
(Taylor & Francis/Routledge/Informaworld, UK & USA)

Submission date: April 2, 2013
Abstract

The importance of sample size, although widely discussed in the literature on structural equation modeling (SEM), has not been widely recognized among applied SEM researchers. To narrow this gap, we focus on second language testing and learning studies and examine the following: (a) Is the sample size sufficient in terms of precision and power of parameters in a model using Monte Carlo analysis? (b) How are the results from Monte Carlo sample size analysis comparable with those from the N≥100 rule and from the N:q≥10 (sample size-free parameter ratio) rule? Regarding (a), parameter bias, standard error bias, coverage, and power were overall satisfactory, suggesting that sample size for SEM models in second language testing and learning studies is generally appropriate. Regarding (b), both rules were often inconsistent with the Monte Carlo analysis, suggesting that they do not serve as guidelines for sample size. We encourage applied SEM researchers to perform Monte Carlo analyses to estimate the requisite sample size of a model.

*Keywords:* structural equation modeling, precision, power, sample size, Monte Carlo, second language studies
Introduction

In test validation, structural equation modeling (SEM) is one of the most widely used multivariate analysis methods for examining the nature of the relationships among observed and latent variables often in a confirmatory, hypothesis-testing approach to the data (e.g., Byrne, 2012; Ullman, 2007). As with statistical methods in general, the proper use of SEM requires researchers to check the univariate and multivariate normality of observed variables and to use and report appropriate parameter estimation methods, model-fit indices, missing data treatment, and sufficient sample size. Determining adequate sample size is vital to ensure an acceptable likelihood of obtaining desirable empirical outcomes, specifically, parameter precision and statistical power. Required sample size is related to the precision and power of a study, and these three issues are prevalently documented in SEM methodological literature. Nevertheless, these issues in SEM are not widely known in substantive areas, although SEM has been a method of choice for many researchers in the fields of testing and assessment when conducting construct validation studies of test instruments. We fill this gap by conducting Monte Carlo simulation studies on SEM models used in second language testing and learning research, to examine whether the sample sizes are appropriate in those models in terms of precision and power of parameters. We draw out some implications for improving
Running head: Review of SEM sample size

the methodological approach of applied SEM studies.

Background

Parameter Precision, Power, and Sample Size in the SEM Literature

Sample Size and Surrounding Issues. Determining the required sample size for a study is very important in any research, including validation studies of tests/instruments/questionnaires, because sample size influences many aspects of a study, including the suitability of methods for use (e.g., parametric/nonparametric methods), model fit, and the precision and power of the model’s parameter estimates. The American Psychological Association (2010) stipulates that researchers should “state how this intended sample size was determined (e.g., analysis of power or precision),” “when applying inferential statistics, take seriously the statistical power considerations,” and “routinely provide evidence that the study has sufficient power to detect effects of substantive interest” (p. 30). Similarly, Wilkinson and Task Force on Statistical Inference (1999) argue that researchers should “provide information on sample size and the process that led to sample size decisions” and “document the effect sizes, sampling and measurement assumptions, as well as analytic procedures used in power calculations” (p. 596).

The precision of the model’s parameter estimates refers to the degree to which parameter estimates in a sample model match those in a population model (e.g., Muthén &
Muthén, 2007). Since we want to generalize findings from a sample at hand to the population, we must obtain parameter estimates that reflect the population parameters as closely as possible within a narrow margin of error. Small error shows high precision of a model’s parameter. Statistical power refers to the probability of rejecting a false null hypothesis (i.e., $1 - \beta$ [Type II error]). The power of 0.80—an 80% chance of rejecting a false null hypothesis—is a commonly used yardstick in the social sciences (e.g., Cohen, 1988). In SEM, power relates to the testing of a model (i.e., the sensitivity of $\chi^2$ to identify model misspecification or to compare alternative models) and to the testing of a parameter (i.e., the probability of identifying a parameter value as different from zero).

**Methods to Determine Sample Size.** Against this backdrop, strategies for determining sample size have been discussed in the SEM literature. One rule of thumb is that a sample size below 100 is often considered small, a sample size between 100 and 200 is medium, and a sample size exceeding 200 is large (Kline, 2005); this is referred to as the $N \geq 100$ rule. A similar criterion reported by Ding, Velicer, and Harlow (1995) is that the minimum sample size adequate for analysis is generally 100 to 150 participants. Another approach is to consider model complexity in terms of the ratio of sample size to the number of free parameters needed to be estimated in a model. A minimum sample size would be at least 10 times the number of free model parameters (Raykov & Marcoulides, 2006); this is referred to as the $N:q \geq 10$ rule. For example, a model with 30 free parameters would require
300 observations (30 × 10). Nevertheless, it should be noted that these authors (Ding et al., 1995; Kline, 2005; Raykov & Marcoulides, 2006) emphasize that these are only rough guidelines, and they do not necessarily recommend using them. In fact, while discussing these guidelines, Brown (2006) stressed that they are not dependable and are unlikely to apply to the researcher’s modeling scenario because requisite sample size is a function of numerous factors, including the amount and patterns of missing data, strength of the relationships among the indicators, types of indicators (e.g., categorical or continuous), estimation methods (e.g., [robust] maximum likelihood, robust weighted least squares), and reliability of the indicators. In order to make recommendations on sample size guidelines, Mundfrom, Shaw, and Ke (2005) attempted to consider various factors and conducted a simulation study to investigate the minimum sample size for conducting factor analysis under a variety of conditions. However, the guidelines developed—as the authors admitted—ended up providing too diverse sample sizes depending on the number of common factors, the ratio of variables per factor, and the degree of communality, suggesting the difficulty of creating universal guidelines for use. Thus, it appears impossible to create definitive rules—or even rules of thumb—concerning necessary sample size, because sample size and many variables affect each other in intricate ways (MacCallum & Austin, 2000).

Instead of elaborating general guidelines for sample size, more empirically grounded, individual-model-focused approaches to determining sample size in relation to parameter
precision and power have been proposed. First, Satorra and Saris (1985) suggested specifying a model representing the null hypothesis (incorrect model), a model representing an alternative hypothesis (correct model), and sample size. The only difference between the two models is that a parameter of interest is set to zero (i.e., it is assumed there is no relationship) in the null model whereas it is freely estimated in the alternative model (i.e., it is assumed that there is a relationship). Since the null model for that parameter is false, the difference in the $\chi^2$ value between the two models reflects the degree of misspecification of the parameter. This $\chi^2$ value is the non-centrality parameter of the non-central $\chi^2$ distribution, and is used, along with degrees of freedom and the critical $\chi^2$ value, to calculate the power, in order to detect the model misspecification (see Brown, 2006, for SAS or SPSS syntax). If the power is found to exceed 0.80, the sample size is considered adequate. Second, MacCallum, Browne, and Sugawara (1996) presented a framework of power analysis employing the root mean square error of approximation (RMSEA)—one of the fit indices used to judge the appropriateness of an SEM model. Three null hypotheses are tested: RMSEA $\leq$ 0.05, RMSEA $\geq$ 0.05, and RMSEA = 0.00. An RMSEA value of less than 0.08 is conventionally considered to suggest good fit. Since an RMSEA value of 0.00 (suggesting perfect fit of the model) is highly unlikely, the first two hypotheses are tested against alternative hypotheses of, for example, RMSEA = 0.08 and 0.01, respectively, with sample size, degrees of freedom, and alpha specified (see Preacher & Coffman, 2006). If the power is found to exceed 0.80, the
sample size is considered adequate.

