pairing laboratory and ecological momentary assessments. *Addiction Biology*, *21*, 171–182. doi:10.1111/adb.12192

Morel, J. G., Bokossa, M. C., & Neerchal, N. K. (2003). Small sample correction for the variance of GEE estimators. *Biometrical Journal*, *45*, 395–409. doi:10.1002/bimj.200390021

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585. doi:10.1007/BF02296397

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*, 338–354. doi:10.1111/j.1745-3984.1991.tb00363.x

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

Muthén, B. O., & Asparouhov, T. (2008). Growth mixture modeling: Analysis with non-Gaussian random effects. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 143–165). Boca Raton, FL: Chapman & Hall/CRC.

Muthen, L.K., & Muthen, B,O. (2010) Growth modeling with latent variables using Mplus: Introductory and intermediate growth models. Retrieved from http://www.statmodel.com/download/Topic3-v.pdf.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267–316. doi:10.2307/271070

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide: Statistical analysis with latent variables* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Pituch, K. A., & Stapleton, L. M. (2011). Hierarchical linear and structural equation modeling approaches to mediation analysis in randomized field experiments. In M. Williams & P. Vogt (Eds.), *The Sage handbook of innovation in social research methods* (pp. 590–619). Thousand Oaks, CA: Sage.

Pituch, K. A., Stapleton, L. M., & Kang, J. Y. (2006). A comparison of single sample and bootstrap methods to assess mediation in cluster randomized trials. *Multivariate Behavioral Research*, *41*, 367–400. doi:10.1207/s15327906mbr4103_5

Pituch, K. A., Whittaker, T. A., & Stapleton, L. M. (2005). A comparison of methods to test for mediation in multisite experiments. *Multivariate Behavioral Research*, *40*, 1–23. doi:10.1207/s15327906mbr4001_1

Polansky, A. M. (1999). Upper bounds on the true coverage probability of bootstrap percentile type confidence intervals. *The American Statistician*, *53*, 362–369.

Preacher, K. J. (2011). Multilevel SEM strategies for evaluating mediation in three-level data. *Multivariate Behavioral Research*, *46*, 691–731. doi:10.1080/00273171.2011.589280

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures*, *6*, 77–98. doi:10.1080/19312458.2012.679848

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

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*, 209–233. doi:10.1037/a0020141

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

Snijders, T. A., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational and Behavioral Statistics*, *18*, 237–259. doi:10.3102/10769986018003237

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, *13*, 290–312. doi:10.2307/270723

Springer, M. D., & Thompson, W. E. (1966). The distribution of products of independent random variables. *SIAM Journal on Applied Mathematics*, *14*, 511–526. doi:10.1137/0114046

Tofighi, D., & MacKinnon, D. P. (2011). RMediation: An R package for mediation analysis confidence intervals. *Behavior Research Methods*, *43*, 692–700. doi:10.3758/s13428-011-0076-x

van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., & van Loey, N. E. (2015). Analyzing small data sets using Bayesian estimation: The case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology*, 6.

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Aken, M. A. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development*, *85*, 842–860. doi:10.1111/cdev.12169

van de Schoot, R., Winter, S., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian papers in

psychology: The last 25 years. *Psychological Methods*. Advance online publication. doi:10.1037/met0000100

van der Leeden, R., Meijer, E., & Busing, F. M. T. A. (2008). Resampling multilevel models. In J. De Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 401–433). New York, NY: Springer.

van Noorden, R., Maher, B., & Nuzzo, R. (2014). The top 100 papers. *Nature*, *514*, 550–553. doi:10.1038/514550a

Wang, J., Carpenter, J. R., & Kepler, M. A. (2006). Using SAS to conduct nonparametric residual bootstrap multilevel modeling with a small number of groups. *Computer Methods and Programs in Biomedicine*, *82*, 130–143. doi:10.1016/j.cmpb.2006.02.006

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, *14*, 301–322. doi:10.1037/a0016972

Zhang, Z., Zyphur, M. J., & Preacher, K. J. (2009). Testing multilevel mediation using hierarchical linear models: Problems and solutions. *Organizational Research Methods*, *12*, 695–719. doi:10.1177/1094428108327450

Zyphur, M. J., & Oswald, F. L. (2015). Bayesian estimation and inference: A user's guide. *Journal of Management*, *41*, 390–420. doi:10.1177/0149206313501200

APPENDIX

Summary Information for Random Sample of Review Studies Citing Preacher, Zyhpur, and Zhang (2010)

