HHS Public Access

Author manuscript

Prev Sci. Author manuscript; available in PMC 2016 October 01.

Published in final edited form as:

Prev Sci. 2015 October ; 16(7): 987–996. doi:10.1007/s11121-014-0489-8.

Sample Size Considerations in Prevention Research Applications of Multilevel Modeling and Structural Equation Modeling

Rick H. Hoyle and **Nisha C. Gottfredson**

Duke University

Abstract

When the goal of prevention research is to capture in statistical models some measure of the dynamic complexity in structures and processes implicated in problem behavior and its prevention, approaches such as multilevel modeling (MLM) and structural equation modeling (SEM) are indicated. Yet the assumptions that must be satisfied if these approaches are to be used responsibly raise concerns regarding their use in prevention research involving smaller samples. In this manuscript we discuss in nontechnical terms the role of sample size in MLM and SEM and present findings from the latest simulation work on the performance of each approach at sample sizes typical of prevention research. For each statistical approach, we draw from extant simulation studies to establish lower bounds for sample size (e.g., MLM can be applied with as few as 10 groups comprising 10 members with normally distributed data, restricted maximum likelihood estimation, and a focus on fixed effects; sample sizes as small as $N = 50$ can produce reliable SEM results with normally distributed data and at least three reliable indicators per factor) and suggest strategies for making the best use of the modeling approach when N is near the lower bound.

Keywords

sample size; multilevel modeling; structural equation modeling

This manuscript focuses on sample size considerations in applications of two statistical methods of particular relevance to prevention research questions—multilevel modeling (MLM) and structural equation modeling (SEM). MLM and SEM are used to fullest advantage when the goal is to model a structure or process as opposed to isolated tests of individual parameters such as correlation coefficients or mean differences. Models of structure focus on the nature of the relations between variables that define a complex construct (e.g., impulsivity, Whiteside & Lynam, 2001) or domain (e.g., problem behavior, Gillmore, Hawkins, Catalano, Day, & Abbott 1991). Models of process focus on causal relations and the mechanisms that account for them as specified by theoretical models (e.g., Shiyko, Lanza, Tan, Li, & Shiffman, 2012). The value of such models to the work of prevention scientists is clear. The suitability of data from small-sample prevention studies for evaluating such models is not always clear.

Correspondence concerning this manuscript should be addressed to Rick Hoyle, Department of Psychology and Neuroscience, Duke University, Durham, NC 27708. Contact: rhoyle@duke.edu.

Our central concern in this manuscript is the use of MLM or SEM with small samples. There is no absolute definition of "small" in the area of statistical analysis; thus, a primary issue is what is considered a small sample when using these statistical methods. Elsewhere, we have defined small samples as "samples that are near the lower bound of the size required for satisfactory performance of the particular statistical model chosen to address the questions that motivated the research" (Hopkin, Hoyle, & Gottfredson, 2013). With regard to MLM and SEM, the question of whether performance is satisfactory concerns multiple features of the analysis. The most basic is whether the model can be estimated at all (i.e., nonconvergence, inadmissible solutions). If the model can be estimated, attention turns to evaluations of the degree to which the model accounts for the data (i.e., fit), and estimates and tests of parameters in the model.

Beginning with MLM, we integrate information from a thorough review of existing simulation studies to touch on each of these concerns for these promising approaches to modeling prevention data. The reason for drawing from simulation studies to provide guidance on the use of MLM and SEM in small samples is the following: Inferences drawn from statistical models depend on a number of assumptions that are likely to be violated to some degree in real data. For instance, the validity of calculated *p*-values depends on the existence of large sample size. Simulations provide a method for assessing the practical impact of violating assumptions to varying degrees by giving researchers experimental control over features of the data such as sample size. We synthesis results and recommendations from the relatively small census of published studies that have used simulation methodology to evaluate MLM and SEM performance as a function of sample size.

The performance of estimators and tests in modeling frameworks such as MLM and SEM typically are evaluated by simulation studies (Bandalos & Gagné, 2012). These studies begin with one or more models for which population values of the parameters are set by the investigators. Many samples are drawn from the population(s); these data sets reflect characteristics on which the performance of estimators or fit statistics are to be evaluated. For example, a simulation study focused on the performance of an estimation method at different sample sizes and degrees of nonnormality might simulate 200 data sets for all combinations of three sample sizes (100, 200, and 400) and three levels of nonnormality (none, moderate, severe). Parameters would be estimated and fit statistics generated for the population model for each of the 1800 data sets. Means for each of the conditions would then be compared in order to determine the effects of sample size, nonnormality, and their interaction.

Multilevel Modeling

Multilevel data might occur by design as a result of a multi-stage sampling technique or as a result of a repeated measures design.[1] Data such as these should be modeled using a

[1]Nested data may also emerge from analyses aimed at accounting for unobserved heterogeneity in outcomes such as in latent class analysis or growth mixture models with longitudinal data. Relatively little is known about sample size requirements for analyses of these types but they almost certainly require samples that are large. For that reason, those models fall outside the scope of this manuscript.

statistical technique that accommodates non-independent observations.[2] Failure to account for non-independence of observations leads to incorrect standard error estimates and a Type I error rate that is either too conservative or too liberal (Raudenbush & Bryk, 2002). Because multilevel models (MLM; Breslow & Clayton, 1993; Goldstein, 1986; Raudenbush & Bryk, 2002) allow researchers to separate contextual effects from intra-individual effects, this is often the preferred technique for modeling nested data (e.g., Aitkin & Longford, 1986).

In this section, we begin with an overview of our notation, which is primarily adapted from Raudenbush and Bryk (2002). After describing a standard two-level model, we briefly discuss issues of sample size that are unique to MLM as they relate to study design. We then move to issues involved in data analysis, describing simulation research that sheds light on performance with small samples. We conclude with practical suggestions for improving power and reducing bias with a small sample of multilevel data.

