Published in final edited form as:
Educ Psychol Meas. 2013 December ; 76(6): 913–934. doi:10.1177/0013164413495237.

Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety

Erika J. Wolf[1,2], **Kelly M. Harrington**[1,2], **Shaunna L. Clark**[3], and **Mark W. Miller**[1,2]

[1]National Center for PTSD at VA Boston Healthcare System, Boston, MA, USA

[2]Boston University School of Medicine, Boston, MA, USA

[3]Center for Biomarker Research and Personalized Medicine, School of Pharmacy, Virginia Commonwealth University, Richmond, VA, USA

Abstract

Determining sample size requirements for structural equation modeling (SEM) is a challenge often faced by investigators, peer reviewers, and grant writers. Recent years have seen a large increase in SEMs in the behavioral science literature, but consideration of sample size requirements for applied SEMs often relies on outdated rules-of-thumb. This study used Monte Carlo data simulation techniques to evaluate sample size requirements for common applied SEMs. Across a series of simulations, we systematically varied key model properties, including number of indicators and factors, magnitude of factor loadings and path coefficients, and amount of missing data. We investigated how changes in these parameters affected sample size requirements with respect to statistical power, bias in the parameter estimates, and overall solution propriety. Results revealed a range of sample size requirements (i.e., from 30 to 460 cases), meaningful patterns of association between parameters and sample size, and highlight the limitations of commonly cited rules-of-thumb. The broad "lessons learned" for determining SEM sample size requirements are discussed.

Keywords

structural equation modeling; confirmatory factor analysis; sample size; statistical power; Monte Carlo simulation; bias; solution propriety

---

Determining sample size requirements for *structural equation modeling* (SEM) is a challenge often faced by investigators, peer reviewers, and grant writers. Advances in approaches to statistical modeling and in the ease of use of related software programs has contributed not only to an increasing number of studies using latent variable analyses but also raises questions about how to estimate the requisite sample size for testing such models.

Corresponding Author: Erika J. Wolf, National Center for PTSD (116B-2), VA Boston Healthcare System, 150 South Huntington Avenue, Boston, MA 02130, USA. erika.wolf@va.gov.

Declaration of Conflicting Interests

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

One of the strengths of SEM is its flexibility, which permits examination of complex associations, use of various types of data (e.g., categorical, dimensional, censored, count variables), and comparisons across alternative models. However, these features of SEM also make it difficult to develop generalized guidelines regarding sample size requirements (MacCallum, Widaman, Zhang, & Hong, 1999). Despite this, various rules-of-thumb have been advanced, including (a) a minimum sample size of 100 or 200 (Boomsma, 1982, 1985), (b) 5 or 10 observations per estimated parameter (Bentler & Chou, 1987; see also Bollen, 1989), and (c) 10 cases per variable (Nunnally, 1967). Such rules are problematic because they are not model-specific and may lead to grossly over-or underestimated sample size requirements. MacCallum et al. (1999) demonstrated that model characteristics such as the level of communality across the variables, sample size, and degree of factor determinacy all affect the accuracy of the parameter estimates and model fit statistics, which raises doubts about applying sample size rules-of-thumb to a specific SEM.

There has been a sharp increase in the number of SEM-based research publications that evaluate the structure of psychopathology and the correlates and course of psychological disorders and symptoms, yet applied information on how to determine adequate sample size for these studies has lagged behind. The primary aim of this study was to evaluate sample size requirements for SEMs commonly applied in the behavioral sciences literature, including *confirmatory factor analyses* (CFAs), models with regressive paths, and models with missing data. We also sought to explore how systematically varying parameters within these models (i.e., number of latent variables and indicators, strength of factor loadings and regressive paths, type of model, degree of missing data) affected sample size requirements. In so doing, we aimed to demonstrate the tremendous variability in SEM sample size requirements and the inadequacy of common rules-of-thumb. Although statisticians have addressed many of these concerns in technical papers, our impression from serving as reviewers, consultants, and readers of other articles is that this knowledge may be inaccessible to many applied researchers and so our overarching objective was to communicate this information to a broader audience.

Sample Size Considerations

When contemplating sample size, investigators usually prioritize achieving adequate *statistical power* to observe true relationships in the data. Statistical power is the probability of rejecting the null hypothesis when it is false; it is the probability of not making a Type II error (i.e., 1 – beta; see Cohen, 1988). Power is dependent on (a) the chosen alpha level (by convention, typically $\alpha = .05$), (b) the magnitude of the effect of interest, and (c) the sample size. However, power is not the only consideration in determining sample size as *bias*[1] in the parameter estimates and standard errors also have bearing. Bias refers to conditions in which an estimated parameter value differs from the true population value (see Kelley & Maxwell, 2003; Maxwell, Kelley, & Rausch, 2008). A standard error may also be biased if it is under- or overestimated (which increases the risk of Type I and II errors, respectively). A

[1]Throughout this article, we refer to the term *bias* in its broadest sense (i.e., a systematic misrepresentation of a parameter estimate or statistic due to methodological problems such as insufficient sample size or poor measurement psychometric characteristics). We also use this term to describe bias in the standard errors and parameter estimates due to insufficient sample size in order to maintain consistency with Muthén and Muthén (2002).

third element of SEMs that is affected by sample size is *solution propriety* (see Gagné & Hancock, 2006). This is whether there are a sufficient number of cases for the model to converge without improper solutions or impossible parameter estimates. Models based on larger samples (Boomsma, 1982; Gagné & Hancock, 2006; Velicer & Fava, 1998), with more indicators per factor (Gagné & Hancock, 2006; Marsh, Hau, Balla, & Grayson, 1998), and with larger factor loadings (Gagneé & Hancock, 2006) are more likely to converge properly.

