International Journal of Accounting Information Systems 14 (2013) 256–274

Contents lists available at ScienceDirect

International Journal of Accounting Information Systems

Sample size requirements in structural equation models under standard conditions

Andreas I. Nicolaou [a,b,*], Michael M. Masoner [c]

[a] *Bowling Green State University, United States*
[b] *University of Cyprus, visiting 2003–04*
[c] *Emeritus, Southern Illinois University at Carbondale, United States*

ARTICLE INFO

Keywords:
Standard conditions in sample size for stability
Barebones for high quality
Relationship barebones standard and ranks quality standards
Multitude of standards both single-study and multi-study perspectives
Research settings where sample size is constrained or collection is costly

ABSTRACT

When a research of risk considers researchers who wish to utilize structured equation modeling (SEM), all users would ask for their sample size. The paper argues for the formulation of a single barebones minimum standard to be accompanied by a hierarchy of higher quality standards. The paper also offers a specific recommendation for such a barebones standard and ranks quality standards by their sample size cost. In sum, the solution to this problem involves integration: over the multitude of standards, over both single-study and multi-study perspectives, and over a broad array of research environments. Our solution deals with the multitude of solutions. It seeks maximum flexibility to accommodate a multi-study perspective, and it is sensitive to the needs of research settings where sample size is constrained or collection is costly. For these reasons we offer important suggestions and summarize recommendations.

1. Introduction

This paper considers the following question: When the researcher wishes to utilize Structural Equation Modeling (SEM) in a study, what sample size should he or she plan on gathering?

The present work began when its authors presented an empirical study utilizing SEM on data containing 160 observations. The authors justified the sample size with power calculations for the RMSEA tests of close and not close fit (review of *SEM NET*). Comments from the audience referred to the rule of thumb found in the EQS manual (review of SEM NET) that a study should have five observations per parameter estimated. A subsequent review of the SEMNET list located a question about sample size requirements. Many list members relied on the traditional regression standard of five to ten observations

* Corresponding author at: Bowling Green State University, United States.
E-mail addresses: anicol@bgsu.edu (A.I. Nicolaou)., mmasoner@siu.edu (M.M. Masoner).

1467-0895/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.accinf.2013.11.001

per parameter, while other members believed that it was appropriate to analyze smaller sample sizes with SEM under certain conditions, as supported by Marsh et al. (1997) and Marsh and Hau (1999). Overall, there was little consensus within the SEM community on the issue. This work is important since it specifically sought to identify minimum sample sizes that could be appropriately analyzed with SEM.

A closely-related body of research concerned the comparison of estimation methods using analysis and simulation. A by-product of this research is guidelines for sample size determination. In contrast, the studies of Marsh, Hau, and Balla focused solely upon the issue of minimum sample size determination. All of this literature however reported upon the properties of SEM statistics as sample size varied. Therefore it is treated in this paper as one body of evidence (and is collectively referred to as the "estimation literature").

Two quite separate and different considerations in sample size determination are the tools for SEM power determination provided by MacCallum et al. (1996) (MBS hereafter) and Saris and Satorra (1993) (SS hereafter). They are viewed as a second and third perspective. A fourth perspective is the traditional set of rules of thumb that include the regression standard of five to ten observations per parameter. These rules were indicative of quality research and therefore are labeled as blue chip standards in this paper.

Naturally, a lack of consensus in the research community would suggest conflict between perspectives or complexities within a perspective. Both of these problems occur and are illustrated in Fig. 1. In this figure, different sample size standards are represented by functions with both positive and negative slopes. In addition, the four perspectives may be viewed as four groups of separate sample size standards. Fig. 1 contains double representation for two of the four groups. Without trying to oversell this problem,