Notation and Model Overview

Let y_{ij} be an outcome that contains variance that can be decomposed into two levels: the Level 1 or *within* portion of the variance (*i*) and the Level 2 or *between* portion of the variance (*j*). For example, y_{ij} might be a measure of alcohol involvement that varies across individuals, but that also varies across neighborhoods within a study (e.g., Duncan, Duncan, & Strycker, 2002). Between-group (e.g., between-neighborhood) variance in y_{ij} is only accounted for by predictors that vary across these independent sampling units (e.g., neighborhood crime rates). Predictors that are measured at Level 1 may contain variance at both the within and between levels (Bollen & Curran, 2006; Kreft, de Leeuw, & Aiken, 1995). Thus, these predictors may explain both group-level variation and within-group variation. Continuing with our example, Level 1 predictors might include gender or having an alcoholic parent. Both of these variables reflect information about the neighborhood (proportion male and alcoholism rates) as well as information about individuals independent from the neighborhood context (e.g., having an alcoholic parent in a neighborhood with lower rates of alcoholism is different from having an alcoholic parent in a neighborhood in which alcoholism is normative).

A generalized two-level MLM with two additive predictors, one varying within group (x_{ij}), and one varying between group (x_j), can be written as follows:

$$y^*_{ij}=\beta_0+\beta_{01}\cdot x_j+u_{0j}+\beta_{10}\cdot x_{ij}+u_{1j}\cdot x_{ij}+r_{ij}. \quad (1)$$

Here, y^*_{ij} is used in the place of y_{ij} to add flexibility to the model so the outcome variable is not restricted to a normal distribution (Breslow & Clayton, 1993; Raudenbush & Bryk, 2002). In the generalized model, y^*_{ij} is associated with y_{ij}, the outcome of interest, via a link function whose form depends on the distribution of y_{ij}. If y_{ij} is normally distributed, then it is assumed that the residual term r_{ij} is also normally distributed with a mean of zero and variance σ^2. In this case, y_{ij} is related to y^*_{ij} via the identity link. If y_{ij} is not normally distributed, then a variety of link functions may be used, and the assumed distribution of r_{ij} changes accordingly. For instance, if y_{ij} is dichotomous, then y_{ij} may be linked with y^*_{ij} via

[2] A number of software packages can handle such analyses, including (but certainly not limited to): HLM, M*plus*, R, SAS, and Stata

a logit or probit link function, and is r_{ij} is assumed to be distributed as logistic (0, $\pi^2/3$) or normal (0,1), respectively (Bauer & Sterba, 2011). If y_{ij} is ordinal, then each category thresholds is typically modeled using a cumulative logit or cumulative probit function.

In Equation (1), β coefficients represent fixed effects, or the effect of a predictor on y^*_{ij} for an average independent sampling unit. That is, fixed effects are the expected predictor effects when the random effects (*u*), which represent systematic variation of independent sampling units around the average, are equal to zero. In the alcohol example, the fixed effect of having an alcoholic parent represents the expected consequence of having a parent with a history of alcoholism on a child's problematic alcohol involvement if it were possible to measure the counterfactual *within* an individual. By contrast, random effects represent the degree to which independent sampling units deviate from the average. For instance, some neighborhoods are characterized by higher rates of problematic alcohol involvement than others (a random intercept), and the effect of having an alcoholic parent might be worse in some neighborhoods than in others (a random effect of parent alcoholism).

Multilevel data are unique in that they involve two distinct sample sizes: the number of independent sampling units (i.e., groups), and the number of secondary sampling units. We call the number of independent sampling units *N*. Because the number of Level 1 units may vary over groups, we will refer to the average number of secondary units per group as $\bar{n}$.

As a general rule, a researcher concerned about power should focus on maximizing *N* to the extent possible because independent sampling units are, by definition, uncorrelated with one another and thus provide more total information than secondary sampling units which are, by definition, correlated with one another. Researchers wishing to draw inferences about contextual or group effects, and particularly about variation in group effects, should be especially concerned with maximizing *N* (Raudenbush & Liu, 2000).

Even though it is important to sample as many Level 2 units as possible, there are many reasons to maximize $\bar{n}$ as well. First, many research questions focus on within-group processes. Longitudinal research designs are an example of this: in longitudinal designs, Level 1 units represent time and Level 2 units represent people. For a researcher wishing to draw inferences about longitudinal processes that occur within individuals (that is, to make claims about development rather than about age/cohort effects), is it essential to have enough over-time information (i.e., a relatively large *n*). Second, some research questions about between-group processes rely on aggregate within-group information for proxy measures of inter-group differences (Snijders & Bosker, 2004). For the latter type of analysis, the Level 1 sample size is important for reliably estimating group-level measures (Lüdtke, Marsh, Robitzsch, Trautwein, Asparouhov, & Muthén, 2008). Third, and most practically, it may be more cost effective to sample Level 1 units than to sample Level 2 units. Raudenbush (1997) presented a sample equation that might be used to optimize the sample sizes at each level given the following information: a) the total monetary resources available; b) the cost of sampling a unit at Level 1; and c) the cost of sampling a unit at Level 2. Raudenbush, Spybrook, Congdon, et al. (2011) provide freely downloadable software for researchers to use when designing longitudinal or cluster-based studies.

Multilevel data differ with respect to the degree to which within-group sampling units correlate with one another. For example, we might expect that children who are nested within the same family will be more correlated with one another than children who are nested within a classroom. Less dependency among Level 1 units is associated with a higher payoff from sampling them (Raudenbush, 1997). The degree of dependency can by computed by calculating the intra-class correlation (ICC; see Raudenbush & Bryk, 2002, or Snijders & Bosker, 2004, for formulae). These values range from zero (no dependency) to 1.0 (total dependency). Because statistical power in MLM partially depends on the ICC, it can be useful to estimate the expected ICC during the study design, relying on extant studies similar to the planned study.