While the methods of both Satorra and Saris (1985) and MacCallum et al. (1996) can test the precision and power of an entire model, they are limited in that they do not indicate the precision and power of individual parameters in a model, nor do they allow modeling various conditions that researchers frequently encounter in their research, such as non-normality or type of indicators. To address these issues, Muthén and Muthén (2002) argue for the use of Monte Carlo analysis, where sample size is estimated under various conditions by taking into account the statistical precision and power of individual parameter estimates and various data conditions. They argue that the Monte Carlo method is flexible in that they can test with pinpoint accuracy the precision and power of individual parameter estimates in a model by taking into account non-normality, type of indicators (i.e., binary, categorical, and continuous), and the amount and patterns of missing data. Similar to the Satorra-Saris method, the Monte Carlo method requires specifying a model with the population (true, correct) values of parameters based on previous studies.

In Monte Carlo analysis, numerous sample datasets of parameters are generated based on those known population parameters. This procedure is iterated a sufficient number of times (for the process of data generation using the Monte Carlo method, see Brown, 2006, pp. 434–437). The results are averaged for each parameter across the samples and are compared to examine divergences (i.e., bias) between the population value and the sample averaged
value. If the results are unsatisfactory and reveal less precise and/or underpowered parameter estimates, researchers can choose to revise a model or to conduct another Monte Carlo study with a larger sample size to investigate how many more samples are required to obtain the desired level of precision and power of parameters. According to Barrett (2007), Brown (2006), and McIntosh (2007), the Monte Carlo approach is currently the best way to evaluate sample size in SEM.

Parameter Precision, Power, and Sample Size in Applied SEM Studies

Although precision, power, and sample size have been consistently discussed in the SEM literature, their application to substantive SEM studies has only recently begun. A review of studies in second language testing and learning research shows that 63 SEM articles have been published (for details, see the Results section of this paper). Of these, only three models reported sample size justification: Eckes and Grotjahn (2006) and Tseng, Dörnyei, and Schmitt (2006) used the N:q≥10 rule; Csizér and Kormos (2009) used Kline’s (2005) criterion of a sample size exceeding 200 as large and sufficient. The picture is similar in the International Journal of Testing: Of 45 SEM articles that have been published (a list of studies is available upon request), two referred to sample size justification: Byrne, Stewart, Kennard, and Lee (2007) used Kim’s (2005) procedures for estimating the adequacy of sample size in examining a multigroup model. The other article was not an applied SEM but
Running head: Review of SEM sample size

a simulation study intended to provide guidelines for sample size: Mundfrom et al. (2005) investigated the minimum sample size for conducting factor analysis under a variety of conditions, only to find that requisite sample size greatly differed depending on the number of common factors, the ratio of variables per factor, and the degree of communality. Lack of attention to sample size in substantive SEM studies warrants investigation into sample size in those studies, especially in terms of the precision and power of parameters.

Current Study

Given the widespread use of SEM and the hitherto insufficient attention paid to sample size in terms of precision and power of parameters in substantive SEM studies, it would be of great interest to address these issues for SEM models. We undertake this task in the case of second language testing and learning research, where we conduct Monte Carlo simulation studies to examine whether sample size for published SEM models is sufficient from the viewpoints of precision and power, and to compare Monte Carlo findings with rule-of-thumb criteria for sample size. We examine two research questions: (a) Is the sample size sufficient in terms of precision and power of parameters in a model using Monte Carlo analysis? (b) How are the results from the Monte Carlo sample size analysis comparable with those from the N\(\geq\)100 rule and from the N:q\(\geq\)10 (sample size-free parameter ratio) rule?

To answer Research Question 1, we examine the relationship between sample size and
Running head: Review of SEM sample size

parameter estimation in actual, published SEM models in the field of second language testing and learning. As this implies, we are more interested in the appropriateness of sample sizes and parameter estimates of the substantive models in the field, not the empirical evaluation of requisite sample sizes of particular types of model or fit indices. Thus, while we examine the relationship between sample size and parameter estimation—a topic also addressed in previous studies—our interest and the purpose of our study differ from those of previous studies. More specifically, some experimental simulation studies have systematically manipulated variables including model complexity, factor loading magnitude, degree of misspecification, normality, missing data, and sample size, and investigated their effects on fit indices or estimators (Bentler & Yuan, 1999; Curran, Bollen, Paxton, Kirby, & Chen, 2002; Enders & Bandalos, 2001; Fan, Thompson, & Wang, 1999; Lei, 2009; Nevitt & Hancock, 2004; Savalei, 2010). Others have examined sampling variability in a model fit index and ensuing model selection uncertainty (Preacher & Merkle, 2012) or sample size requirements for particular types of model, including growth mixture models (Kim, 2012). These issues are not addressed in the current research.

Regarding Research Question 2, the appropriateness of the N≥100 rule and the N:q≥10 rule does not seem to have been examined, although these rules-of-thumb are widely used among SEM practitioners. We assess this appropriateness below.

Method
Article Collection

We searched for studies using SEM in second language testing and learning in June 2011, in two ways. First, we retrieved studies in the following 20 representative journals that are internationally available, through ERIC, LLBA, and search engines available at journal homepages: *Annual Review of Applied Linguistics*, *Applied Language Learning*, *Applied Linguistics (AL)*, *Assessing Writing*, *ELT Journal*, *Foreign Language Annals*, *International Journal of Applied Linguistics*, *International Review of Applied Linguistics in Language Teaching*, *Language Assessment Quarterly (LAQ)*, *Language Learning (LL)*, *Language Learning & Technology*, *Language Teaching*, *Language Teaching Research*, *Language Testing (LT)*, *Modern Language Journal (MLJ)*, *RELC Journal (RELC)*, *Second Language Research*, *Studies in Second Language Acquisition (SSLA)*, *System*, and *TESOL Quarterly (TQ)*. We used the following keywords independently: *causal analysis (analyses)*, *causal model(s)*, *causal modeling (modelling)*, *confirmatory factor analysis (analyses)*, *covariance structure(s)*, *covariance structure analysis (analyses)*, *covariance structure model(s)*, *simultaneous equation model(s)*, *simultaneous equation modeling (modelling)*, *structural equation model(s)*, and *structural equation modeling (modelling)*. We compiled these keywords based on the keywords and synonyms supplied in books, articles, authors’ experiences, and feedback from colleagues. We used abstract, title, and article keyword searches without imposing a date range restriction. Although SEM includes path and
regression analyses where, by definition, observed variables are focused on, we excluded
studies using such analyses from our search, because SEM is often viewed as a method for
incorporating both observed and unobserved variables and correcting for measurement error
(Byrne, 2012). Second, we performed manual searches in these journals to cross-check the
studies electronically identified. We further inspected the reference list of empirical,
theoretical, and review papers for additional relevant materials.