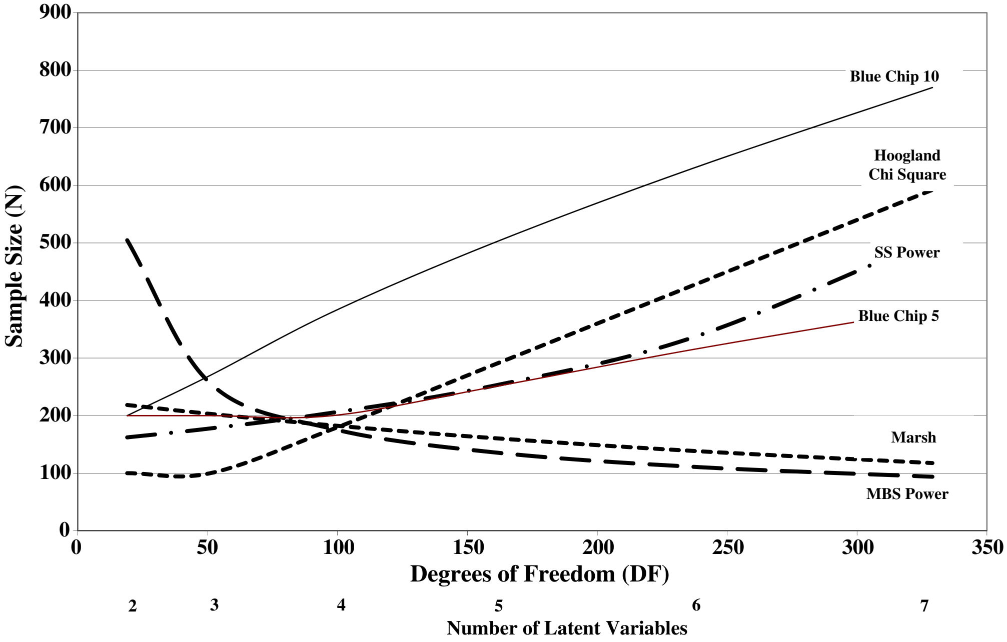


Fig. 1. Four standards: Estimation, MBS power, SS power, Blue Chip. Estimation parameter SD is based upon a regression of the Marsh and Hau (1999) data (Table 1, Fac Corr SD). The underlying data, standard deviations of parameter estimates, for a CFA Monte Carlo study with measured variable loading (λ) centered on .6. The graph also assumes the average value of SD exhibited in the data. Estimation chi square is based upon Hoogland (1999, p 143, ML Rule of Thumb). The graph assumes medium saturation (measured variable loading) (λ = .6) normal data. MBS power is based upon a regression of N and DF for 80% power (MacCallum et al., 1996). SS power is based upon a regression of N and DF for 80% power for testing a medium sized effect of an independent latent variable on a dependent latent variable using the two-step method (Saris and Satorra, 1993) and under the additional condition of medium saturation (λ = .6). Blue Chip 5 has four rules: a & b) Min (5*# of Parameters, 200), c) 4 items per latent variable, and d) a unique loading (λ) and disturbance (θ) for each measured variable. The graph also assumes one latent dependent variable. Blue Chip 10 is similar to Blue Chip 5 with the requirement of ten observations per parameter instead of five.

the representative rules graphed in Fig. 1 were selected more for their similarity than differences. Other standards would be off the graph.

Two ways of simplifying Fig. 1 would be an "all" solution and an "any" solution. The "all" solution is to select the maximum sample size requirement for each degree of freedom over the range of standards in Fig. 1. If other standards that were not represented are considered however, this solution would recommend SEM to only a small fraction of the problems to which it has previously been applied with success.

This paper seeks an "any" solution. Society does not value all information the same. Obtaining a sample incurs a cost. The SEM methodology literature should offer an array of options concerning sample size within a cost/benefit framework. A bare minimum sample size standard should be set where justification exists but where reasonable risks still exist. Increased confidence could be rationally purchased by incurring the cost of greater sample size required by a higher standard. Higher standards should significantly reduce sample size sensitive risks.

The purpose of this paper is to offer a simplification of Fig. 1. Unfortunately, this figure cannot be replaced by a single curve. Some simplification however can be made. In general, the paper argues for the formulation of a single barebones minimum standard to be accompanied by a hierarchy of higher quality standards. The paper also offers a specific recommendation for such a barebones standard and ranks quality standards by their sample size cost. If these were adopted, Fig. 1 would be replaced by the graph shown later as Fig. 5.

A solution to this problem involves integration: over the multitude of standards, over both single-study and multi-study perspectives, and over a broad array of research environments. Our solution deals with the multitude of solutions. It seeks maximum flexibility to accommodate a multi-study perspective, and it is sensitive to the needs of research settings where sample size is constrained or collection is costly.

This paper has nine sections. The first three sections provide an overview of the problem. The next four sections consider the four standard areas individually. The eighth section returns to an overview, and the conclusion then summarizes the recommendations.