Estimation Considerations

MLM estimation entails a number of important decision points for the data analyst. Whereas estimation is fairly straightforward when y_{ij} is normally distributed (Demidenko, 2004), estimates are less clear-cut when y_{ij} is non-normal. We first describe estimation with a normally distributed outcome variable and then move to the more complex scenario.

If y_{ij} is normally distributed then the only estimation decision involves whether to rely on full maximum likelihood (FML; Anderson, 1957) or restricted maximum likelihood (REML; Dempster, Laird, & Rubin, 1977).[3] FML was designed to be unbiased when sample size is large; this method inherently results in downwardly biased estimates of random effect variances and in confidence intervals that are too narrow around the fixed effect estimates when the sample is small (Raudenbush & Bryk, 2002; Singer & Willet, 2003). REML was designed to correct for this bias and thus it is a natural choice when small samples are a concern. There are two caveats to this conclusion. First, Kreft and deLeeuw (1998) illustrated a trade-off between bias and precision with small samples: FML estimates are downwardly biased but more precise; REML estimates are unbiased but less precise. Second, FML is the only approach that can be used for constructing likelihood ratio tests to compare nested models. An analyst wishing to compare models should rely on alternative model comparison techniques if REML is used for model estimation (e.g., Bayesian Information Criterion, Schwarz, 1978).

Maas and Hox (2005) presented a simulation study testing REML performance as a function of sample size. They tested bias, efficiency, and coverage of fixed effects and random effect variances using with 30, 50, or 100 groups.[4] They found that REML estimates were always unbiased but that standard error estimates for variance components were downwardly biased when 30 groups were present. Given these findings, an analyst with a normally distributed outcome variable should use REML estimation and trust that all point estimates are unbiased and that inference around fixed effects is correct. However, inference about random effects should be approached with caution.

[3]We omit a discussion of least squares approaches because these tend to be less efficient than maximum likelihood (Singer & Willet, 2003).

[4]We recognized that 30 groups is unrealistic for many prevention studies. This simulation study did not consider fewer than 30 groups given that the results were already somewhat problematic for that number. Findings from work considering 10 Level 2 units is presented below.

If y_{ij} is not normally distributed, then the maximum likelihood solution cannot be found analytically because it is not in closed form, and so it must be approximated in one of two ways. Either the likelihood function itself must be approximated (e.g., by using a Taylor series to approximation to the likelihood function that can itself be maximized directly; Rodriguez & Goldman, 2001), or the maximum of the likelihood function must be found computationally (e.g., Rabe-Hesketh, Skrondal, & Pickles, 2002). The former method is referred to as *Quasi-Maximum Likelihood*. Computationally intensive estimation techniques can be implemented using SAS Proc NLMIXED, SAS Proc GLIMMIX, xtlogit or gllamm in STATA, or the glmer routine in the lme4 package in R (Austin, 2010). It is generally accepted that the true FML estimates found using this exact technique are superior to the alternative when y_{ij} is binary or when it is ordinal with relatively few response categories. Both techniques perform well when y_{ij} is Poisson distributed or when it is ordinal with many response categories (Rabe-Hesketh et al., 2002). However, as is the case with straightforward linear mixed models (i.e., with normally distributed y_{ij}), estimates generated using FML are biased in small samples, particularly as the number of parameters increase (Bauer & Sterba, 2011; Raudenbush & Bryk, 2002). REML corrections for generalized linear mixed models have not yet been perfected or implemented in standard MLM software packages.[5]

Modern software programs that rely on approximation estimations, such as HLM (Raudenbush, Bryk, & Congdon, 2004), SAS Proc GLIMMIX, and the glmmPQL function in R use the widely accepted approach of penalized quasi-likelihood (PQL) or a variant thereof. An advantage of PQL estimation relative to true FML is that it enables the implementation of a REML correction to the quasi-likelihood function (Schabenberger, 2005).

Bauer and Sterba (2011) challenged the widely held belief that computationally intensive estimation techniques are superior to the approximation technique. In a simulation study with ordered categorical items, Bauer and Sterba found that PQL estimates were more biased but were also much more efficient than FML estimates when the number of groups was small (i.e., between 25 and 50 groups), particularly if the number of response categories was also small.[6]

In order to provide practical advice to applied researchers with non-normal outcomes, Austin (2010) compared estimator performance with a binary outcome and relatively small samples using six software packages that implement generalized MLM. All of the packages using the estimators described above recovered fixed effects well when there were at least 20 Level 1 units per group. All except for the PQL estimators recovered fixed effects well with more than 10 groups. Mirroring the findings of Bauer and Sterba (2011), Austin found that confidence interval coverage for fixed effects was accurate with quasi-likelihood estimators used by Proc GLIMMIX and in HLM (even though the point estimates were

[5]However, progress is being made on this front. See Noh & Lee (2007) for an example.

[6]Bauer and Sterba (2011) also found that increasing the number of response categories resulted in less bias and greater efficiency. This result is not surprising given the well-known consequences of dichotomization (MacCallum, Zhang, Preacher, & Rucker, 2002). Whenever possible, researcher should maximize the number of response categories or use a continuous response scale to maximize power.

biased). FML methods used in xtlogit and glmer also resulted in correct confidence interval coverage. For variance component point estimates, Austin found that, as expected, most of the software programs had difficulty estimating these parameters with fewer than 10 units per group. The quasi-likelihood estimators in Proc GLIMMIX, the adaptive quadrature technique in Proc NLMIXED, the Gauss Hermite estimator in xtlogit, and the adaptive quadrature estimator glmer performed reasonably well as the number of units per group increased. HLM performed well only as the number of groups reached 15 or higher. Random effect confidence intervals were not assessed. Austin concluded that it is generally safe to rely on generalized MLM estimates with at least 10 groups, or with fewer than 10 groups as long as there are at least 30 within-group units. His findings also suggest that it is not safe to rely on estimates if there are 5 or fewer units per group.