The literature search was restricted to internationally available, published journal
articles, since we intended to examine precision and power of parameter estimates in SEM
models in representative journals. However, this could produce bias if we were to generalize
the findings beyond those journals. Such generalization requires careful interpretation of the
findings.

Analyses

**Research Question 1.** We collected from each published article the following
information necessary for parameter precision, power, and sample size estimates: the number
of observed and latent variables, sample size, parameters including (preferably
unstandardized) path coefficients, covariances, and measurement errors. When parameters
were only partially reported, the model was re-estimated using available variance/covariance
or correlation matrices. When matrices or the aforementioned pieces of information were
missing, we requested them by e-mailing (co-)authors and explaining the purpose of our research. When we did not receive a response, we did not send a reminder.

We estimated precision and power for each parameter in a model using Mplus (Version 6.12; Muthén & Muthén, 2011). Given no access to raw data, we could not take into account univariate or multivariate normality or missing data. Models based on non-normal data and/or missing data in the primary studies were run as if they contained normal and full data. This was because Mplus requires the creation of a mixture of two subpopulations or classes of individuals (e.g., normal vs. outlier groups) to analyze non-normal data, and the specification of the patterns of missing data to analyze models with missing data. The analyses of non-normal data and/or missing data are only possible when raw data are available or when researchers have knowledge about the situation under which the primary study was conducted, neither of which was available or known in secondary studies like the current one. Parameter precision, power, and sample size estimation was therefore conducted in all cases using maximum likelihood methods. Additionally, when unstandardized parameters were unavailable or not reported, standardized ones were used as population

---

1Of the 18 models we analyzed (see the Results section), five were reported to have non-normal data: four of the five models were analyzed by Satorra-Bentler robust maximum likelihood (ML) methods and one was analyzed by ML methods. These five models also had missing data and the authors deleted them listwise. There was another model analyzed by full maximum ML methods while containing missing data. The remaining 12 models were unclear as to normality and/or missing data. Nevertheless, since the information we used in Monte Carlo studies is parameter values and not the standard errors of the parameters or fit indices, only the latter two of which are influenced by non-normality, our findings would be robust even if those 12 models had been based on non-normal data. On the other hand, missing data affect parameter estimates (Allison, 2003) and our inability to model the pattern of data gaps, if any, is the limitation of the current study.
Running head: Review of SEM sample size

parameter values. This would probably be close enough for parameter precision, power, and sample size estimates (personal communication, Linda Muthén, May 29, 2011).

We followed Muthén and Muthén (2002) to estimate precision and power for each parameter in a model. We specified the parameters of the population model, and generated 10,000 samples (replications) in each run. Results over the 10,000 replications were summarized. Note that the population model refers to the model reported in each article—we viewed these published models as the correct, true model. A run was considered invalid when the following two situations arose: First, the number of unusable data generated—data showing non-positive-definite matrices and/or non-convergence—was more than 5% of the total replications (i.e., 500 = 10,000*0.05); second, the population variance/covariance matrix that was used as an input was found to be not positive-definite itself, resulting in generating no sample (replication).

To determine if the sample size for a model is sufficient in terms of precision and power, precision of parameter estimates was examined using four criteria, and power using one criterion, following Muthén and Muthén (2002). First, parameter bias and standard error bias should not exceed $|10\%|$ for any parameter in the model. Second, the standard error bias for the parameter for which power is of particular interest should not exceed $|5\%|$. Third, 95% coverage—the proportion of replications for which the 95% confidence interval covers the population parameter value—should fall between 0.91 and 0.98. One minus the coverage
value equals the alpha level of 0.05. Coverage values should be close to the correct value of 0.95. Next, power was evaluated as to whether it exceeded 0.80—a commonly accepted value for sufficient power (e.g., Cohen, 1988). Results based on these five criteria—parameter bias and standard error bias, standard error bias for the parameter of interest, 95% coverage, and power—were first calculated for each parameter for each model. Then the results for each criterion were averaged across models. These four criteria are all important because standard errors, parameter coverage, and the power of detecting effects may be questionable at small sample sizes despite good precision of parameter estimates (Muthén, 2000).

Table 1 shows an example of the Mplus output excerpt for Woodrow’s (2006a) model of motivation that consisted of three intercorrelated factors of performance approach, performance avoidance, and task. Columns 2 and 3 show population and sample parameters. Using these values, the parameter bias for observed variable X1 was calculated as follows:

\[ \frac{|0.6781 - 0.68|}{0.68} \times 100 = 0.279\% \]

This was far below the criterion of 10%, suggesting good estimation of the parameter. Column 4 shows the standard deviation of the parameters across 10,000 replications. Column 5 shows the average of the standard errors across replications. The standard error bias for X1 was

\[ \frac{|0.0556 - 0.0565|}{0.0565} \times 100 = 1.593\% \]

This was far below the criterion of 10%, suggesting good estimation of the parameter. Column 6 provides the mean square error of parameter estimates, which equals the variance of the estimates across replications plus squared bias (Muthén & Muthén, 2007).
Running head: Review of SEM sample size

Column 7 shows coverage, or the proportion of replications where the 95% confidence interval covers the true parameter value. The value of 0.946 for X1 was very close to 0.95, suggesting good estimation of the parameter. The last column shows the percentage of replications for which the parameter is significantly different from zero (i.e., the power estimate of a parameter). The power for X1 was 1.000, which exceeded 0.80 and suggests sufficient power for the parameter. These results altogether provide favorable evidence for parameter precision and power for observed variable X1 and suggest that the sample size for X1 is sufficient. The same process was repeated for the remaining parameters. Appendix A shows the Mplus syntax used for analyzing Woodrow’s (2006a) model.

[Insert Table 1 about here]

**Research Question 2.** For each SEM model, the sample size reported in a primary study was coded. The number of free parameters in a model was calculated by drawing a model with Amos (Version 18.0.0; Arbuckle, 2009) and entering (a) the corresponding correlation (with standard deviations) or variance/covariance matrix, or (b) the corresponding size (row × column) of a hypothetical matrix for a model when (a) was not available. The number of free parameters in each model was multiplied by 10, following Raykov and Marcoulides (2006).
Across the research questions, we considered the article as a unit of analysis. When one article included multiple models, we coded information on the final model. When one article included several final models (i.e., it investigated multiple research questions, each of which was answered with a final model, resulting in more than one final model in a single article), we selected one of the final models randomly.