2. Investigator's decision process

The alternative standards in Fig. 1 represent different relationships involving a small number of variables — mainly three in total: sample size, model size, and phenomena size. The discussion will temporarily widen its focus to describe the setting in which the three interact with other factors that influence sample size requirements.

In conducting a research study, the investigator's decision process is summarized in Fig. 2 with sample size relevant factors identified. There are two critical points in time in conducting a research study — when the study is designed and then after the data has been collected and initially examined. The factors that are controllable versus uncontrollable differ at the two points in time. Fig. 2 more accurately portrays the first of the two points in time in terms of what is and is not controllable.

In Panel 2A of Fig. 2, the researcher forms expectations concerning saturation (or the measured variable loading on the latent variable along with its accompanying measurement error), effect size (or the latent-to-latent loading along with the latent disturbance), normality, and specification error — all of which are represented as the Uncontrollables in Panel 2A. He or she then combines these expectations with sample size standards obtained from the SEM methodological literature. These sample size standards, listed in Panel 2B, relate to the four Fig. 1 sample size considerations previously described. Combining expected values of uncontrollables with sample size standards provides a setting of possible choices concerning the decision variables of the study. From this setting of possible choices, specific values of the decision variables (controllables) are selected. The controllables are manipulated in order to gain varying levels of achievement with respect to the statistical objectives.

These controllables especially include the planned size of the sample to be collected, the size of the model to be estimated (number of latent variables, measured variables per latent variable, and parameters per measured variable), and the estimation method along with other design choices — all of which are represented as the "controllables" in Panel 2A and listed in Panel 2C.

Communication is facilitated by two groupings. Saturation and effect size are bundled into the concept of phenomena size. Three variables (number of latent variables, measured variables per latent variable,