Summary of MLM estimation

When y_{ij} is normally distributed, REML is the preferred estimator and inferences about fixed effects can be trusted. Inferences about random effects can be trusted only if the number of groups is moderate.

When y_{ij} is not normally distributed, there is no clear-cut choice between using computationally intensive FML methods and using quasi-likelihood approximations. Austin's 2010) simulation work suggests that the quasi-likelihood estimator used by SAS Proc GLIMMIX is a good choice, and that the FML estimators used in STATA xtlogit, and R glmer are best for small samples. Whereas inference about fixed effects can generally be trusted with 10 or more groups or with fewer than 10 groups and 30 or more Level 1 units (e.g., Simon et al., 2008), inference about random effects is unstable with small samples and non-normal outcomes. Even when y_{ij} is not normally distributed, meaningful variation in y_{ij} should be maximize through the use of as many response categories as are reasonable (Bauer & Sterba, 2011).

Rules of Thumb

A number of rules of thumb about sample size with MLM are available in the literature. Many of these are based on sound simulation or analytical work. Although we present some findings and rough guidelines from the literature, we do not advocate strict adherence to rules of thumb. Multilevel models are complex and each study has a unique combination of features that influence inferential ability. Instead of relying on rules of thumb, we encourage analysts with small samples to do the following: 1) consider conducting a power analysis that incorporates information about data features for your study (Optimal Design software is a good place to start; Raudenbush, et al., 2011); 2) maximize inferential precision through study design and statistical analysis; and 3) consider which point estimates and standard errors are likely to be biased or unbiased and limit inference to those estimates that you trust.

Structural Equation Modeling

Structural equation modeling (SEM) is a very general statistical approach for modeling multivariate data. Evidence of its generality is the observation that any of the analyses discussed in the previous section could be accomplished using SEM. For instance, latent

curve modeling (Bollen & Curran, 2006; McArdle & Nesselroade, 2003; Meredith & Tisak, 1990) permits growth modeling of longitudinal data using SEM that can produce equivalent results to MLM (Bauer, 2003; Curran, 2003), but that also provides the analyst with more flexibility in modeling choices (e.g., Curran, Lee, Howard, Lane, & MacCallum, 2012). Importantly, SEM can also be used to model certain types of relations that cannot be modeled using MLM or other methods commonly used by prevention scientists. Perhaps most prominent among these is the relations between observed variables and the underlying, or latent, construct they were intended to measure. Somewhat like embedding a factor analysis in a multiple regression model, relations can be modeled between constructs rather than variables. In addition, SEM is useful when the model prescribes multiple dependent variables that are themselves directionally related to each other. A simple example is the three variable mediation model, in which an independent variable influences both a mediator and an outcome, which are related to each other through a directional path from the mediator to the outcome. The ability to model latent variables, the relations between them, and directional relations between dependent variables makes SEM an attractive analytic option for many prevention research questions and designs.[7]

Yet, the generality and flexibility of SEM come at a price. The estimators typically used to derive parameter estimates, standard errors, and model fit statistics are asymptotic in nature; that is, they are unbiased and efficient when sample size is large (Bollen & Noble, 2011). Given the constraints on sample size that are typical of the behavioral and health sciences, the question of how large a sample is necessary for valid estimation and testing has received considerable attention (e.g., Tanaka, 1987). Addressing the question is complicated by the fact that the minimum sample size varies as a function of a number of data and model features.

In this section, we summarize a rich and growing literature on sample size considerations in SEM. We begin with an overview of estimation and testing in SEM. We then review the research to date on the performance of estimators and test statistics at different sample sizes.

Overview of SEM

SEM analyses concern the correspondence between observed data and the data implied by one or more models, which typically reflect a set of logic- or theory-based relations between variables (for overviews, see Hoyle, 2011; Weston & Gore, 2006). The values of free parameters (e.g., factor loadings, regression weights, error variances) are estimated from the observed data, after which the estimated and fixed parameters can be used to generate a theoretical matrix. This implied covariance matrix contains the data we would expect to observe were the specified model correct in the population. A comparison of the observed and implied covariance matrices is the basis for a goodness of fit test, reflected in the null hypothesis,

$$H_0: \sum = \sum \left(\hat{\theta}\right), \quad (2)$$

[7]Software packages that can estimate SEMs include, but are not limited to: EQS, LISREL, M*plus*, R, and STATA.

where Σ corresponds to the population data and $\Sigma(\hat{\theta})$ to the population data implied by the model. Although the test of this multivariate hypotheses seems straightforward—consult a reference distribution given the model degrees of freedom and interpret *p* values greater than .05 as indicative of fit—it is fraught with problems ranging from questions about the reference distribution to concerns about a null hypothesis of exact fit. As a result, a large number of indices have been developed to index goodness (or badness) of fit, with performance when sample size is small varying from one to the next

Sample Size Considerations in SEM

Although statistical power is a significant concern for SEM analyses, other concerns related to sample size are also important. Perhaps the most basic of these concerns is the degree to which the observed covariance matrix, **S**, reflects the population covariance matrix, Σ. Strategies that ensure representativeness and retention of all participants sampled are basic concerns that apply regardless of sample size. Yet, assuming these concerns are adequately addressed, the likelihood of a departure of **S** from Σ increases with smaller sample sizes. To the extent that **S** is not representative of Σ, a model that offers a satisfactory account of the data in one study might not do so for data from a different sample from the same population. Put differently, as *N* gets smaller, the confidence interval around the observed covariances gets larger. The more observed covariances to be estimated the greater this concern, leading to rules of thumb based on the ratio of participants to variables—10:1 is a commonly proposed ratio (Tanaka, 1987). This logic suggests that, with smaller sample sizes, the number of observed variables should be small. In short, a fundamental consideration is whether the observed covariances are a valid reflection of the covariances in the population so that the target of fit reflects the assumption evident in the null hypothesis (Eq. 2).