Monte Carlo Analyses for Sample Size Determinations

Three major approaches to evaluating sample size requirements in SEMs have been proposed: (a) the Satorra and Saris (1985) method, which estimates power based on the noncentrality parameter (i.e., the amount of model misspecification); (b) the MacCallum, Browne, and Sugawara (1996) method, which is based on the power of the model to obtain a root mean square error of approximation value that is consistent with good model fit; and (c) the Monte Carlo simulation method (Muthén & Muthén, 2002), the focus of this research. With this method, associations among variables are set by the user based on a priori hypotheses. The specified associations are akin to the population estimates of the true relationships among the variables. A large number of data sets are then generated to match the population values; each individual data set is akin to a sample and is based on a user-determined number of cases. The hypothesized model is then evaluated in the generated data sets and each parameter estimate is averaged across the simulations to determine if the specified number of cases is sufficient for reproducing the population values and obtaining statistically significant parameter estimates. This approach does not address the power of the overall model—it provides estimates of power and bias for individual effects of interest (i.e., individual factor loadings, correlations, or regressive paths). This is important as an investigator may only be interested in having sufficient sample size for select aspects of a given model. Additional details and script for conducing Monte Carlo analyses in MPlus can be found in Muthén and Muthén (2002), Muthén and Asparouhov (2002), and Muthén (2002). Readers are also referred to Paxton, Curran, Bollen, Kirby, and Chen (2001), who provide a useful step-by-step discussion of conducting Monte Carlo analyses.

It is important to distinguish between two overarching approaches to the use of Monte Carlo analyses. The first, termed *proactive* (see Marcoulides & Chin, in press; Marcoulides & Saunders, 2006), involves conducting simulation studies of a hypothesized model with relationships among the variables specified based on the available research literature; this approach is the focus of this investigation, as described above (see also Paxton et al., 2001). The second approach, termed *reactive* (Marcoulides & Chin, in press; Marcoulides & Saunders, 2006), involves analyzing existing data after a study has been completed to evaluate the hypothesized model many times over by taking repeated random draws of cases from the larger sample (see also Marcoulides, 1990). The model is fit in each subsample and the resulting parameter estimates and fit statistics are then examined across all the random draws. The former approach is a prospectively designed one that has its basis in theory and the relevant literature. The latter approach is a post hoc method that is limited by the quality of the existing data and may lead to unwarranted confidence in the stability of the results or the appropriateness of the sample for the planned analyses. Throughout this article, we

describe the importance of conducting proactive Monte Carlo simulation studies for the purposes of sample size planning.

Aims

The primary aim of this study was to provide applied behavioral science researchers with an accessible evaluation of sample size requirements for common types of latent variable models and to demonstrate the range of sample sizes that may be appropriate for SEM. We hope this will help researchers better understand the factors most relevant to SEM sample size determinations and encourage them to conduct their own Monte Carlo analyses rather than relying on rules-of-thumb. A second aim was to examine how sample size requirements change as a function of elements in an SEM, such as number of factors, number of indicators, strength of indicator loadings, strength of regressive paths, degree of missing data, and type of model.[2] We evaluated these aspects of structural models in one study and included several permutations and combinations of each model characteristic to have broad applicability. Finally, this study also afforded the opportunity to compare sample size requirements for SEMs versus for similar models based on single indicators (i.e., traditional path models).

Method

Procedure

We conducted Monte Carlo simulation studies for several types of CFAs and SEMs following the guidelines described by Muthén and Muthén (2002). For each model, we systematically varied the number of indicators of the latent variable(s) and the strength of the factor loadings and structural elements in the model to examine how these characteristics would affect statistical power, the precision of the parameter estimates, and the overall propriety of the results. We examined models that we thought would have broad applicability and included models with the absolute minimum number of indicators required to yield overidentified measurement models (i.e., models with positive degrees of freedom, meaning that the total number of variances and covariances in the data set exceeded the number of parameter estimates in the analysis).

CFA models—Figure 1 provides an overview of the one-, two-, and three-factor CFAs evaluated. For illustrative purposes, suppose that the one-factor model tests the latent construct of depression, the two-factor model tests correlated latent constructs of depression and anxiety, and the three-factor model tests latent constructs of depression, anxiety, and substance use. We varied the number of indicators of these factors such that the one-factor model was indicated by four, six, or eight indicators and the two- and three-factor models were indicated by three, six, or eight indicators. We did not evaluate a three-indicator, one-factor model because it would be just-identified (i.e., have 0 *df* and hence would estimate all the associations among the data perfectly and yield perfect fit; see Brown, 2006). We also varied the factor loadings (and hence the unreliability of the indicators) using standardized

[2]These characteristics of a model have also been shown to influence other important aspects of model output that are not under consideration in this study, such as the performance and stability of fit indices (see Jackson, 2001; Kim, 2005; Velicer & Fava, 1998).

loadings of .50, .65, or .80. All loadings and structural paths in the models were completely standardized (factor variances were fixed to 1.0, factor means fixed to 0, and the squared indicator loading plus the residual indicator variance equaled 1.0). Within a model, factor loadings were held constant across indicators. For models with more than one factor, the factor intercorrelations were set to $r = .30$; for some models, we also evaluated the effect of a stronger factor intercorrelation ($r = .50$). We did so because determining the sample size required to observe a factor correlation is often an important consideration. For example, this is crucial when evaluating convergent or discriminant validity, in determining the longitudinal stability of a trait, or in determining the similarity of scores across non-independent observations (e.g., married couples, twins, etc). Appendix A provides sample Mplus script for the first CFA model that was evaluated (i.e., the one-factor model with four indicators, each loading at .50).

Structural path models—We also evaluated a three-factor latent variable *mediation* model with regressive direct and indirect paths between factors (i.e., implying directional relationships) and varied the variance explained in the dependent variable. We focused on the mediation model because of its popularity in the behavioral science literature. This model is shown in Figure 2. As an applied example, suppose this model tested if chronic stress (the independent variable) predicted functional impairment (the dependent variable) via symptoms of depression (the mediator). In many mediation models, the researcher is primarily interested in the size and significance of the indirect effect (i.e., stress → depression → impairment) because this effect informs the understanding of the mechanism of action. Therefore, we focused these analyses on the minimum sample size required to observe this indirect effect. In these models, each factor was indicated by three observed variables, and all factor loadings were set to .65. This model was structurally saturated (i.e., no unanalyzed relationships among the latent variables), although the measurement portion of it remained overidentified because all indicators loaded on only one factor and there were no cross-loadings. We varied the total variance explained in the latent dependent variable from 16% to 45% to 75% (corresponding to standardized *indirect* effects of $\beta = .06$, .16, and .25, respectively), kept the magnitude of all structural relationships in the model equal to one another, and held the total variance of all latent and observed variables at 1.0.[3] Sample Mplus script for the first SEM that was evaluated (i.e., the model explaining 16% of the variance in the dependent variable) is included in Appendix A.