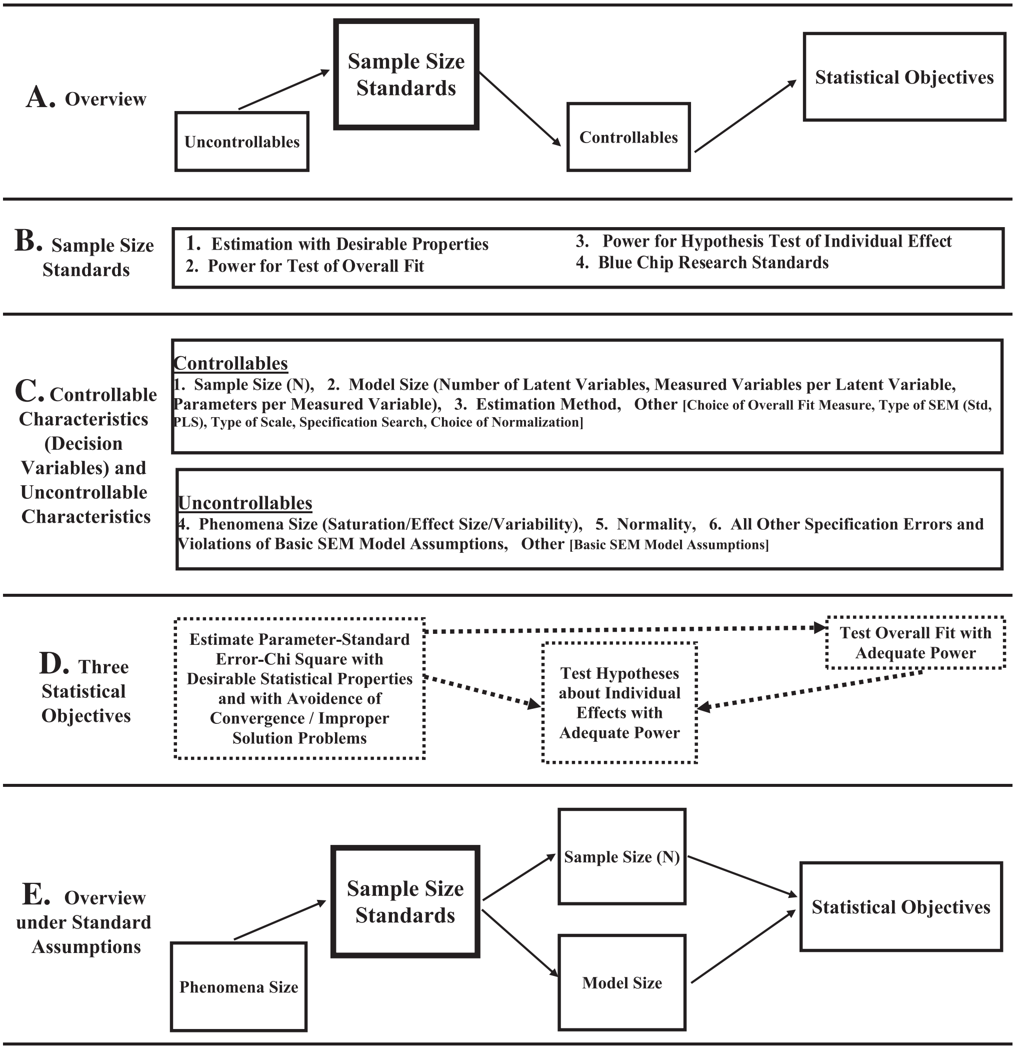

A. Overview
Uncontrollables
Sample Size Standards
Controllables
Statistical Objectives
B. Sample Size Standards
1. Estimation with Desirable Properties
2. Power for Test of Overall Fit
3. Power for Hypothesis Test of Individual Effect
4. Blue Chip Research Standards
C. Controllable Characteristics (Decision Variables) and Uncontrollable Characteristics
Controllables
1. Sample Size (N), 2. Model Size (Number of Latent Variables, Measured Variables per Latent Variable, Parameters per Measured Variable), 3. Estimation Method, Other [Choice of Overall Fit Measure, Type of SEM (Std, PLS), Type of Scale, Specification Search, Choice of Normalization]
Uncontrollables
4. Phenomena Size (Saturation/Effect Size/Variability), 5. Normality, 6. All Other Specification Errors and Violations of Basic SEM Model Assumptions, Other [Basic SEM Model Assumptions]
D. Three Statistical Objectives
Estimate Parameter-Standard Error-Chi Square with Desirable Statistical Properties and with Avoidence of Convergence / Improper Solution Problems
Test Hypotheses about Individual Effects with Adequate Power
Test Overall Fit with Adequate Power
E. Overview under Standard Assumptions
Phenomena Size
Sample Size Standards
Sample Size (N)
Model Size
Statistical Objectives

Fig. 2. Sample size standards and their context.

and parameters per measured variable) are similarly bundled into the concept of model size. Degrees of freedom is a function of the three sub-variables of model size and therefore serves as the overall measure of model size.

Phenomena size and the other uncontrollables are exogenous. Model size is viewed not only as a controllable (or decision variable) but also as the only real policy variable in the process. Using the term model size as a communication tool, has a serious limitation since there is generally a negative relationship between its third sub-variable, parameters per measured variable, and required sample size. Also viewing the choices concerning the model only in terms of the “size of the model” is a shallow view that is augmented in a later section.

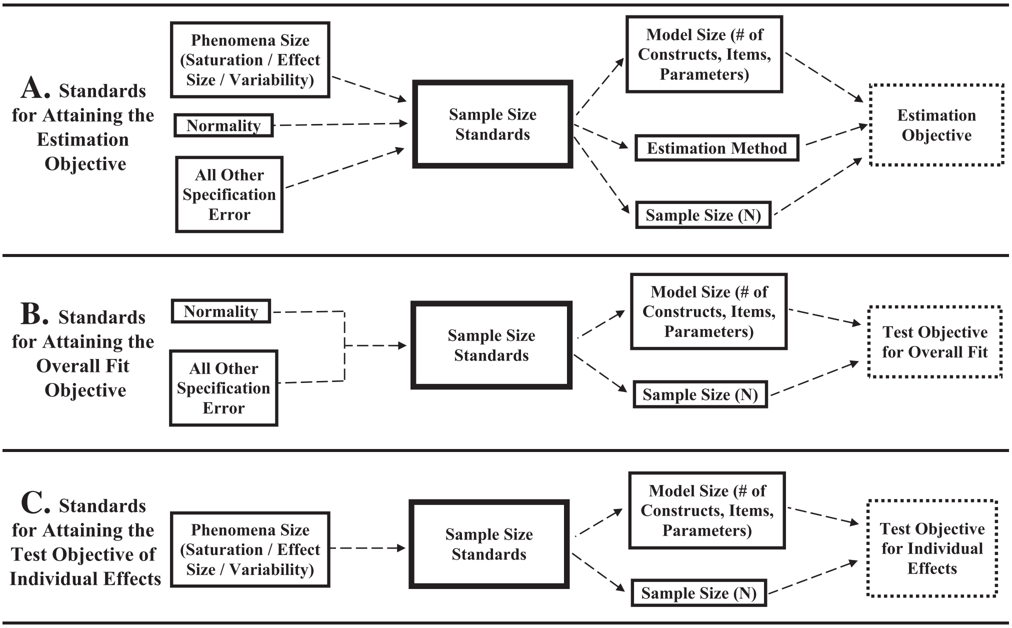
Three statistical objectives, listed in Panel 2D, coincide with three of the four areas of sample size standards: desirable properties of SEM statistics, MBS power, and SS power. The blue chip standards are characterized here as general standards not directed at any specific statistical objective.