As noted earlier, estimation in SEM analyses yields parameter estimates, standard errors, and test statistics that have asymptotic properties. That is to say, their values do not depend directly on sample size as do, for example, the components of the *F* and *t* statistics used in general linear modeling analyses. Instead they assume a sample that is sufficiently large to ensure the theoretical properties of the estimates and tests. Related to this concern is the influence of sample size on estimation. Estimators such as maximum likelihood, the most widely used method in applications of SEM, are iterative. They begin the search for parameter estimates that minimize the difference between the observed and implied data with a set of starting values. These are updated after each iteration until it is no longer possible to improve the quality of the parameter estimates. At this point, the estimation is said to have converged. As discussed later in this section, small sample data are associated with nonconvergence. In such cases, the parameter estimates and/or standard errors cannot be interpreted. The likelihood of nonconvergence when *N* is small is increased by nonnormal data and misspecified models.

Beyond these sample-size concerns specific to estimation in SEM analyses is the typical concern regarding statistical power. The challenges associated with power analysis in SEM are numerous. First is the distinction between overall, or omnibus, fit and the significance of specific parameter estimates. Focusing first on tests of individual parameters, there is the problem that parameter estimates in models are interdependent—the magnitude of each is, to

some degree, contingent on the magnitude of the others (Kaplan, 1995). Thus, the evaluation of statistical power for a given parameter must account for other parameters in the model. The challenges are greater still for evaluations of omnibus fit. Returning to the null hypothesis discussed earlier, because the goal is to not reject the null hypothesis, the investigator would appear to benefit from low power. Of course, the problem with this logic is that low power may lead to the equivalent of a Type I error by failing to detect meaningful differences between the observed and implied data. An additional problem is that, because the null hypothesis specifies a perfect match between the observed and implied data, it is always the case that, with a sufficiently large *N*, this hypothesis would be rejected, resulting in the equivalent of a Type II error. As noted earlier, these drawbacks to the straightforward goodness-of-fit test led to the development of a number of alternative indices for judging omnibus fit. Sample size also is a concern when using these indices, affecting their performance in direct and indirect ways (Bollen, 1990).

Estimation problems associated with small *N*s—As noted earlier, the estimators typically used in SEM analyses are iterative, updating parameter estimates after each iteration until the difference between the observed and implied data is at its minimum given the model. On occasion, iteration is unable to reach a minimum, resulting in nonconvergence and a set of parameter estimates and tests that cannot be interpreted. Convergence does not always guarantee an interpretable solution, as estimation sometimes yields out of range values for parameters (e.g., variances less than zero) or implausible values for standard errors. Each of these undesirable outcomes of estimation is more likely with data from small samples. For example, in a simulation study of the effects of sample size, unreliability, and specification strategy (composites vs. latent variables) on models of simple mediation, Hoyle and Kenny (1999) found that 14% of solutions were problematic when *N* was very small (25 or 50) and reliability was low (α = .60). With a minimum *N* of 100 and moderate reliability (α = .75), problematic solutions were very rare. Marsh, Hau, Balla, and Grayson (1998) showed that sample size and number of indicators per factor could each compensate for small size of the other, leading to the surprising conclusion that, when *N* is small, more, not fewer, indicators are to be preferred. Focusing specifically on the smallest size they considered, *N* = 50, the percentages of proper solutions were 14, 55, 87, and 100 for 2, 3, 4, and 6 indicators, respectively. Improper solutions were very rare at *N* = 100 with four or more indicators per factor. In short, at sample sizes under 100, nonconvergence and improper values are frequent occurrences. For these small sample sizes, more highly reliable indicators can improve, but not eliminate, the likelihood of estimation problems (Gagné & Hancock, 2006; Jackson, Voth, & Frey, 2013).

Statistical power and sample size—If estimation results in a proper solution, the concern shifts to evaluation of fit. As noted earlier, the evaluation of fit, though straightforward in a conceptual sense, is quite complex in a technical sense. The chi-square test of the null hypotheses presented earlier does not perform well in realistic modeling situations (e.g., Bearden, Sharma, & Teel, 1982; Tanaka, 1987). As a result, a large number of alternative fit indices have been developed. As most are indices rather than statistics, there is no strong basis for particular cutoff values that would serve as targets for evaluations of power (Hu & Bentler, 1999). Moreover, simulation studies of power that use popular

rule-of-thumb cutoffs find that the magnitude of these indices is influenced by factors other than model fit such as estimation method (e.g., Fan, Thompson, & Wang, 1999). These caveats aside, we can draw some general conclusions from the simulation work on sample size and statistical power in SEM.

In the simulation study of simple mediation models referenced earlier, Hoyle and Kenny (1999) found the power to detect the indirect effect through a single mediator of a single predictor on a single outcome to be unacceptably low at *N*s of 100 or less, peaking at .65 for $N = 100$ and $\alpha = .90$. At $N = 200$, power exceeded the standard target of .80 when indicators were at least moderately reliable ($\alpha = .75$). Kim (2005) examined the power of several fit indices as a function of sample size, number of variables, and the magnitude of the relations between variables. Kim's results give a general sense of the degree of power typical for confirmatory factor models at different *N*s. For the comparative fit index (CFI; Bentler, 1990), acceptable power was evident for *N*s of about 70 when factor loadings were high (λs = .8) but rose to more than 200 when factor loadings were moderate (λs = .6). Power for the root mean square of approximation (Steiger & Lind, 1980) varies as a function of both sample size and the number of degrees of freedom, which is related to the number of variables. For the hypothesis of close fit (see MacCallum, Browne, & Sugawara, 1996), *N*s required to achieve acceptable power were 294, 147, and 73 for models with 6, 9, and 15 variables, respectively. Although these values are within the range of typical prevention studies, the optimism they bring must be tempered by the knowledge that substantially larger *N*s are required for more complex models (e.g., Kim, 2012).