First Author	*Year*	*Sample*	*Bayes*	*<25*	*<50*	*< 100*	*Primary Model*	*Second Model*	*Third Model*	*Repeated Measures*	*Level 2 Unit*
Allen	2016	78	0	0	0	1	211			0	Schools
Bai	2015	43	0	0	1	1	211			0	Firms
Berg	2016	389	0	0	0	0	211			0	Schools
Carr	2014	22	1	1	1	1	211			0	Countries
Chen	2016	42	0	0	1	1	211			0	Companies
Chun	2015	23	0	1	1	1	222			0	Teams
de Clerq	2015	32	0	0	1	1	211	111	221	0	Employers
de Cock	2016	20	0	1	1	1	111			0	Schools
Delhey	2014	30	0	0	1	1	211			0	Countries
Demblon	2016	32	0	0	1	1	111			1	
Dierdorff	2013	230	0	0	0	0	211			0	Occupations
Donati	2016	28	0	0	1	1	121			0	Teams
Elorza	2016	51	0	0	0	1	211			0	Managers
Ewen	2013	190	0	0	0	0	211			0	Team leader
Flinchbaugh	2016	25	0	1	1	1	212			0	Teams
Friedrich	2015	73	0	0	0	1	111			0	Teachers
Graham	2014	11	0	1	1	1	211			0	Schools
Green	2010	26	0	0	1	1	112			0	Swiss cantons
Gregory	2014	12	0	1	1	1	211			0	Schools
Greijdanus	2015	27	0	0	1	1	211	221		0	Discussion groups
Hesser	2014	67	0	0	0	1	111			1	
Holman	2016	49	0	0	1	1	111			1	
Huang	2015	84	0	0	0	1	111			1	
Kang	2015	39	0	0	1	1	121			0	CEOs
Knowles	2015	14	0	1	1	1	111			0	Countries
Koch	2015	83	0	0	0	1	211			0	Schools
Koopman	2016	82	0	0	0	1	111			1	
Kranzler	2014	122	0	0	0	0	211			1	
Lam	2016	30	0	0	1	1	111			0	Managers
Lehmann-Willenbrock	2015	30	0	0	1	1	221			0	Teams
lepine	2016	74	0	0	0	1	212			0	Leader
Li	2015	81	0	0	0	1	111			0	Teams
Liao	2015	126	0	0	0	0	221			0	Teams
Liden	2014	71	0	0	0	1	211			0	Employers
Lin	2016	98	0	0	0	1	111			0	Managers
Lu	2015	104	0	0	0	0	122			0	Schools
Massenberg	2015	34	0	0	1	1	211	221		0	Teams
Meleady	2013	12	0	1	1	1	211			0	Experimental groups
Miranda	2016	96	1	0	0	1	211			1	
Mok	2014	26	0	0	1	1	111			0	Schools
Narine	2013	45	0	0	1	1	221	111		0	Neighborhoods
Nohe	2013	33	0	0	1	1	111			1	
Paustian-UnderDahl	2014	47	0	0	1	1	211			0	Health care units
Peter (study1)	2012	67	0	0	0	1	111			0	Teachers

(*Continued*)

(Continued)

First Author	*Year*	*Sample*	*Bayes*	*<25*	*<50*	*< 100*	*Primary Model*	*Second Model*	*Third Model*	*Repeated Measures*	*Level 2 Unit*
Peter (study2)	2012	48	0	0	1	1	111			0	Teachers
Petitta	2015	38	0	0	1	1	122			0	Sports teams
Prati	2013	10	0	1	1	1	111			0	Health organizations
Prati	2012	6	0	1	1	1	121			0	Schools
Reizer	2012	133	0	0	0	0	111			0	Couples
Reyes	2012	63	0	0	0	1	211			0	Teachers
Rosen	2016	107	0	0	0	0	111			1	
Ruzek	2016	68	0	0	0	1	111			0	Teachers
Saarento	2015	77	0	0	0	1	211			0	Schools
Schulte	2015	54	0	0	0	1	111			0	Teams
Shen	2014	35	0	0	1	1	111			0	Firms
Shen	2016	30	0	0	1	1	211			0	Firms
Shi	2013	53	0	0	0	1	111			0	Teams
Super	2016	60	0	0	0	1	221			0	Experimental groups
Terry-McElrath	2015	42	0	0	1	1	221	111		0	Schools
Valdimarsdottir	2015	83	0	0	0	1	211			0	Schools
Vauclair	2015	28	0	0	1	1	211			0	Countries
Vermeeren	2014	41	0	0	1	1	212			0	Managers
Way	2014	61	0	0	0	1	221			0	Leader
Weisman	2016	66	0	0	0	1	211			0	Caregiver
Wijnia	2016	22	0	1	1	1	212	112		0	Teams
Wilson	2012	23	0	1	1	1	111			0	Teachers
Wilt	2016	40	0	0	1	1	111			1	
Wohrmann	2013	32	0	0	1	1	111			0	Jobs
Zhao	2016	22	0	1	1	1	111			0	Teams
Zhu	2016	89	0	0	0	1	211			0	Teams