Summary of Minimum Sample Size Recommendations for SEM

Below is a table summary of some minimum sample size recommendations for structural equation modeling commonly noted in the literature and online. Minimum sample size recommendations are based on having sufficient sample size to reduce the likelihood of convergence problems and to obtain unbiased estimates or standard errors. The question of whether sample size is large enough to achieve sufficient power for significance tests, overall fit, or likelihood ratio tests is a separate question that is best answered by power analysis for specific circumstances (see the handout " Power Analysis for SEM: A Few Basics" for this class, and Hancock, 2013 and Lee, Cai, & MacCallum, 2012, for reviews). Please be cautioned that the numbers below are rough approximations based on recommendations from authors who have conducted simulation studies and that they may not apply equally to all circumstances. **They should not be taken as definitive, infallible, or exact.** Simulation studies can only examine a few conditions at a time and often involve simplified conditions compared with actual practice.[1]

Estimator	Recommended Minimum N	Example Citations	Comments
ML with multivariate normal data	[a] > 100 [b] ≥ 200-400 [c] 5:1 ratio of cases to free parameters [d] 10:1 ratio of cases to free parameters	[a] Anderson & Gerbing (1984) [b] Jackson (2001) [c] Tanaka (1987) [c,d] Bentler & Chou (1987)	These suggested sample sizes are based on ML estimation with multivariate normal data, which may be somewhat rare in practice, and correctly specified models. For analyses with fewer than 100 or so cases, some authors would suggest using *t* critical values instead of *z* critical values for parameter significance tests. The complexity of the model is important and the minimum needed for simple path models, which are equivalent to regression models, may be different from complex full structural models with latent variables. Though even the 10:1 ratio is often considered safe, simulation work by Nevitt and Hancock (2004) suggest that there are some circumstances when this is not sufficient.
MLM for nonnormal continuous variables (ML with robust standard errors and Satorra-Bentler scaled chi-square)	≥ 250	Hu & Bentler (1999); Yu & Muthén (2002)	When data are multivariate normal (and no missing data), standard ML and MLM will have the same estimates. Overcorrection of standard errors can occur if sample sizes are too small (e.g., < 250).
Bootstrap for nonnormal continuous variables	≥ 200-1000	Nevitt & Hancock (2001)	When data are multivariate normal, standard ML is preferable in terms of unbiased and efficient standard errors. Standard errors were well below true values when sample sizes were < 1000 for moderately nonnormal data. Bootstrap estimates of standard errors do not perform well with small sample sizes (< 200), but performance may depend on the complexity of the model. They note that a sample size of 100 could be sufficient for simple models. Nevitt and Hancock recommend 250 or more bootstrap samples be used for estimation, although many sources recommend 500-1000 bootstrap samples in various contexts, Nevitt and Hancock find that more than 250 bootstrap samples did not improve estimates.
Bootstrap tests of indirect effects	>50-500	Creedon & Hayes (2015); Fritz, Taylor, & MacKinnon (2012); Tofighi & MacKinnon (2015)	Percentile bootstrap confidence intervals for indirect effects do not show seriously inflated Type I error even for very small sample sizes of 50 or 100 (Creedon & Hayes, 2015; Fritz, Taylor, & MacKinnon, 2012; Tofighi & MacKinnon, 2015), but bias-corrected and accelerated bias-corrected methods required at least 500 cases to avoid problematic Type I error. Results from Fritz and colleagues showed that the Type I error problems depended on the effect sizes of the a and b effects, with larger effects sometimes showing more problematic rates or lower than expected power. Although Taylor and colleagues and Tofighi and MacKinnon found that the Monte Carlo numerical integration approach performed similarly to the percentile bootstrap method, results reported by Creedon and Hayes suggested percentile bootstrap was superior for the smallest sample sizes.
MLR for continuous nonnormal missing data (robust ML)	> 400	Savalei & Bentler (2005); Yuan & Bentler (2000)	Though more simulation work is probably needed, the robust adjustments with full information maximum likelihood appear to work well when data are MAR and sample sizes are 400 or above.
Robust DWLS for with binary or ordinal variables (WLSMV in Mplus and lavaan)	≥ 200-500	Bandalos (2014); Forero, Maydeu-Olivares, & Gallardo-Pujol (2009)	Unadjusted categorical WLS does (diagonal or full weight matrix) does not perform as well as the mean and variance adjusted (robust) version of DWLS (Bandalos, 2014). 500 or more cases may be needed for sufficient power to reject models. Less than 200 seems to be associated with serious standard error bias and inflated Type I errors; 500 cases may be needed for nominal Type I error rate. Generally, more powerful than MLR for binary and ordinal variables.
Robust ML for binary and ordinal variables (MLR with categorical designation in Mplus)	≥ 200-500	Bandalos (2014)	Unadjusted marginal ML for binary and ordinal variables (full information ML) does not perform as well as the mean and variance adjusted (robust) version. Like robust DWLS, less than 200 seems to be associated with serious standard error bias and inflated Type I errors; 500 cases may be needed for nominal Type I error rate. Computationally more intensive, but performs comparably to WLSMV in most cases. May have less bias in standard errors than WLSMV for small sample sizes with asymmetric distributions in some cases (Bandalos, 2014).