A more general treatment of power for the omnibus test of close fit was offered by MacCallum et al. (1996). Their power tables (e.g., Table 2, p. 142) show clearly the power advantage achieved by reducing the number of parameters to be estimated in a model and, in so doing, increasing the number of degrees of freedom. For example, at $N = 100$, the likelihood of detecting close fit is .65 for a model with df = 100 but only .13 for a model with df = 5. In general, their work suggests the need for samples of size 200 or greater with at least 50 degrees of freedom for ample power.

M*plus* softeware now has a MonteCarlo feature that permits users to conduct their own power analysis, both for individual effects and for omnibus model fit. Examples of power calculations are available on the M*plus* website

Validity of fit indices when *N* is small—Power considerations notwithstanding, the performance of some fit indices is problematic at small, or even modest, sample sizes. The literature on this topic is large and far ranging, but an example will serve to illustrate the point. Bentler (1990) evaluated the performance of five comparative fit indices under a variety of data and model conditions, including sample size, which ranged from 50 to 1600. To reinforce the point addressed earlier, he observed about 12% improper solutions at $N = 50$, a trivial number at $N = 100$, and none at all at *N*s larger than 200. Although the performance of the CFI was acceptable at $N = 50$, the nonnormed fit index (also referred to as the Tucker-Lewis index) was highly variable at *N*s of 400 and lower. These findings point to the need to carefully consider which fit indices to use when *N* is small. Some indices are

unreliable at low *N*s and may lead to rejection of a model that is satisfactory or acceptance of a model that is not.

Using SEM When *N* is Small

Our review of the simulation work on SEM and sample size offers a mixed message. The performance of some fit indices and the power of tests of some parameters within models may be acceptable with samples as small as 50 when the variables are normally distributed and the reliability of indicators at least moderate in magnitude. Yet, the performance of estimators with samples in the 50-100 range can be problematic, and to achieve desired levels of power for models of typical complexity requires samples sizes of 200 or more. We recommend that reports of uses of SEM for modeling data from samples smaller than 200 include a justification and reference to limitations given the findings from the simulation research we have summarized.

General Recommendations

Although any recommendation regarding sample size when using MLM or SEM must account for features of the data and the model, we can offer some general recommendations for maximizing the yield of these analyses when *N* is small.

Leave no data unmodeled

Because the initial sample size of many prevention studies is near the minimum for effective use of these methods, it is essential that all cases be retained in the analysis sample. A combination of diligence in retention efforts and the use of missing data methods as needed is recommended when the data are to be analyzed using MLM or SEM.

Optimize the observed data

We noted that the estimators typically used in these methods assume multivariate normality and made brief mention of the fact that the minimum sample size increases as the data depart from normality. Any efforts at achieving normal data are likely to pay off with improvements in estimation and testing. These may concern measurement, scoring, or transformations. We also highlighted the role of unreliability in estimation problems and statistical power for SEM. More reliable indicators of latent variables are associated with fewer estimation problems and increased statistical power. When sample size is small, reliable, normally distributed variables are critical to success in modeling data using MLM or SEM.

Fix and constrain

The power and performance of SEM are improved by increasing the number of degrees of freedom associated with a model. Degrees of freedom can be increased by increasing the number of variables (e.g., number of indicators per latent variable) and decreasing the number of parameters to be estimated. A reduction in the number of parameters to be estimated can be achieved by (1) fixing free parameters to a value, (2) constraining parameters to equal (or correspond to some other function of) other parameters in the model. Because both of these adjustments to a model could lead to a deterioration in fit, they must

be used wisely, typically with reference to knowledge gleaned from prior research with the variables.

A Note on the Limitations of Simulation Studies

Key concerns associated with simulation studies are the choice of which factors to manipulate (e.g., sample size, distribution of the variables, functional form of the model, effect sizes) and levels thereof. In reality, the effects of many factors are moderated by other factors, and those moderated effects can only be studied when the relevant factors are included in the same study. Moreover, conclusions can only be reached regarding the levels of the factors included in the study. These concerns are relevant for simulation work on sample size in MLM and SEM because the effects of small samples on performance vary as a function of various features of model (mis)specification and data; and the performance with very small samples can only be evaluated if they are considered when sample size is a factor. Fortunately, simulation studies are increasingly likely to include sample sizes that historically would have been considered too small to warrant study (e.g., $N = 50$).

Summary and Conclusions

A primary concern of prevention science is determination of the complex and dynamic structures and processes involved in problem behavior and its prevention. Advances in measurement and statistical methods now allow for the specification and evaluation of models that approximate the complexity of those structures and processes. MLM and SEM are two such statistical methods. They share in common is the need for samples of sufficient size to ensure valid tests of model fit and estimates and tests of parameters within models of adequate fit. Our concern has been the ways in which sample size affects estimation and testing in MLM and SEM, lower bounds of sample size for different aspects of data analysis using these methods, and strategies for optimizing applications of MLM and SEM when samples are small.

We have drawn attention to the sample size considerations for each of these methods in turn. To conclude, we turn our attention to considerations that apply to MLM, SEM, and other approaches to modeling prevention research data. The first such consideration is whether any model beyond a two- or three-variable system should be estimated at all. Clearly there are absolute lower bounds for sample size that determine whether any meaningful results can be obtained from modeling. We have identified those lower bounds for different modeling situations when MLM and SEM are used for estimation and testing. When lower bounds cannot be met, the alternatives are simpler analyses that focus on estimating and testing parameters in contexts as close as possible to that of the guiding theoretical model. Examples include analysis of covariance and multiple regression analysis. When sample size is small but larger than the lower bound for use of the modeling method, the considerations are for optimization of estimation and testing. We have reviewed the effectiveness of different estimation methods at different sample sizes for different types of models. Optimization in this way may require moving away from software defaults (e.g., maximum likelihood in SEM software) to alternatives that require user specification.