**Results and Discussion**

**Preliminary Analysis**

Table 2 shows that 63 articles using SEM were found in nine of the 20 journals we reviewed. Forty-two articles out of 63 (67%) contained enough information for parameter precision, power, and sample size estimates. When we e-mailed the authors of these remaining 21 articles missing information to ask for these estimates, three authors responded, and one of them provided information. Although not shown in Table 2, articles often lacked information on part of or the whole parameter estimates, such as factor loadings from latent to observed variables, covariances among measurement errors, and covariance or correlation matrices that would have been used to reproduce models. In total, 43 articles—and 43 SEM models therein—were analyzed with Mplus. Those 63 articles are marked with asterisks (*) in the references, among which the 43 articles that provided information for parameter precision, power, and sample size estimates, are marked with double asterisks (**)
During preliminary analyses, 25 of the 43 models (58%) faced problems with either model un-identification (1 model), non-positive-definite matrices (23 models), or the categorical nature of data (1 model). This last-named categorical model was a single-group, single-time point model, and Monte Carlo simulation studies are not possible for such a model. Only in a multiple group or multiple-time point models, residual variances for categorical outcomes are identified, which enables Monte Carlo simulation studies (Linda Muthén, personal communication, June 3, 2011). In the end, the remaining 18 models without analytical problems were the sources of parameter precision, power, and sample size estimates. They were Chen, Warden, and Chang (2005), Choi, Kim, and Boo (2003), Csizér and Dörnyei, (2005), Eckes and Grotjahn (2006), Fushino (2010), Gorsuch (2000), Llosa (2007), Ockey (2011), Peng and Woodrow (2010), Purpura (1997), Shin (2005), Tseng, Dörnyei, and Schmitt (2006), Tseng and Schmitt (2008), Woodrow (2006a, 2006b), Xi (2005, 2010), and Yashima (2002). Note that analyses for Research Question 2 also included the 25 models that were excluded and the 20 models that lacked information on parameters—both of which were excluded from Research Question 1, in order to examine the representativeness of the 18 models that remained in the final analysis.
Research Question 1: Is the Sample Size Sufficient in Terms of Precision and Power of Parameters in a Model Using Monte Carlo Analysis?

Table 3 shows the descriptive statistics of precision of parameter estimates and power across the 18 models, which included 301 parameters. The average parameter bias and standard error bias were 0.355% and 1.022% respectively—far below the criterion of 10%—suggesting good estimation of parameters and standard errors in SEM models. Nevertheless, one parameter estimate in one model was problematic, showing a parameter bias of 10.125%. Although the data are not shown in Table 3, a close look shows that this bias came from the divergence in the population parameter and the model parameter (–0.08 and –0.0881). This can be negligible in practice. Further, the bias percentage of 10.125 was very close to the criterion of 10%. Thus, overall, few concerns were found for the parameter bias and standard error bias.

Further, the average standard error bias of parameters of particular interest (1.022%) was well below 5% for each parameter in the models. The average coverage value was also satisfactory as the mean (0.948) was very close to the correct value of 0.95 with a negligible
standard deviation (0.004), and all the values were within 0.91 and 0.98. Finally, the average power (0.962) exceeded 0.80, suggesting overall a sufficient level of power across models. However, the smallest power estimate was 0.057, and there were 17 parameters with power below 0.80 across nine models in total. Although not shown in Table 3, these 17 parameters were small in size (M = 0.167) and power (M = 0.459), suggesting less than 50% probability of identifying a parameter value as different from zero.

In summary, the parameter bias was found in only one of the 301 parameters, no standard error bias was found, coverage was satisfactory in all cases, and power was lacking in only 17 of the parameters (5.648%). Thus, it would be reasonable to argue that sample size for SEM models used in the second language testing and learning research that were analyzed in the current study, is overall adequate enough for these statistics to be well-behaved. Nevertheless, close analysis of these 17 underpowered parameters that were found in the nine models would be worthwhile. In the primary studies, nine were reported to be statistically significant, six non-significant, and two unclear as to significance due to lack of information reported. Since these parameters were all very small in size and power, the results—regardless of significance or non-significance in the primary studies—would encourage caution concerning the accuracy of these models. In particular, it is possible that the underpowered parameters reported as statistically significant were significant by chance and that the underpowered parameters reported as statistically non-significant would have
been found to be significant if the sample sizes had been larger. Since these 17 parameters were of substantive interest and related to research questions in the primary studies, caution should be exercised in the interpretation of the parameters in question.

Although we have argued that sample size in SEM models used in second language testing and learning research that were analyzed in the current study, is overall adequate, the 17 underpowered parameters across the nine models lead to the question of how much larger samples would have been needed for these parameters to be well-powered. The original sample sizes in these nine models and the optimal sample sizes we calculated through power analysis are shown in Table 4. An original sample size was increased by 50 until all underpowered parameter(s) in the model reached power over 0.80. For example, the sample size for Model D \((N = 275)\) was tested for 275, 325, 375, and 425.

[Insert Table 4 about here]

The underpowered parameters ranged from \(-0.080\) to 0.626 in size and from 0.057 to 0.746 in power. In general, the smaller the parameter in size, the larger is the requisite sample size in order to achieve an adequate power: Model H included the smallest parameter size of \(-0.023\) and required 31,832 samples—23 times larger than the original sample size—to test the significance of the parameter. Model B included the largest three parameters of 0.626,
0.515, and 0.361 along with one of the smallest parameters of –0.026, resulting in the requisite sample size of 54,269—247 times larger than the original sample size—to identify those parameters as distinguishable from zero.

**Research Question 2: How Are the Results from Monte Carlo Sample Size Analysis Comparable with those from the N≥100 Rule and from the N:q≥10 Rule?**

Table 5 shows a comparison of the sample sizes of the SEM models, using Kline’s (2005) N≥100 guidelines. Among the 18 models, only one model was found to be based on a small sample size (99 or less), while another model had a medium sample size (100–199) and the remaining 16 models included large sample sizes (200 or above). Note that we also show in Table 5 the results for the 63 models and the 43 models, because we were interested in the degree of the representativeness of the 18 models in comparison with the 63 and 43 models: if models in each category—small, medium, and large—were proportionally similar across three datasets of sample sizes—63, 43, and 18—we would be confident with the generalizability of the current finding. For the result from the 18 models, a lesser proportion of models was found in the small and medium categories than that in the 64 and 43 models. For example, in the case of a small sample size, 5.556% of the 18 models belonged to the small category in contrast to 11.111% of the 63 models and 11.628% of the 43 models. These results show that the results from the 18 models tend to underrepresent models in small and
medium sample sizes. This limits the generalizability of the current finding to the whole family of SEM models in second language testing and learning research. Since the result from the 63 models can be considered to represent the whole family of models in second language testing and learning research, we conclude that 90% of the SEM models in this field had an adequate sample size of 100 or above when judged by Kline’s (2005) N≥100 guidelines.