Missing data—We next evaluated the effects of missing data on sample size requirements for one CFA and the latent mediation model. Prior work has evaluated the effects of missing data on statistical power analyses in structural equation models (i.e., Davey & Salva, 2009a, 2009b; Dolan, van der Sluis, & Grasman, 2005), but to our knowledge, the specific models under investigation in this study have not been evaluated under missing data conditions with respect to effects on power, parameter bias, and solution propriety. The CFA that was

[3]The model which explained 16% of the variance was based on standardized direct structural paths of $\beta = .25$; the model with 45% variance explained included standardized structural paths of $\beta = .40$, and the model with 75% of the variance explained in the latent dependent variable was based on standardized structural paths of $\beta = .50$. The size of the unstandardized paths was determined through the following equation: total variance explained in dependent variable = c^2* Variance$_X$ + b^2* Variance$_M$ + 2bc* Covariance$_{(X, M)}$ + Residual Variance$_Y$, where c = the parameter estimate for the dependent variable (Y) regressed on the independent variable (X); b = the parameter estimate for the dependent variable regressed on the mediator (M) and variance explained plus residual variance = 1.0.

evaluated was a two-factor model with each factor indicated by three observed variables loading on their respective factors at .65 and a correlation of .30 between the factors, permitting direct comparison to the same model without missing data. We also evaluated the effects of missing data on the mediation model (Figure 2), where each factor was indicated by three observed variables loading at .65, and all direct paths in the model were set to .40. We systematically varied the percentage of missing data *on each indicator* in the models such that each indicator was missing 2%, 5%, 10%, or 20% of its data. We did not model a covariate as a predictor of missingness; thus, in accord with Muthén and Muthén (2002), we assumed the data were missing completely at random (Little & Rubin, 1989).

Single indicators—Finally, we also compared power, bias, and solution propriety of path models that were based on single indicators versus latent variables (i.e., we evaluated the effects of unspecified measurement error in single indicator designs). To do so, we specified a path analysis in which the association between the observed *X* and *Y* variables was mediated by an observed third variable (i.e., a single indicator version of the model depicted in Figure 2). The standardized magnitude of the three direct paths in the model was set to .40, and the total variance of each variable was set to 1.0 (and all means to 0). In total, 45% of the variance in the dependent variable was explained by the direct paths in this model. The reliability of each of the variables was specified in the data simulation phase of the analysis. Specifically, we set observed variable reliability to .42, .81, or .90, which, in the case of the use of completely standardized factor loadings, is equivalent to factor loadings of .65, .90, and .95, respectively.[4] After specifying the degree of indicator reliability for data generation, we then purposefully misspecified the model by setting the reliability of the indicators to 1.0 in the data analysis phase. Essentially, this is the assumption of a single indicator analysis, as measurement error is not removed from the variance of the measure and the indicator is treated as if it is a perfect measure of the latent construct.

Data Analysis

Overview—All models were evaluated using Mplus version 5.2 (Muthen & Muthen, 1998-2008). All models were based on a single group with 10,000 replications of the simulated data. We set the sample size of a given model and then adjusted it (upwards or downwards) based on whether the results met our criteria for acceptable precision of the estimates, statistical power, and overall solution propriety, as detailed below. In the first model, we started with a sample size of 200 because that has previously been suggested as the minimum for SEMs (Boomsma, 1982); starting sample sizes for subsequent models were determined based on the results of prior models. To determine the minimum sample size required, we tested the effects of increasing or decreasing the sample size of each model by $n = 10$ (or $n = 20$ when it was clear that increments of 10 were too fine to yield meaningful differences among the models). Next, we continued to increase the sample size by a unit of 10 or 20 to test stability of the solution, as defined by both the minimum sample size and the next largest sample size meeting all a priori criteria. Finally, as recommended by Muthén and Muthén (2002), we tested the stability of the results by running the analyses again with a

[4]Reliability = the proportion of true score variance (i.e., the squared factor loading) to total score variance (i.e., true score + error variance).

new, randomly selected seed number (i.e., so that data generation began at a different point, resulting in a different set of 10,000 data sets compared to the analysis with the initial seed number).

All analyses were conducted using the maximum likelihood (ML) estimator. The simulated data were dimensional (i.e., continuous) and normally distributed. Models with missing data invoked full information ML estimation for cases with missingness as this approach performs well when data are missing completely at random or missing at random (Enders & Bandalos, 2001).

Criteria for evaluation of sample size requirements—We used several criteria to evaluate the minimum sample size required to achieve minimal bias, adequate statistical power, and overall propriety of a given model, following recommendations by Muthén and Muthén (2002). All criteria had to be met in order to accept a given *N* as the minimum sample size.

Bias—First, we evaluated (a) the degree of bias in the parameter estimates, (b) the degree of bias in the standard error estimate, and (c) the 95% confidence interval for the parameter estimates. Specifically, we examined the mean parameter estimates (i.e., factor loadings and regressive paths) across simulations in comparison to the specified population values. This index of parameter bias was derived by subtracting the population value from the mean estimated value and dividing by the population value, following Muthén and Muthén (2002). We specified that the most discrepant mean parameter estimate in a model had to be within 5% of the population value. We also evaluated the potential for standard error bias in an analogous way, by comparing mean standard errors in the generated data sets to the *SD* for each parameter estimate (the *SD* of the estimates over many replications is akin to the population standard error; Muthén & Muthén, 2002). Acceptable standard error estimates also had to be within 5% of the population standard error. Finally, we specified that the 95% confidence interval for each parameter estimate had to include the population value in at least 90% of the analyses of the simulated data (i.e., achieve adequate accuracy).

Power—Second, we specified that estimated power had to be 80% or greater (with $\alpha = .05$) for all parameters of interest (in this case, factor loadings, correlations, regressive paths) in the model. Thus, a model in which any one of the parameters of interest fell below 80% power would be rejected (this approach is conceptually similar to the “all-pairs power” approach described by Maxwell, 2004).

Overall solution propriety—Third, we specified that the analysis could not yield any errors (i.e., improper solutions or failures to converge). This is a conservative criterion as a single error in the analysis of 10,000 data sets is not particularly worrisome, assuming that all other criteria are met. However, we adopted this rule for several reasons. First, Mplus provides a description of any errors in the analysis of the generated data but omits these cases from the overall statistics summarized across replications. Depending on the number and type of errors, this can potentially yield overly optimistic summary statistics and also raises the possibility that a similarly sized sample of real data might produce improper solutions. Second, as is evident in this study, the number of errors in the analysis of the

generated data increased with decreasing sample size, which further suggests the need to consider such instances when evaluating sample size requirements.