Beyond this concern is the standard concern in treatments of sample size and statistical analyses—statistical power. Our treatment of power in this manuscript has been conceptual and strategic, owing largely to the fact that the issue of power in MLM and SEM is multidimensional and multiply determined in more ways than is typical of statistical analyses in prevention science. We presented findings from simulation work on models of general interest to give a sense of the number of cases generally required for adequate power. Collectively, these considerations lead to the conclusion that data from a few dozen cases, particularly when they are clustered, are not suitable for modeling with MLM or SEM. However, in light of the potential MLM and SEM offer for modeling the structures and processes implicated in prevention research, when possible, the investment required to assemble the larger, though still modest-sized, samples required for responsible use of these methods, is well justified.

Acknowledgments

During the writing of this manuscript, the authors were supported by National Institute on Drug Abuse (NIDA) Grant P30 DA023026. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NIDA.

References

Aitkin MA, Longford NT. Statistical modeling issues in school effectiveness studies. Journal of Royal Statistical Society. 1986; 149:1–26. Retrieved from http://www.jstor.org/stable/2981882.

Anderson TW. Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. Journal of the American Statistical Association. 1957; 52:200–203. Retrieved from http://www.jstor.org/stable/2280845.

Austin PC. Estimating multilevel logistic regression models when the number of clusters is low: A comparison of different statistical software procedures. The International Journal of Biostatistics. 2010; 6:1–18.10.2202/1557-4679.1195

Bandalos, DL.; Gagné, P. Simulation methods in structural equation modeling. In: Hoyle, RH., editor. Handbook of structural equation modeling. New York: Guildford Press; 2012. p. 92-108.

Bauer DJ. Estimating multilevel linear models as structural equation models. Journal of Educational and Behavioral Statistics. 2003; 28:135–167.10.3102/10769986028002135

Bauer DJ, Sterba SK. Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. Psychological Methods. 2011; 16:373–390.10.1037/a0025813 [PubMed: 22040372]

Bearden WO, Sharma S, Teel JE. Sample size effects on chi square and other statistics used in evaluating causal models. Journal of Marketing Research. 1982; 19:425–430.10.2307/3151716

Bentler PM. Comparative fit indexes in structural models. Psychological Bulletin. 1990; 107:238–246.10.1037/0033-2909.107.2.238 [PubMed: 2320703]

Bollen KA. Overall fit in covariance structure models: Two types of sample size effects. Psychological Bulletin. 1990; 107:256–259.10.1037/0033-2909.107.2.256

Bollen, KA.; Curran, PJ. Latent curve models: A structural equation approach. Hoboken NJ: Wiley; 2006.

Bollen KA, Noble MD. Structural equation models and the quantification of behavior. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108(Suppl 3):15639–15646.10.1073/pnas.1010661108 [PubMed: 21730136]

Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association. 1993; 88:9–25. Retrieved from http://www.jstor.org/stable/2290687.

Curran PJ. Have multilevel models been structural equation models all along? Multivariate Behavioral Research. 2003; 38:529–569.10.1207/s15327906mbr3804_5

Curran, PJ.; Lee, T.; Howard, AL.; Lane, S.; MacCallum, R. Disaggregating within-person and between-person effects in multilevel and structural equation growth models. In: Harring, JR.; Hancock, GR., editors. Advances in longitudinal models in the social and behavioral sciences. Charlotte, NC: Information Age Publishing; 2012. p. 217-253.

Demidenko, E. Mixed models: Theory and applications. Hoboken, NJ: Wiley; 2004.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B. 1977; 44:1–38. Retrieved from http://www.jstor.org/stable/2984875?origin=JSTOR-pdf.

Duncan SC, Duncan TE, Strycker LA. A multilevel analysis of neighborhood context and youth alcohol and drug problems. Prevention Science. 2002; 3:125–133.10.1023/A:1015483317310 [PubMed: 12088137]

Erdfelder E, Faul F, Buchner A. GPOWER: A general power analysis program. Behavior Research Methods, Instruments, & Computers. 1996; 28:1–11.10.3758/BF03203630

Fan X, Thompson B, Wang L. Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. Structural Equation Modeling. 1999; 6:56–83.10.1080/10705519909540119

Gagné P, Hancock GR. Measurement model quality, sample size, and solution propriety in confirmatory factor models. Multivariate Behavioral Research. 2006; 41:65–83.10.1207/s15327906mbr4101_5

Gillmore MR, Hawkins JD, Catalano RF Jr, Day LE, Moore M, Abbott R. Structure of problem behaviors in preadolescence. Journal of Consulting and Clinical Psychology. 1991; 59:499–506.10.1037/0022-006X.59.4.499 [PubMed: 1918552]

Goldstein H. Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika. 1986; 73:43–56.10.1093/biomet/73.1.43

Hopkin CR, Hoyle RH, Gottfredson NC. Maximizing the yield of small samples in prevention research: A review of general strategies and best practices. 2013 Manuscript submitted for publication.

Hoyle, RH. Structural equation modeling for social and personality psychology. London, UK: Sage Publications; 2011.

Hoyle, RH.; Kenny, DA. Sample size, reliability, and tests of statistical mediation. In: Hoyle, H., editor. Statistical strategies for small sample research. Thousand Oaks, CA: Sage Publications; 1999. p. 195-222.

Hu L-T, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling. 1999; 6:1–55.10.1080/10705519909540118

Jackson DL, Voth J, Frey MP. A note on sample size and solution propriety for confirmatory factor analytic models. Structural Equation Modeling. 2013; 20:86–97.10.1080/10705511.2013.742388

Kaplan, D. Statistical power in structural equation modeling. In: Hoyle, RH., editor. Structural equation modeling: Concepts, issues, and applications. Newbury Park, CA: Sage Publications; 1995. p. 100-117.