Table 6 shows the descriptive statistics of actual and minimum desired sample sizes calculated based on Raykov and Marcoulides (2006) for three datasets of sample sizes (63, 43, and 18 models). We interpreted the median rather than mean, standard deviation, and mode, because mean and standard deviation were misleading given the large skewed and kurtotic distribution of actual sample sizes and because the mode was based only on two or three models with the same sample size. The median sample size of the 18 models was 296.5, which exceeded 200 and is thus considered a large sample size according to Kline (2005). The smallest sample size was 93.
Table 6 also shows that the median number of free parameters that needed to be estimated in each model was 27 for the 18 models. The minimum desirable sample size based on the number of free parameters was 270 (27 × 10). A comparison of the actual (296.5) and minimum desirable sample size (270) across the 18 models shows that, on average, the sample size used for 18 SEM models was sufficient. However, a closer look at each model shows that four (22.222%) of the 18 models had sample sizes that were less than the minimum sample size desired by the number of free parameters. Further, for the 63 and the 43 models, the median of the actual sample size was below the minimum desirable sample size (275 versus 320; 275 versus 310). A closer analysis shows that 34 (53.968%) of the 63 models and 21 (48.837%) of the 43 models had sample sizes that were less than the minimum sample size desired by the number of free parameters. Based on the results of the 63 models, we conclude that approximately half of the SEM models in the field had an adequate sample size when judged according to the N:q≥10 guidelines of Raykov and Marcoulides (2006).

More importantly, recall that nine models were found to include underpowered parameters, using Monte Carlo analysis (Models A to I; see Table 4); one of the 18 models model had a small (N < 100) sample size (Model J; see Table 5); and four of the 18 models had a sample size-free parameter ratio of less than 10 (Models A, D, J, and L; see Table 6). A comparison of these three criteria for sample size is summarized in Table 7. It shows that the
results were not always consistent across the criteria. For example, Model C lacked sample size in terms of power, while it was adequate in terms of the $N \geq 100$ rule and the $N:q \geq 10$ rule. Of the 18 models, seven models (38.889%) were categorized as having sufficient sample sizes across the three criteria, whereas none of the models were consistently found to include insufficient sample size across the criteria. Comparison of the Monte Carlo analysis against the $N \geq 100$ rule and the $N:q \geq 10$ rule respectively is noteworthy. Comparison of the Monte Carlo analysis and the $N \geq 100$ rule shows that ten models (44.444%) were judged as having either sufficient or insufficient sample size according to the two guidelines: nine models (50.000%) were judged as having insufficient sample size using Monte Carlo analysis but as sufficient by the $N \geq 100$ rule; one model (5.556%) was judged as having sufficient sample size using Monte Carlo analysis but as insufficient by the $N \geq 100$ rule. The results for judging half of the 18 models as inadequate using Monte Carlo analysis, but as adequate by the $N \geq 100$ rule, suggest that when results from the two guidelines diverge, the $N \geq 100$ rule is more lenient than the Monte Carlo analysis. To the extent to which Monte Carlo analysis can accurately incorporate various conditions into models when estimating sample size, we can conclude that the $N \geq 100$ rule may suggest that the model has a sufficient sample size although in fact it does not.

[Insert Table 7 about here]
Similarly, comparison of Monte Carlo analysis and the N:q≥10 rule shows that they produced the same results for half of the cases (nine models; 50.000%), whereas they produced divergent results for the other half of the cases. For divergent results, the N:q≥10 rule provided more lenient or more severe results. Seven models (38.889%) were categorized as inadequate using Monte Carlo analysis, but as adequate by the N:q≥10 rule. Two models (11.111%) were categorized as adequate using Monte Carlo analysis, but as inadequate by the N:q≥10 rule.

**Summary and Discussion**

Regarding Research Question 1 on the adequacy of sample size in terms of the precision and power of parameters in an SEM model, Monte Carlo simulation results showed that parameter bias was observed in only one of the 301 parameters across 18 models, no standard error bias was found, coverage was adequate in all cases, and power was lacking (less than 0.80) in only 17 of the 301 parameters (5.648%). These favorable results for parameter bias, standard error bias, and coverage suggest the high precision of parameters. The mostly favorable results for power of parameters suggest a high probability of identifying a parameter value as different from zero. It follows that sample size for these SEM models may be said to be overall adequate. One caution is that these results are limited to the studies and
models that could be analyzed in terms of precision and power. Twenty-five models had statistical problems in the preliminary analyses and had to be removed from the final analysis, and other 20 models failed to report necessary information for Monte Carlo analysis. These models tended to have smaller sample size. If these models had been included, sample size in second language testing and learning studies might not have been sufficient.

Considering Research Question 2 in the light of the comparison of results from the Monte Carlo sample size analysis with those from the N≥100 rule and from the N:q≥10 rule, some inconsistencies were observed across the criteria. In particular, we obtained two relevant findings. First, the N≥100 rule was more lenient than the Monte Carlo analysis, suggesting that models considered on the basis of the former to have a sufficient sample size may actually need a larger sample. Second, the N:q≥10 rule was either more lenient or severe compared with the Monte Carlo analysis. These results suggest that both these rules may produce results different from those of Monte Carlo analysis and cannot always serve as guidelines for sample size.

Recall that only three of the 63 SEM articles in second language testing and learning research were reported with sample size justification by referring to either the N≥100 rule or the N:q≥10 rule. Based on the literature and our current findings, we argue that a Monte Carlo approach provides more accurate estimates of sample size than the two rules, as it allows researchers to estimate requisite sample size under various conditions and to indicate
parameter and standard error biases, 95% coverage, or power of each parameter. To be specific, Monte Carlo analysis differs from the two rules in its ability to model various variables, including number of free parameters to be estimated, factor loadings, covariances, and measurement error variances, among others, whereas the two rules are merely rules-of-thumb. In other words, requisite sample size in Monte Carlo analysis is estimated based on a larger amount of information on the characteristics of a model of interest, making sample size estimates under Monte Carlo analysis more informative than those under the two rules-of-thumb. We have also shown that a Monte Carlo approach is practical and applicable to each research context, prompting researchers to pay more attention to sample size in SEM models and examine it using this approach.

Additionally, the smallest sample size among the 18 models was 93. Although parameter precision and power for this model were all acceptable—note that this model (Model J) does not appear among underpowered-parameter models in Table 4—we do not necessarily recommend the use of small-sample SEM models. Small sample size often causes convergence failures and may not represent a population well (e.g., Brown, 2006). Thus, despite favorable results for parameter precision and power for a model, we would be better off having larger samples when resources allow.