Results

Orientation to Figures and Tables

Figure 3 shows the minimum sample size (meeting all a priori criteria) required for each of the models. Supplementary tables appear online. Although not shown in the figures or tables, all models yielded acceptable coverage (i.e., the 95% confidence interval for each parameter estimate included the population value in at least 90% of the replications), with the exception of the single indicator models (discussed below).

CFAs

Effect of number of factors—Within the CFAs, increasing the number of latent variables in a model resulted in a significant increase in the minimum sample size when moving from one to two factors, but this effect plateaued, as the transition from two to three factors was not associated with a concomitant increase in the sample size. For example, as shown in Figure 3 (Panels A–C) and Supplemental Tables 1–3, sample size requirements at least doubled when comparing the simplest possible one-factor, four-indicator model, which required a sample of 190, 90, and 60 participants at factor loadings of .50, .65, and .80, respectively, relative to the simplest possible two-factor model (with three indicators per factor), which required a minimum sample of 460, 200, and 120, respectively. However, this was not true for comparisons between two- and three-factor models. In these cases, sample sizes were relatively unchanged, or in some cases decreased, with the addition of the third factor.

Effect of number of indicators—Overall, models with fewer indicators required a larger sample relative to models with more indicators. A one-factor, four-indicator model with loadings of .50, .65, and .80 required sample sizes of 190, 90, and 60, respectively, while a one-factor, six-indicator model required sample sizes of 90, 60, and 40, respectively (see Figure 3, Panels A–C). However, this effect also leveled as the transition from six to eight indicators did not result in as dramatic a decrease in sample size (i.e., this dropped from 60 to 50 when increasing from six to eight indicators in the one-factor model with factor loadings of .65) and, in some instances, yielded no reduction in minimum sample size (i.e., both the six- and the eight-indicator one-factor model with loadings of .50 were associated with a minimum sample size of 90). The same general pattern held for the two- and three-factor models at factor loadings of .50 and .65 but the effect of increasing the number of indicators was less evident in the models with factor loadings of .80 (see Figure 3, Panel C, and Supplemental Table 3).

Effect of magnitude of factor loadings—Models with stronger factor loadings required dramatically smaller samples relative to models with weaker factor loadings. For example, the one-factor, four-indicator model with loadings of .50, .65, and .80 required a minimum of 190, 90, and 60 observations, respectively; decreasing the strength of the factor loadings from .80 to .50 necessitated a threefold increase in the sample size. On average,

factor loadings of .50 were associated with nearly 2.5-fold increases in required sample size relative to an identical model with loadings of .80.

Effect of magnitude of factor correlations—As described in the notes to Supplementary Tables 1, 2, and 3, we also evaluated the effect of increasing the factor intercorrelation from .30 to .50. On average, the increased factor relationship was associated with 78 fewer participants in the minimum acceptable sample size for the CFA models with .50 factor loadings, 55 fewer participants for CFA models with .65 factor loadings, and 37 fewer participants for CFA models with .80 factor loadings.

SEMs

Effect of magnitude of regressive paths—Mediation models with larger effects tended to achieve adequate statistical power for the direct and indirect effects in the model with smaller sample sizes. For example, the model in which the direct effects accounted for 16% of the variance in the dependent variable required greater than 2.4 times more participants to achieve adequate statistical power relative to the model in which 45% of the variance was explained (i.e., n = 440 vs. 180). However, the model that explained 75% of the variance in the dependent variable actually required more participants than the model that explained 45% of the variance. This was not due to statistical power (which was sufficient even at relatively small sample sizes) but instead primarily due to bias of the parameter estimates in the model with larger effects. This likely reflects larger standard errors (evidenced by larger standard error estimates for the regressive paths) as the true value becomes more extreme, thus necessitating increased sample size.

Effect of Missing Data—Greater amounts of missing data in the two-factor CFA and mediation models generally necessitated larger sample sizes (see Figure 3, Panel E, and Supplemental Table 5). This was primarily due to problems with errors in the analysis of generated data sets. For example, the two-factor CFA model with three indicators per factor loading at .65 and a factor intercorrelation of .30 was associated with a minimum sample size of 200 when there were no missing data or only a small amount of missing data (i.e., 2% per indicator). However, the minimum sample size increased to 260 with 5% and 10% missing data per indicator and to 320 with 20% missing data per indicator because of errors that occurred at smaller sample sizes. In contrast, power was not dramatically affected by the addition of missingness. The same basic pattern emerged with respect to the effect of missing data in the mediation models.

Effect of Latent Variables—The comparison of latent versus observed variables in the mediation model demonstrated that no amount of increasing the sample size of single indicator models could account for bias in those results; the attenuation in the magnitude of the parameter estimates was directly related to the degree of unspecified unreliability in the measure (see Cole & Maxwell, 2003). Specifically, the results showed that a measure with 42% reliability (equivalent to a factor loading of .65) would estimate a direct path parameter estimate of .17 when the true population estimate was .40 (biased by 58%), whereas a measure with 90% reliability (equivalent to a factor loading of .95) would estimate a direct path parameter estimate of .36 when the true estimate was .40 (biased by 10%). Statistical

power was also affected because the model with reliability of 42% estimated the direct effect as .17 (instead of the true effect of .40) and a sample size of 180, which was sufficient for the equivalent latent variable mediation model, was inadequate (power = 68%) to detect an effect of .17. The single-indicator models with more reasonable amounts of indicator reliability (.81 and .90) achieved at least 99% power for all direct and indirect paths because the parameter estimates in these models, while still attenuated, were not so small as to affect statistical power. If the three observed variables in the single-indicator mediation model contained only true score variance (i.e., unreliability = 0), then a sample size of 50 would achieve 86% power to detect the smallest direct effect, and a sample size of 70 would be the minimum required to detect the projected indirect effect (and would achieve 81% power).

Stability of Results—With a few exceptions, solutions that met all a priori criteria at a given sample size were stable relative to the results of the analysis at the next largest sample size. We observed somewhat greater instability of results, however, with respect to the reanalysis of the best solutions using a new seed number. Specifically, in 34% of the analyses, the minimum sample size increased (average increase in sample size = 24 cases, range = 10–50); in 16% of the analyses, the minimum sample size decreased (average decrease in sample size = 18 cases, range = 10–30); and in 50% of the analyses, the minimum necessary sample size was equal across analyses using the two seed numbers. When reevaluated with the new seed number, 41% of the CFA models, 67% of the SEM models (note that there were only three total), and 75% of the missingness models required a change in sample size. In the majority of the cases, errors caused the increase in sample size in the models with the new seed number; however, in a minority of cases, power (when it was just on the threshold of .80 using the first seed number), and bias were the unstable factors that necessitated increased sample size.