Kim KH. The relation among fit indexes, power, and sample size in structural equation modeling. Structural Equation Modeling. 2005; 12:368–390.10.1207/s15328007sem1203_2

Kim S-Y. Sample size requirements in single- and multiphase growth mixture models: A Monte Carlo simulation. Structural Equation Modeling. 2012; 19:457–476.10.1080/10705511.2012.687672

Kreft, IGG.; de Leeuw, J. Introducing multilevel modeling. Thousand Oaks, CA: Sage Publications; 1988.

Kreft IGG, de Leeuw J, Aiken LS. The effect of different forms of centering in hierarchical linear models. Multivariate Behavioral Research. 1995; 30:1–21.10.1207/s15327906mbr3001_1

Lüdtke O, Marsh HW, Robitzsch A, Trautwein U, Asparouhov T, Muthén B. The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. Psychological Methods. 2008; 13:203–229.10.1037/a0012869 [PubMed: 18778152]

Maas CJM, Hox JJ. Sufficient sample sizes for multilevel modeling. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences. 2005; 1:85–91.10.1027/1614-1881.1.3.86

MacCallum RC, Browne MW, Sugawara HM. Power analysis and determination of sample size for covariance structure modeling. Psychological Methods. 1996; 1:130–149. Retrieved from http://doi.apa.org/journals/met/1/2/130.pdf.

MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. Psychological Methods. 2002; 7:19–40.10.1037/1082-989X.7.1.19 [PubMed: 11928888]

Marsh HW, Hau K-T, Balla JR, Grayson D. Is more ever too much? The number of indicators per factor in confirmatory factor analysis. Multivariate Behavioral Research. 1998; 33:181–220.10.1207/s15327906mbr3302_1

McArdle, JJ.; Nesselroade, JR. Growth curve analysis in contemporary psychological research. In: Schinka, J.; Velicer, W., editors. Comprehensive handbook of psychology: Research methods in psychology. New York: Wiley; 2003. p. 447-80.

Meredith W, Tisak J. Latent curve analysis. Psychometrika. 1990; 55:107–122.10.1007/BF02294746

Noh M, Lee Y. REML estimation for binary data in GLMMs. Journal of Multivariate Analysis. 2007; 98:896–915.10.1016/j.jmva.2006.11.009

Rabe-Hesketh S, Skrondal A, Pickles A. Reliable estimation of generalized linear mixed models using adaptive quadrature. The Stata Journal. 2002; 2:1–21. Retrieved from http://www.stata-journal.com/article.html?article=st0005.

Raudenbush SW. Statistical analysis and optimal design for cluster randomized trials. Psychological Methods. 1997; 2:173–185.10.1037/1082-989X.2.2.173

Raudenbush, SW.; Bryk, AS. Hierarchical linear models: Applications and data analysis methods. Thousand Oaks, CA: Sage Publications; 2002.

Raudenbush, SW.; Bryk, AS.; Congdon, R. HLM 6 for Windows [Computer software]. Lincolnwood, IL: Scientific Software; 2004.

Raudenbush SW, Liu X. Statistical power and optimal design for multisite randomized trials. Psychological Methods. 2000; 5:199–213.10.1037//1082-989X.5.2.199 [PubMed: 10937329]

Raudenbush, SW.; Spybrook, J.; Congdon, R.; Liu, X.; Martinez, A. Optimal Design Software for multi-level and longitudinal research (version 3.01) [Software]. 2011. Available from www.wtgrantfoundation.org

Rodríguez G, Goldman N. Improved estimation procedures for multilevel models with binary response: A case study. Journal of the Royal Statistical Society, Series A (Statistics in Society). 2001; 164:339–355.10.1111/1467-985X.00206

Schabenberger, O. Introducing the GLIMMIX procedure for generalized linear mixed models. Cary, NC: SAS Institute; 2005. p. 196-30.Retrieved from http://www2.sas.com/proceedings/sugi30/

Schwarz GE. Estimating the dimension of a model. Annals of Statistics. 1978; 6:461–464. Retrieved from http://www.jstor.org/stable/2958889.

Shiyko MP, Lanza ST, Tan X, Li R, Shiffman S. Using the time-varying effect model (TVEM) to examine dynamic associations between negative affect and self-confidence on smoking urges: Differences between successful quitters and relapsers. Prevention Science. 2012; 13:288–299.10.1007/s11121-011-0264-z [PubMed: 22246429]

Simon TR, Ikeda RM, Smith EP, Reese LE, Rabiner DL, Miller-Johnson S, Winn DM, Dodge KA, Asher SR, Home AM, Orpinas P, Martin R, Quinn WH, Tolan PH, Gorman-Smith D, Henry DB, Gay FN, Schoeny M, Farrell AD, Meyer AL, Sullivan TN, Allison KW. The Multisite Violence Prevention Project: Impact of a universal school-based violence prevention program on social-cognitive outcomes. Prevention Science. 2008; 9:231–244.10.1007/s11121-008-0101-1 [PubMed: 18780181]

Singer, JD.; Willett, JB. Applied longitudinal data analysis: Modeling change and event occurrence. New York: Oxford University Press; 2003.

Snijders, TAB.; Bosker, RJ. Multilevel analysis: An introduction to basic and advanced multilevel modeling. 2. London: Sage Publications; 2004.

Steiger, JH.; Lind, JC. Statistically based tests for the number of common factors; Paper presented at the Meeting of the Psychometric Society; Iowa City, IA. 1980 May.

Tanaka JS. "How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. Child Development. 1987; 58:134–146. Retrieved from http://www.jstor.org/stable/1130296.

Westland JC. Lower bounds on sample size in structural equation modeling. Electronic Commerce Research and Applications. 2010; 9:476–487.10.1016/j.elerap.2010.07.003

Weston R, Gore PA Jr. A brief guide to structural equation modeling. The Counseling Psychologist. 2006; 34:719–751.10.1177/0011000006286345

Whiteside SP, Lynam DR. The Five Factor Model and impulsivity: Using a structural model of personality to understand impulsivity. Personality and Individual Differences. 2001; 30:669–689.10.1016/S0191-8869(00)00064-7