Power analyses on the 17 underpowered parameters showed that huge sample sizes were required to raise the power of the parameters to 0.80 to reject zero effects. Some of
these estimates for sample size—although accurate and desirable for testing the statistical significance of parameters—would obviously cause logistic concern and make a study difficult to conduct. Although underpowered parameters, significant or non-significant, can be retained in a model if they are substantively sensible or if the overall model fit is appropriate, caution is necessary in interpreting findings.

Based on these results, we recommend conducting Monte Carlo analysis on a model of interest in the planning phase of a study. Researchers can do this, without raw data, by writing syntax specifying the number of observed and latent variables, sample size, and parameter estimates derived by theory or previous studies. If underpowered parameters are identified, we need to conduct a series of analyses by systematically increasing sample sizes to find out the sample size sufficient for these parameters to exceed the power of 0.80. If researchers cannot obtain large enough sample sizes given budget or ethical considerations, we recommend using Monte Carlo analysis to simulate how the requisite sample size is a function of factors such as the number of parameters in a model, and the number of missing data. To address this issue, Muthén’s (2007) post to the Mplus discussion list is useful: He states that critical factors in determining sample size in Mplus Monte Carlo simulation are the categorical/continuous outcome variables, the skewness of the outcomes, the number of parameters in a model, and the number of missing data. Muthén and Muthén (2002) illustrate the effect of these factors by showing that the sample size requirement for a 10-item
confirmatory factor analysis model is 150 for normal and complete data, 175 for normal and missing data, 265 for non-normal and complete data, and 315 for non-normal and missing data. Further, larger sample size is required for a model with categorical, skewed variables while containing a greater number of parameters that need to be estimated (Brown, 2006). A model with continuous and normally distributed variables and a complete dataset, while containing a small number of parameters, is expected to require a smaller sample size than a model with categorical, non-normal, and incomplete data that include many parameters to estimate. It would be advisable to change the design of a study or consider alternative structures for the model, provided that such an altered model would still serve the purpose of the study well.

While an a priori use of Monte Carlo analysis to estimating sample size is preferable, a posthoc analysis is still possible. In fact, the limitation of this study includes the posthoc approach to parameter precision, power, and sample size estimates. We are aware that an a priori (prospective), not posthoc (retrospective), approach would be better with plausible population values supplied by theory. However, this is not always easy in practice—the related literature may report only part of the model parameters, offer little help in hypothesizing plausible values particularly in new areas, or a pilot test cannot be conducted due to logistic issues. In those cases, we are best served by using values from previous studies as population parameter values for posthoc parameter precision, power, and sample size
estimates. This is what we did in the present study. The Monte Carlo approach can be used regardless of a priori or posthoc issues, and an a priori analysis is better if conditions allow. When that is difficult, a posthoc analysis would still be useful as it provides invaluable insight into the quality of a model.

The Monte Carlo approach to parameter precision, power, and sample size estimates is now readily available thanks to the Monte Carlo facilities implemented in Mplus. A more accurate estimation of these parameters can prove useful in grant applications and thesis proposals. We encourage applied SEM researchers in the area of testing and assessment to perform Monte Carlo analyses more often to check the quality of the models in a study.

**Acknowledgement**
Running head: Review of SEM sample size

**References**


**Arbuckle, J. L. (2009). Amos (Version 18.0.0) [Computer software]. Crawfordville, FL: Amos Development Corporation.**


Running head: Review of SEM sample size


**Carr, N. T. (2006).** The factor structure of test task characteristics and examinee
Running head: Review of SEM sample size


35


**Fushino, K. (2010). Causal relationships between communication confidence, beliefs about group work, and willingness to communicate in foreign language group work. TESOL Quarterly, 44, 700–724.


**Gorsuch, G. J. (2000). EFL educational policies and educational cultures: Influences on
Running head: Review of SEM sample size


Running head: Review of SEM sample size

225–250.


Carlo simulation [Online forum comment]. Retrieved from


Muthén, B. O. (2007, May 20). Mplus discussion > Structural equation modeling > Sample

size [Online forum comment]. Retrieved from


Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample


Muthén & Muthén.


Angeles, CA: Muthén & Muthén.


statistics in structural equation modeling. *Multivariate Behavioral Research*, 39,

439–478.

**Ockey, G. (2011). Self-consciousness and assertiveness as explanatory variables of L2 oral


measuring anxiety at different stages of the foreign language learning process: The input anxiety scale, the processing anxiety scale, and the output anxiety scale. *Language Learning*, 50, 87–117.


Running head: Review of SEM sample size

47, 289–325.


Shin, S.-K. (2005). Did they take the same test? Examinee language proficiency and the


**Vandergrift, L., Goh, C. C. M., Mareschal, C. J., & Tafaghodtari, M. H. (2006).** The
Running head: Review of SEM sample size


**Xi, X. (2010). Aspects of performance on line graph description tasks: Influenced by graph familiarity and different task features. Language Testing, 27, 73–100.**


*Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses?
Appendix A

*Mplus Input for the Monte Carlo Analysis to Determine the Precision and Power of Parameters for Woodrow (2006a)*

**TITLE:** CFA THREE-FACTOR, NORMAL DATA, NO MISSING

**MONTECARLO:**

NAMES ARE X1-X15;
NOBSERVATIONS = 275; ! SAMPLE SIZE
NREPS = 10000; ! NUMBER OF REPLICATIONS
SEED = 53567;

MODEL POPULATION:
Approach BY X1*.68 X2*.71 X3*.75 X4*.84 X5*.72;
Avoid BY X6*.63 X7*.72 X8*.50 X9*.63 X10*.67;
Task BY X11*.60 X12*.66 X13*.48 X14*.49 X15*.58;
Approach@1; Avoid@1 Task@1;
X1*.53; X2*.50; X3*.44; X4*.29; X5*.48; X6*.60; X7*.48; X8*.75;
X9*.60; X10*.55; X11*.64; X12*.56; X13*.77; X14*.49; X15*.66;
Approach WITH Avoid*.59; Approach WITH Task*.17; Avoid WITH Task*.31;

MODEL:
Approach BY X1*.68 X2*.71 X3*.75 X4*.84 X5*.72;
Avoid BY X6*.63 X7*.72 X8*.50 X9*.63 X10*.67;
Task BY X11*.60 X12*.66 X13*.48 X14*.49 X15*.58;
Approach@1; Avoid@1 Task@1;
X1*.53; X2*.50; X3*.44; X4*.29; X5*.48; X6*.60; X7*.48; X8*.75;
X9*.60; X10*.55; X11*.64; X12*.56; X13*.77; X14*.49; X15*.66;
Approach WITH Avoid*.59; Approach WITH Task*.17; Avoid WITH Task*.31;