Discussion

In this study, we systematically evaluated sample size requirements for common types of SEMs by performing Monte Carlo analyses that varied by type of model, number of factors, number of indicators, strength of the indicator loadings and regressive paths, and the amount of missing data per indicator. We evaluated the extent to which statistical power, parameter estimate bias, and the overall propriety of the results affected sample size requirements for these different models. In so doing, we aimed to demonstrate to applied researchers the broad variability in sample size requirements for latent variable models and show how the sample size estimates vary greatly from model to model. This is important as the results of methods-focused articles are not always accessible and interpretable to applied researchers, thus misinformation about sample size requirements persists in the literature and in reviewer feedback on manuscripts and grant applications. In the paragraphs that follow, we describe our results with respect to the broad “lessons learned” from this study.

Optimizing Model Characteristics

Lesson 1A: One size does not fit all—The results of this study demonstrate the great variability in SEM sample size requirements and highlight the problem with a “one size fits all” approach, consistent with the observations of MacCallum et al. (1999). More

specifically, required sample sizes ranged from 30 cases (for the one-factor CFA with four indicators loading at .80) to 460 (for the two-factor CFA with three indicators loading at .50). In comparison, the 10 cases per variable rule-of-thumb would have led to sample size recommendations ranging from 40 to 240, respectively. Furthermore, rather than increasing linearly with number of estimated parameters or number of variables, we found that sample size requirements actually decreased when the number of indicators of a factor increased. This was likely a result of the increase in information available for use in solving the simultaneous regression equations. This effect was particularly evident in moving from three or four indicators to six, but less so when transitioning from six to eight indicators. This is consistent with prior work suggesting that increasing the number of indicators per factor may be one way to compensate for an overall small sample size and preserve statistical power (Marsh et al., 1998). This suggests that researchers may wish to design models in which the number of indicators per factor is greater than the minimum number of indicators required for model identification. Investigators should consider the tradeoff between the simplicity of having fewer indicators per factor versus the benefits of enhanced power and precision that come with having more than the minimum number of indicators per factor.

Lesson 1B: More is not always better—While the number of indicators in a model had an inverse effect on sample size requirements (i.e., models with more indicators per factor required fewer cases), other types of increases to the complexity of the model necessitated increased sample size requirements. For example, increasing the number of factors yielded a commensurate increase in the minimum sample size required. This was likely a function of the need to have adequate power to detect the correlation between the factors. This is an important point as it is not uncommon to have a measurement model in which the correlation between the factors is the focus of investigation and is expected to be small to moderate in magnitude (i.e., when evaluating the discriminant validity of constructs). Thus, evaluating a multifactor measurement model in which the factors are thought to be distinct but correlated may require larger samples compared to a model in which the factor correlation is not of interest and the researcher is focused only on obtaining adequate sample size to estimate the factor loadings.

While models with stronger effects generally required fewer cases, this effect was nonlinear, as SEM models with large path coefficients actually required substantially more cases relative to the same models with more moderate strengths of association due to bias in the parameter estimates. In general, models with indicators with strong relationships to the latent variables yielded more problems with biased and error-prone solutions, even after power reached acceptable levels. In contrast, models with weaker factor loadings tended to have problems with power, even when bias and errors were not a problem.

The broad take home message is that, with respect to the strength of the factor loadings and the magnitude of regressive effects in the model, effects that are very weak or very strong may require larger samples relative to effects that are more moderate in magnitude; albeit this effect is more pronounced in models with very weak effects. This implies that researchers need to carefully consider these parameters when planning an SEM and to keep in mind that having large effects, while associated with generally improved statistical power,

yields new concerns about parameter bias that may ultimately result in increased sample size requirements in order to obtain sufficiently accurate estimates.

The Perils of Ignoring Bias, Errors, Small Effects, and Missing Data

Lesson 2A: Power is not enough—This study highlights how attending only to statistical power is problematic when contemplating sample size requirements. In many models, statistical power was not the limiting factor driving sample size requirements, but rather, bias or errors (i.e., solution propriety) were the culprit. It is our impression that applied researchers often consider sample size needs only in relationship to achieving adequate statistical power.

Lesson 2B: Latent variable models have benefits—Results of this study also reveal the advantages associated with the use of latent variables and the inclusion of error theory in the model. Although a path model composed of only single indicators certainly requires fewer participants than the equivalent model evaluated with latent variables, this is based on the almost certainly false tenet that the single indicators are perfect measures of the underlying construct (i.e., in the behavioral sciences, most phenomena of interest are not directly observable and thus are measured with error). To the extent that a single indicator is not reliable, then it is clear from these results that bias in the parameter estimates is introduced (in direct relation to the degree of error in the indicator). Prior work has arrived at this same general conclusion (Cole & Maxwell, 2003), although to our knowledge, no prior study has systematically evaluated and compared sample size requirements with respect to statistical power, bias, and solution propriety under varying parameter conditions across latent-variable and single-indicator models. No amount of increasing the sample size in a single-indicator model will compensate for inadequate reliability of the measures. In most single-indicator designs, concerns regarding bias are not considered when evaluating minimum sample size as only statistical power is typically evaluated. If an investigator has no choice but to use single indicators of a given construct, then it is generally advisable to attempt to account for measurement error in the indicator by estimating it (see Brown, 2006). Doing so will likely improve the strength and accuracy of the parameter estimates, and hence, statistical power to detect them.

Lesson 2C: Don’t ignore the weakest link—In addition, these results demonstrate the need to attend to sample size requirements for the smallest effect of interest in the model. For example, in a mediation model, the strength of the indirect effect (often the effect of greatest interest to researchers) is always weaker than its component parts and this indirect effect may not be a focus of the Monte Carlo modeling. In fact, the Monte Carlo simulation could proceed without inclusion of the line in the model script that estimates the size and significance of the indirect effect, and a researcher might erroneously determine sample size requirements based only on the weakest projected direct effect. For example, note that in the simulations in which an impressive 75% of the variance in the dependent variable was explained, the magnitude of the indirect effect was still modest at best ($\beta = .25$). Failing to focus on the magnitude of the indirect effect could lead to a study with insufficient sample size.