[1] See handouts "Alternative Estimation Methods," "SEM with Nonnormal Continuous Variables," and “SEM with Categorical Variables” for more information on each of the estimation approaches described in the table.

There are many issues when considering minimum sample sizes. The minimum sample size recommendation of 100 comes from simulation studies (e.g., Anderson & Gerbing, 1984) that indicate an unacceptable number of models failed to converge when the sample size was 50 and a much more acceptable number (5% or less) failed to converge if the sample size was 100. Sufficient power to reject a model based on the chi-square test of the model is another important consideration, and how alternative fit indices perform with different sample sizes is another (e.g., Hu & Bentler, 1999). Then there is sufficient power for individual parameter tests (loadings, paths). The ratio of cases to free parameters, or *N:q*, which is sometimes stated in terms of indicators in the context of CFA, is commonly employed for minimum recommendations, but may not be as important as other considerations such as the overall sample size (> 200-400) and magnitude of the loadings (e.g., standardized value > .60), which may be more important (Jackson, 2007). In fact, Wolf and colleagues (Wolf, Harrington, Clark, & Miller, 2013) show that having more indicators per factor leads to smaller required sample sizes rather than larger required sample sizes in general. Whether the model is misspecified—whether the true model differs from the one tested—is also critical to how many tests perform under various sample size conditions. Absolute fit indices (e.g., chi-square, RMSEA) appear to be more sensitive to misspecification than relative fit indices (e.g., CFI). Hu and Bentler (1999) suggested that there may be a tendency for the combination rules of absolute and relative fit indices to over reject models if sample size is less than or equal to 250. Jackson's results suggest a highly complex set of interactions among specific fit index, loading magnitude, misspecification, and the *N:q* ratio, making clear that there is no simple rule to go by.

References

Anderson, J. C.,& Gerbing, D.W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49,* 155–173.

Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(1), 102-116.

Bentler, P. M., & Chou, C.-P. (1987). Practical issues in structural equation modeling. *Sociological Methods and Research, 16,* 78–117.

Creedon, P. S., & Hayes, A. F. (2015). *Small sample mediation analysis: How far can we push the bootstrap.* Presented at the Annual Conference of he the Association for Psychological Science.

Jackson, D. L. (2001). Sample size and number of parameter estimates in maximum likelihood confirmatory factor analysis: A Monte Carlo investigation. *Structural Equation Modeling, 8,* 205–223.

Jackson, D. L. (2003). Revisiting sample size and the number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling, 10,* 128–141.

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, *16*(4), 625-641.

Fritz, M. S., Taylor, A. B., & MacKinnon, D. P. (2012). Explanation of two anomalous results in statistical mediation analysis. *Multivariate behavioral research. Multivariate Behavioral Research, 47*(1), 61-87.

Hancock, G.R. (2013). Power analysis in structural equation modeling. In G.R. Hancock & R.O. Mueller (Eds.), *Structural Equation Modeling: A second Course, 2nd Edition* (pp.117 -162). Charlotte, NC: Information Age Publishing.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Lee, T., Cai, L., & MacCallum, R.C. (2012). Power analysis for tests of structural equation modeling. In Hoyle, R. H. (Ed.). *Handbook of structural equation modeling* (pp. 181-194). New York: Guilford Press.

Nevitt, J., & Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling*, *8*(3), 353-377.

Nevitt, J., & Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research*, *39*(3), 439-478.

Tanaka, J. S. (1987). “How big is big enough?”: Sample size and goodness of fit in structural equation models with latent variables. *Child Development, 58,* 134–146.

Tofighi, D., & MacKinnon, D. P. (2016). Monte Carlo confidence intervals for complex functions of indirect effects. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(2), 194-205.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73*, 913-934.

Yu, C.-Y., & Muthén, B. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* Doctoral dissertation. Retrieved from http:// www.statmodel.com/download/Yudissertation.pdf