**ANALYSIS:** ESTIMATOR = ML;
**OUTPUT:** TECH9;
Table 1

*Mplus Output for the Monte Carlo Analysis to Determine the Precision and Power of Parameters for Woodrow (2006a)*

<table>
<thead>
<tr>
<th>Population parameter</th>
<th>Sample parameters averaged</th>
<th>SD of sample parameters</th>
<th>Standard error of sample parameters</th>
<th>Mean square error of parameter s</th>
<th>95% Coverage</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Approach By</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X1</td>
<td>0.68</td>
<td>0.6781</td>
<td>0.0565</td>
<td>0.0556</td>
<td>0.0032</td>
<td>0.946</td>
</tr>
<tr>
<td>X2</td>
<td>0.71</td>
<td>0.7074</td>
<td>0.0550</td>
<td>0.0553</td>
<td>0.0030</td>
<td>0.952</td>
</tr>
<tr>
<td>X3</td>
<td>0.75</td>
<td>0.7480</td>
<td>0.0539</td>
<td>0.0541</td>
<td>0.0029</td>
<td>0.949</td>
</tr>
<tr>
<td>X4</td>
<td>0.84</td>
<td>0.8381</td>
<td>0.0510</td>
<td>0.0513</td>
<td>0.0026</td>
<td>0.949</td>
</tr>
<tr>
<td>X5</td>
<td>0.72</td>
<td>0.7179</td>
<td>0.0552</td>
<td>0.0548</td>
<td>0.0031</td>
<td>0.945</td>
</tr>
<tr>
<td><strong>Avoidance By</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X6</td>
<td>0.63</td>
<td>0.6292</td>
<td>0.0599</td>
<td>0.0600</td>
<td>0.0036</td>
<td>0.949</td>
</tr>
<tr>
<td>X7</td>
<td>0.72</td>
<td>0.7189</td>
<td>0.0583</td>
<td>0.0582</td>
<td>0.0034</td>
<td>0.950</td>
</tr>
<tr>
<td>X8</td>
<td>0.50</td>
<td>0.4991</td>
<td>0.0637</td>
<td>0.0627</td>
<td>0.0041</td>
<td>0.945</td>
</tr>
<tr>
<td>X9</td>
<td>0.63</td>
<td>0.6279</td>
<td>0.0598</td>
<td>0.0600</td>
<td>0.0036</td>
<td>0.954</td>
</tr>
<tr>
<td>X10</td>
<td>0.67</td>
<td>0.6684</td>
<td>0.0588</td>
<td>0.0593</td>
<td>0.0035</td>
<td>0.953</td>
</tr>
<tr>
<td><strong>Task By</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X11</td>
<td>0.60</td>
<td>0.5983</td>
<td>0.0659</td>
<td>0.0650</td>
<td>0.0043</td>
<td>0.947</td>
</tr>
<tr>
<td>X12</td>
<td>0.66</td>
<td>0.6582</td>
<td>0.0659</td>
<td>0.0644</td>
<td>0.0043</td>
<td>0.946</td>
</tr>
<tr>
<td>X13</td>
<td>0.48</td>
<td>0.4788</td>
<td>0.0661</td>
<td>0.0667</td>
<td>0.0044</td>
<td>0.949</td>
</tr>
<tr>
<td>X14</td>
<td>0.49</td>
<td>0.4881</td>
<td>0.0560</td>
<td>0.0559</td>
<td>0.0031</td>
<td>0.949</td>
</tr>
<tr>
<td>X15</td>
<td>0.58</td>
<td>0.5787</td>
<td>0.0661</td>
<td>0.0651</td>
<td>0.0044</td>
<td>0.948</td>
</tr>
<tr>
<td><strong>Approach With</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avoidance</td>
<td>0.59</td>
<td>0.5904</td>
<td>0.0536</td>
<td>0.0529</td>
<td>0.0029</td>
<td>0.942</td>
</tr>
<tr>
<td>Task</td>
<td>0.17</td>
<td>0.1698</td>
<td>0.0754</td>
<td>0.0740</td>
<td>0.0057</td>
<td>0.947</td>
</tr>
<tr>
<td><strong>Avoidance With</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task</td>
<td>0.31</td>
<td>0.3101</td>
<td>0.0764</td>
<td>0.0745</td>
<td>0.0058</td>
<td>0.939</td>
</tr>
</tbody>
</table>

*Note.* The column labels were partially changed from original Mplus outputs to enhance clarity. *Approach By X1* refers to a path from the Approach factor to the X1 variable. *Approach With Avoidance* refers to a correlation between the Approach factor and Avoidance factor. We report parameter values to four decimal places at most—a default setting in Mplus—for accuracy.

Table 2

*Number of Articles by Journal*

<table>
<thead>
<tr>
<th>Journal</th>
<th>AL</th>
<th>LAQ</th>
<th>LL</th>
<th>LT</th>
<th>MLJ</th>
<th>RELC</th>
<th>SSLA</th>
<th>System</th>
<th>TQ</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of articles</td>
<td>5</td>
<td>1</td>
<td>13</td>
<td>25</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>5</td>
<td>63</td>
</tr>
<tr>
<td>Number of articles reporting information for power estimatesa</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>18</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>42</td>
</tr>
</tbody>
</table>

*Note.* aThese articles are marked with double asterisks (**) in the references,
Running head: Review of SEM sample size

Table 3  
**Precision and Power of Parameters Across 18 Models**

<table>
<thead>
<tr>
<th>Parameter bias (%)</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>Mode</th>
<th>Outlier n</th>
</tr>
</thead>
<tbody>
<tr>
<td>[10% or less]</td>
<td>0.000</td>
<td>10.125</td>
<td>0.355</td>
<td>0.739</td>
<td>0.191</td>
<td>0.000</td>
<td>1 (0.003%)</td>
</tr>
<tr>
<td>SE bias (%) [10% or less]</td>
<td>0.000</td>
<td>4.465</td>
<td>1.022</td>
<td>0.829</td>
<td>0.877</td>
<td>0.000</td>
<td>0</td>
</tr>
<tr>
<td>SE bias for parameter in focus (%) [5% or less]</td>
<td>0.000</td>
<td>4.465</td>
<td>1.022</td>
<td>0.829</td>
<td>0.877</td>
<td>0.000</td>
<td>0</td>
</tr>
<tr>
<td>95% coverage [.91–.98]</td>
<td>0.936</td>
<td>0.963</td>
<td>0.948</td>
<td>0.004</td>
<td>0.948</td>
<td>0.948</td>
<td>0</td>
</tr>
<tr>
<td>Power [0.80 or more]</td>
<td>0.057</td>
<td>1.000</td>
<td>0.962</td>
<td>0.140</td>
<td>1.000</td>
<td>1.000</td>
<td>17 (5.648%)</td>
</tr>
</tbody>
</table>

**Note.** Outlier n = the number of corresponding parameters exceeding the criteria of Muthén and Muthén (2002) (thus judged to be outliers, or unfavorable results); the percentage of outlier n was calculated by dividing the number of outlier parameters by 301 (the total number of parameters across the 18 models). [ ] = Criteria for judgment. SE = standard error.