Lesson 2D: Missing data matter—Results of this study also highlight problems with not accounting for missing data in determining sample size requirements. Models with 20% missing data per indicator necessitated, on average, nearly 50% increase in sample size requirements (due to problems with error and bias). This is a substantial relative increase in the sample size requirement and implies that failure to account for missing data when determining sample size requirements may ultimately lead to sample size benchmarks that are not sufficient. Prior work suggests that researchers must consider if missing data are missing at random versus missing completely at random, as this distinction will likewise affect statistical power and sample size requirements (Davey & Salva, 2009a; Dolan et al., 2005). Although judging the effects of missing data may add an additional layer of complexity and time to the planning phase of research, the results of this study (and prior work) suggest that attending to the relationship between missing data and sample size requirements is time well spent.

To summarize, failure to consider potential measurement error, the effects of parameter bias, unevaluated weak effects, and the amount of missing data is likely to lead to sample size estimates that are inappropriate for the planned analyses. Participant and investigator time and effort, and research dollars, may be wasted if these contributors to sample size needs are overlooked. Attending to potential sources of bias and error is important for ensuring the validity of the results and their interpretation.

Ensuring Stability of Sample Size Requirements

Lesson 3A: Sample size estimates may vary—This study also demonstrated that the minimum sample required for a given model was not always stable, with respect to (a) having the next largest sample size yield acceptable results and (b) evaluating the same model with a new seed number. This problem was particularly evident in models that included missing data. Researchers may want to assess the cause of the instability of sample size requirements before determining how to respond to it. If reanalysis with a new seed number reveals bias in a primary parameter estimate, it may warrant choosing the most conservative sample size estimate, whereas a single error resulting from reanalysis may not warrant increasing the sample size. The difference in sample size requirements when evaluated with different seed numbers was sometimes quite meaningful—in one instance, there was a discrepancy of 50 cases across the two simulations. Depending on the participant pool, 50 additional cases may be unattainable. Investigators may want to combine across sample size estimates, for example, by taking the mean minimum sample size based on analyses with several different seed numbers. The lesson here is that, rather than focusing on a single minimum sample size requirement, it is preferable for investigators to think of acceptable sample size estimates as falling within a range.

Limitations and Conclusion

This study was limited in that we did not evaluate all of the possible factors that could have bearing on sample size requirements. For example, we did not evaluate the effects of nonnormal data (see Enders, 2001), or of the use of categorical, count, censored, or item parceled data (see Alhija & Wisenbaker, 2006). We did not evaluate models in which indicators loaded very weakly on the latent variables. In addition, for the sake of simplicity

and consistency, we held the size of all equivalent parameter estimates in a model consistent, but of course this pattern of associations is unlikely in real data sets. We did not evaluate the sample size necessary to obtain overall good model fit (e.g., Kim, 2005) and there are many other types of models including longitudinal designs (see Hertzog, von Oertzen, Ghisletta, & Lindenberger, 2008), multigroup designs (see Hancock, Lawrence, & Nevitt, 2000), and power equivalence designs (see MacCallum, Lee, & Browne, 2010) that we did not evaluate. We limited the models under consideration to those that we thought would have the widest appeal and that would be most readily interpretable.

We hope that researchers will find the examples in this study relevant to their own work and use them to help arrive at initial estimates for sample size that are then evaluated fully through additional analyses matched to the researcher's specific study. Use of popular rules-of-thumb for such purposes may be quick and handy, but they are insufficient. Just as clinicians must take an individualized approach to the assessment and treatment of the client in the therapy office, so too, must researchers individualize their sample size planning to the specific model under consideration. The final lesson learned is that determining sample size requirements for SEM necessitates careful, deliberate evaluation of the specific model at hand.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Drs. Dan and Lynda King for their helpful review and commentary on a draft of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a VA Career Development Award to Erika J. Wolf, and by National Institute on Alcohol Abuse and Alcoholism grant K01AA021266 awarded to Shaunna L. Clark.

References

Alhija FNA, Wisenbaker J. A Monte Carlo study investigation the impact of item parceling strategies on parameter estimates and their standard errors in CFA. Structural Equation Modeling. 2006; 13:204–228.

Bentler PM, Chou CH. Practical issues in structural modeling. Sociological Methods & Research. 1987; 16:78–117.

Bollen, KA. Structural equations with latent variables. New York, NY: John Wiley; 1989.

Boomsma, A. Robustness of LISREL against small sample sizes in factor analysis models. In: Joreskog, KG.; Wold, H., editors. Systems under indirection observation: Causality, structure, prediction (Part I). Amsterdam, Netherlands: North Holland; 1982. p. 149-173.

Boomsma A. Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. Psychometrika. 1985; 50:229–242.

Brown, TA. Confirmatory factor analysis for applied research. New York, NY: Guilford Press; 2006.

Cohen, J. Statistical power analysis for the behavioral sciences. 2. Hillsdale, NJ: Erlbaum; 1988.

Cole DA, Maxwell SE. Testing meditational models with longitudinal data: Questions and tips in the use of structural equation modeling. Journal of Abnormal Psychology. 2003; 112:558–577. [PubMed: 14674869]

Davey A, Salva J. Estimating statistical power with incomplete data. Organizational Research Methods. 2009a; 12:320–346.

Davey, A.; Salva, J. Statistical power analysis with missing data: A structural equation modeling approach. New York, NY: Routledge; 2009b.

Dolan C, van der Sluis S, Grasman R. A note on normal theory power calculation in SEM with data missing completely at random. Structural Equation Modeling. 2005; 12:245–262.

Enders CK. The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. Psychological Methods. 2001; 6:352–370. [PubMed: 11778677]

Enders CK, Bandalos DL. The relative performance of full information maximum likelihood estimation for missing data in structural equation models. Structural Equation Modeling. 2001; 8:430–457.

Gagné P, Hancock GR. Measurement model quality, sample size, and solution propriety in confirmatory factor models. Multivariate Behavioral Research. 2006; 41:65–83.

Hancock GR, Lawrence FR, Nevitt J. Type I error and power of latent mean methods and MANOVA in factorially invariant and noninvariant latent variable systems. Structural Equation Modeling. 2000; 7:534–556.

Hertzog C, von Oertzen T, Ghisletta P, Lindenberger U. Evaluating the power of latent growth curve models to detect individual differences in change. Structural Equation Modeling. 2008; 15:541–563.

Jackson DL. Sample size and number of parameter estimates in maximum likelihood confirmatory factor analysis: A Monte Carlo investigation. Structural Equation Modeling. 2001; 8:205–223.

Kelley K, Maxwell SE. Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. Psychological Methods. 2003; 8:305–321. [PubMed: 14596493]

Kim KH. The relation among fit indexes, power, and sample size in structural equation modeling. Structural Equation Modeling. 2005; 12:368–390.

Little RJ, Rubin DB. The analysis of social science data with missing values. Sociological Methods and Research. 1989; 18:292–326.

MacCallum RC, Browne MW, Sugawara HM. Power analysis and determination of sample size for covariance structural modeling. Psychological Methods. 1996; 1:130–149.

MacCallum RC, Lee T, Browne MW. The issue of isopower in power analysis for tests of structural equation models. Structural Equation Modeling. 2010; 17:23–41.

MacCallum RC, Widaman KF, Zhang S, Hong S. Sample size in factor analysis. Psychological Methods. 1999; 4:84–99.

Marcoulides GA. Evaluation of confirmatory factor analytic and structural equation models using goodness-of-fit indices. Psychological Reports. 1990; 67:669–670.

Marcoulides, GA.; Chin, W. You write, but others read: Common methodological misunderstandings in PLS and related methods. In: Abdi, H.; Chin, W.; Vinzi, VE.; Russolillo, G.; Trinchera, L., editors. New perspectives in partial least squares and related methods. Berlin, Germany: Springer-Verlag; in press

Marcoulides GA, Saunders C. PLS: A silver bullet? MIS Quarterly. 2006; 30:iii–ix.

Marsh HW, Hau KT, 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.

Maxwell SE. The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. Psychological Methods. 2004; 9:147–163. [PubMed: 15137886]

Maxwell SE, Kelley K, Rausch JR. Sample size planning for statistical power and accuracy in parameter estimation. Annual Review of Psychology. 2008; 59:537–563.

Muthén, B. Using Mplus Monte Carlo simulations in practice: A note on assessing estimation quality and power in latent variable models. Mplus Web Notes, No 1. 2002. Retrieved from http://www.statmodel.com/download/webnotes/mc1.pdf

Muthén, B.; Asparouhov, T. Using Mplus Monte Carlo simulations in practice: A note on non-normal missing data in latent variable models. Mplus Web Notes, No 2. 2002. Retrieved from http://www.statmodel.com/download/webnotes/mc2.pdf

Muthén, LK.; Muthén, BO. Mplus user's guide. 5. Los Angeles, CA: Muthén & Muthén; 1998–2008.

Muthén LK, Muthén BO. How to use a Monte Carlo study to decide on sample size and determine power. Structural Equation Modeling. 2002; 4:599–620.

Nunnally, JC. Psychometric theory. New York, NY: McGraw-Hill; 1967.

Paxton P, Curran PJ, Bollen KA, Kirby J, Bhen F. Monte Carlo experiments: Design and implementation. Structural Equation Modeling. 2001; 8:287–312.

Satorra A, Saris WE. Power of the likelihood ratio test in covariance structure analysis. Psychometrika. 1985; 50:83–90.

Velicer WF, Fava JL. Effects of variable and subject sampling on factor pattern recovery. Psychological Methods. 1998; 3:231–251.

Appendix A

Annotated Mplus Monte Carlo Script for the First Confirmatory Factor Model Evaluated

```
TITLE: Monte Carlo simulation of 1 factor CFA with 4 items and .50 loadings
MONTECARLO:
NAMES ARE A1-A4;
NOBSERVATIONS = 200; !* specifies sample size
NREPS = 10000;
SEED = 53487;
NGROUPS = 1;
MODEL POPULATION: ! specifies population parameters
F1 BY A1*.50; ! sets factor loadings
F1 BY A2*.50;
F1 BY A3*.50;
F1 BY A4*.50;
F1@1.0; ! specifies factor variance
A1*.75; ! specifies residual variance of the indicators
A2*.75;
A3*.75;
A4*.75;
MODEL : ! specifies model to be tested
F1 BY A1*.50;
F1 BY A2*.50;
F1 BY A3*.50;
F1 BY A4*.50;
F1@1.0; ! this parameter is not estimated as it is set using the “@” symbol
A1*.75;
A2*.75;
A3*.75;
A4*.75;
OUTPUT: TECH9;
*! Is used for annotation purposes; Mplus will not read script beyond the !
```

Annotated Mplus Monte Carlo Script for the First Regressive Structural Equation Model Evaluated

```
TITLE: Monte Carlo simulation for mediation model explaining 16 percent
variance in F3
MONTECARLO: NAMES ARE A1-A3 B1-B3 C1-C3;
NOBSERVATIONS = 200;
NREPS = 10000;
SEED = 53487;
NGROUPS = 1;
MODEL POPULATION:
F1 BY A1-A3*.65;
F2 BY B1-B3*.65;
F3 BY C1-C3*.65;
F1@1.0; ! sets factor variance
F2@.94; ! sets residual factor variance
F3@.84;
[F1-F3@0]; ! sets factor means to 0
A1-A3*.5775;
B1-B3*.5775;
C1-C3*.5775;
F3 ON F1*.25; ! sets regressive path
F3 ON F2*.25;
F2 ON F1*.25;
[A1-A3@0]; ! sets indicator intercepts to 0
[B1-B3@0];
[C1-C3@0];
Model Indirect: F3 IND F2 F1*.06; ! sets magnitude of indirect effect
MODEL:
F1 BY A1-A3*.65;
F2 BY B1-B3*.65;
F3 BY C1-C3*.65;
F1@1.0;
F2@.94;
F3@.84;
[F1-F3@0];
A1-A3*.5775;
B1-B3*.5775;
C1-C3*.5775;
F3 ON F1*.25;
F3 on F2*.25;
F2 on F1*.25;
[A1-A3@0];
```

```
[B1-B3@0];
[C1-C3@0];
Model Indirect: F3 IND F2 F1*.06;
OUTPUT: TECH9;
```

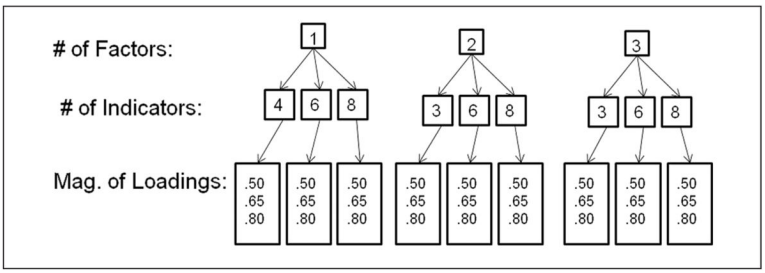

Figure 1. Schematic diagram of CFA model permutations

Note. The figure shows a representation of the model characteristics that were varied in the Monte Carlo analyses of the CFAs. We evaluated models with one to three factors and each factor in the model was indicated by three to eight indicators, which loaded on their respective factors at .50, .65, or .80. The factor correlation(s) was set to r = .30 or .50. CFA = confirmatory factor analysis; mag = magnitude.