Table 4  
**Sample Size Requirements for the Nine Models With 17 Underpowered Parameters**

<table>
<thead>
<tr>
<th>Underpowered parameter</th>
<th>Original sample size</th>
<th>The smallest sample size for power &gt;= 0.80</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Size</td>
</tr>
<tr>
<td>Model A</td>
<td>5</td>
<td>0.071</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.080</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.098</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.024</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.041</td>
</tr>
<tr>
<td>Model B</td>
<td>4</td>
<td>0.626</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.515</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.361</td>
</tr>
<tr>
<td>Model C</td>
<td>2</td>
<td>0.160</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.210</td>
</tr>
<tr>
<td>Model D</td>
<td>1</td>
<td>0.170</td>
</tr>
<tr>
<td>Model E</td>
<td>1</td>
<td>0.150</td>
</tr>
<tr>
<td>Model F</td>
<td>1</td>
<td>0.140</td>
</tr>
<tr>
<td>Model G</td>
<td>1</td>
<td>0.070</td>
</tr>
<tr>
<td>Model H</td>
<td>1</td>
<td>-0.023</td>
</tr>
<tr>
<td>Model I</td>
<td>1</td>
<td>-0.080</td>
</tr>
</tbody>
</table>

Table 5  
**Sample Sizes of SEM Models Classified by Kline’s Guidelines (2005)**

<table>
<thead>
<tr>
<th>Small: 99 or less</th>
<th>Medium: 100–199</th>
<th>Large: 200 or above</th>
</tr>
</thead>
<tbody>
<tr>
<td>63 models</td>
<td>7 (11.111%)</td>
<td>14 (22.222%)</td>
</tr>
<tr>
<td>43 models</td>
<td>5 (11.628%)</td>
<td>10 (23.256%)</td>
</tr>
<tr>
<td>18 models</td>
<td>1 (5.556%; Model J)</td>
<td>1 (5.556%; Model K)</td>
</tr>
</tbody>
</table>

**Note.** The 63 models refer to the initial number of models we collected. The 43 models refer to the number of models reporting information for parameter precision, power, and sample size estimates. Both models as classified by journals are shown in Table 2. The 18 models refer to the number of models based on which the current study was able to calculate precision and power to evaluate sample size.
Table 6  

<table>
<thead>
<tr>
<th>Models</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>Mode</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Outlier n</th>
</tr>
</thead>
<tbody>
<tr>
<td>63 models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>59</td>
<td>15,000</td>
<td>787.651</td>
<td>2036.533</td>
<td>275</td>
<td>275</td>
<td>5.934</td>
<td>39.601</td>
<td>--</td>
</tr>
<tr>
<td>2.</td>
<td>10</td>
<td>132</td>
<td>45.339</td>
<td>29.806</td>
<td>32.5</td>
<td>27</td>
<td>1.140</td>
<td>0.661</td>
<td>--</td>
</tr>
<tr>
<td>3.</td>
<td>100</td>
<td>1,320</td>
<td>453.387</td>
<td>298.064</td>
<td>320</td>
<td>270</td>
<td>1.140</td>
<td>0.661</td>
<td>34</td>
</tr>
</tbody>
</table>

| 43 models |      |        |            |        |        |      |          |          |           |
| 1.        | 59   | 4,765  | 502.419    | 777.139 | 275    | 275  | 4.337    | 22.202   | --        |
| 2.        | 10   | 105    | 40.581     | 23.172  | 31     | 27   | 1.096    | 0.570    | --        |
| 3.        | 100  | 1,050  | 405.814    | 231.724 | 310    | 270  | 1.096    | 0.570    | 21        | (48.837%) |

| 18 models |      |        |            |        |        |      |          |          |           |
| 1.        | 93   | 4,765  | 664.889    | 1069.301| 296.5  | 275  | 3.703    | 14.549   | --        |
| 2.        | 10   | 63     | 30.444     | 15.378  | 27     | 18   | 0.756    | -0.016   | --        |
| 3.        | 100  | 630    | 304.444    | 153.784 | 270    | 180  | 0.756    | -0.016   | 4a        | (22.222%) |

Note. Outlier n = the number of models whose sample size was below the desirable sample size judged by the N:q≥10 rule. For example, four of the 18 models had less than the desirable sample size judged by the N:q≥10 rule.

aModels A, D, J, and L.
## Table 7

**Comparison of Three Criteria for Sample Size Across 18 Models**

<table>
<thead>
<tr>
<th>Model (original N; minimum N estimated using the Monte Carlo method)</th>
<th>Power of parameter(s): 0.80 or above</th>
<th>Sample size: 100 or above (N ≥ 100)</th>
<th>Sample size-free parameter ratio: 10 or above (N:q ≥ 10)</th>
<th>No. of models (out of 18)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model K (172; --)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>7 (38.889%)</td>
</tr>
<tr>
<td>Model M (4765; --)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>7 (38.889%)</td>
</tr>
<tr>
<td>Model N (259; --)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>7 (38.889%)</td>
</tr>
<tr>
<td>Model O (296; --)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>7 (38.889%)</td>
</tr>
<tr>
<td>Model P (567; --)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>7 (38.889%)</td>
</tr>
<tr>
<td>Model Q (360; --)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>7 (38.889%)</td>
</tr>
<tr>
<td>Model R (258; --)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>7 (38.889%)</td>
</tr>
<tr>
<td>Model B (219; 54,269)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>7 (38.889%)</td>
</tr>
<tr>
<td>Model C (408; 908)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>7 (38.889%)</td>
</tr>
<tr>
<td>Model E (503; 653)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>7 (38.889%)</td>
</tr>
<tr>
<td>Model F (297; 747)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>7 (38.889%)</td>
</tr>
<tr>
<td>Model G (738; 1,288)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>7 (38.889%)</td>
</tr>
<tr>
<td>Model H (1,382; 31,832)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>7 (38.889%)</td>
</tr>
<tr>
<td>Model I (876; 9,526)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>7 (38.889%)</td>
</tr>
<tr>
<td>None</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>None</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>Model J (93; --)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>1 (5.556%)</td>
</tr>
<tr>
<td>Model L (275; --)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>1 (5.556%)</td>
</tr>
<tr>
<td>Model A (225; 4,625)</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>2 (11.111%)</td>
</tr>
<tr>
<td>Model D (275; 425)</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>2 (11.111%)</td>
</tr>
</tbody>
</table>

---

**Note.** N = sample size. -- = Not estimated because the original sample size was already found to be sufficient in the Monte Carlo analysis.