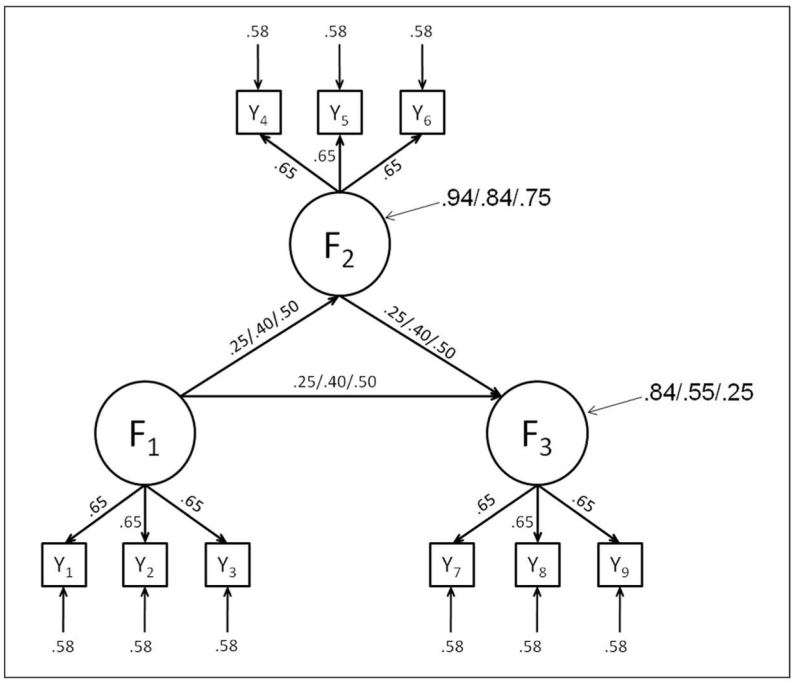

Figure 2. Schematic diagram of mediation model permutations

Note. The strength of all the direct regressive paths was varied from .25 to .40 to .50 to account for 16%, 45%, and 75% of the variance in the dependent variable, respectively. All variables were set to have a variance of 1.0 and mean of 0. Arrows pointing toward the dependent variable show the amount of residual variance in each variable.

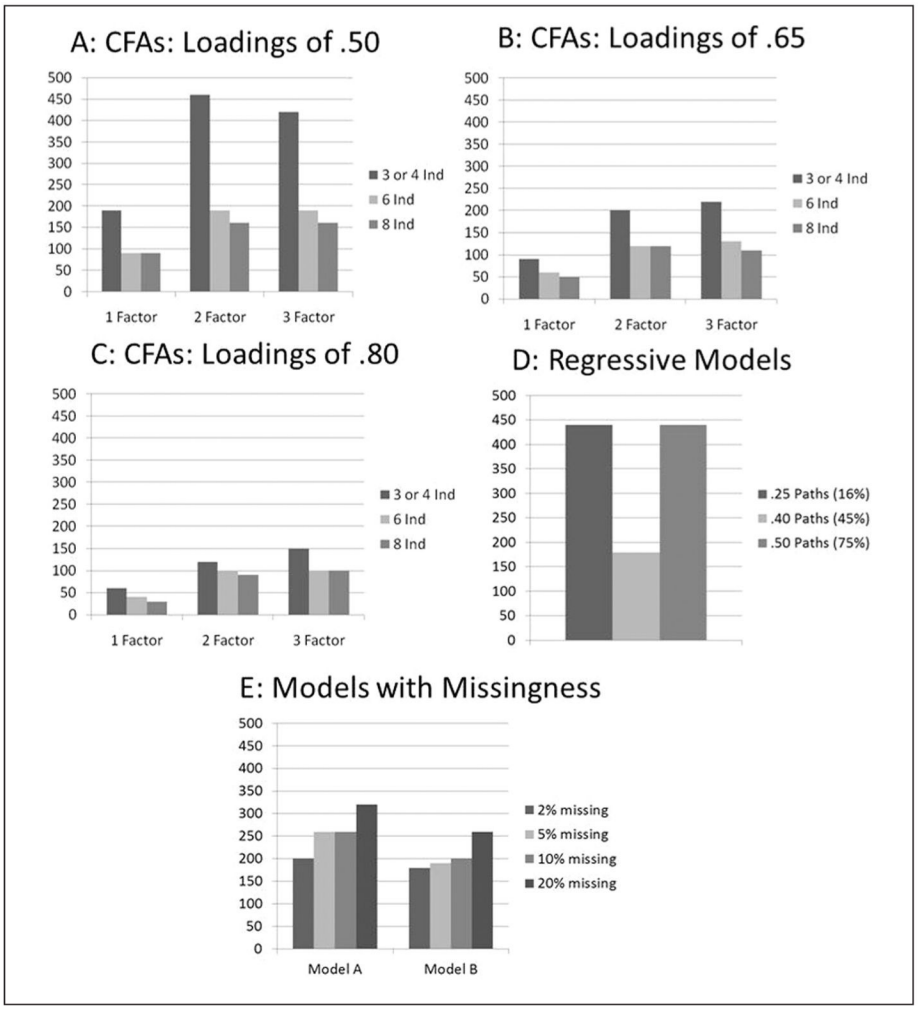

Figure 3. Minimum sample size required for CFA, SEM, and missingness models

Note. Ind = indicator; CFA = confirmatory factor analysis. Panels A to E show the results of the Monte Carlo simulation studies. Each panel shows the minimum sample size required for each permutation of each type of model that was evaluated. For panel E, Model A is the CFA model and Model B is the SEM model. The permutations of the regressive model (i.e., mediation model) differed in both the magnitude of the structural parameters and the total variance explained in the dependent variable (as shown in parentheses).