Fig. 2 is an abstraction and simplification of the diversity of research settings. In many situations, sample size is fixed or constrained. Such constraints are not exhibited in Fig. 2. In unconstrained cases, the monetary and time cost of collection differs greatly on a per observation basis between research settings. These costs are also not found in Fig. 2. Nevertheless, this paper was written from the point of view that such costs are often a significant consideration in research.

Sample size is related to a great many SEM aspects. All aspects where sample size is an independent variable instead of a dependent variable are ignored in Fig. 2. For example, all measures of overall fit are influenced by sample size (Fan et al.). Concerning other variables, Marsh et al. treat saturation as a controllable through careful pilot testing and scale refinement. Fig. 2 is therefore not a perfect picture of this complex situation.

For most of the variables that are explicitly shown in Fig. 2, their actual values are constrained to a "standard" set of conditions in this paper. The tables and arguments presented are therefore narrow. Generalization from these restrictions occurs in the Conclusion.

The calculations underlying all of the tables to follow are based upon the maximum likelihood fit function, and the exposition implicitly implies maximum likelihood estimation. Other assumptions are: 1) RMSEA is the primary statistic for assessment of overall model fit; 2) measured variables have continuous scales; 3) the model is relatively well specified for the data; 4) the data conforms reasonably to normality; and 5) the basic SEM model assumptions hold with the exception that the small sample strategies of constraining measured variable loadings to be equal and constraining measured variable disturbances to have equal variance are considered. Finally, Tables 3 and 5 are based upon the normalization rule that each latent variable is measured in units of one of its corresponding measured variables. The alternative normalization of constraining latent variable variances to one would have changed the entries in the two tables when the small sample equality constraints were imposed (the implications drawn from the tables would not however change).


A. Standards for Attaining the Estimation Objective
Phenomena Size (Saturation / Effect Size / Variability)
Normality
All Other Specification Error
Sample Size Standards
Model Size (# of Constructs, Items, Parameters)
Estimation Method
Sample Size (N)
Estimation Objective
B. Standards for Attaining the Overall Fit Objective
Normality
All Other Specification Error
Sample Size Standards
Model Size (# of Constructs, Items, Parameters)
Sample Size (N)
Test Objective for Overall Fit
C. Standards for Attaining the Test Objective of Individual Effects
Phenomena Size (Saturation / Effect Size / Variability)
Sample Size Standards
Model Size (# of Constructs, Items, Parameters)
Sample Size (N)
Test Objective for Individual Effects

Note: Model Size (# of Constructs, Items, Parameters) refers to Number of Latent Variables, Measured Variables per Latent Variable, and Parameters per Measured Variable in Figure 2 and the rest of the paper.

Fig. 3. Sample size standards by objective. Note: Model size (# of constructs, items, parameters) refers to number of latent variables, measured variables per latent variable, and parameters per measured variable in Fig. 2 and the rest of the paper.

Under these standard conditions and assumptions, the lists of controllables and uncontrollables in Panel 2C of Fig. 2 that are the focus of this paper are reduced to two controllables, sample size and model size, and one uncontrollable, phenomena size. Panel 2E of Fig. 2 shows this reduction. With sample size as the dependent variable, the lone controllable is model size with its three sub-variables: number of latent variables, measured variables per latent variable, and parameters per measured variable.

With the above mentioned variables set to standard conditions, with phenomena size also held constant, and with model size measured by degrees of freedom, Fig. 1 expresses the problem addressed in this paper. Even within this restricted context, there are diverse sample size standards that need narrowing and simplification.

To momentarily stay with Fig. 1, the question can be asked if SEM is intended for the analysis of small models or large models. In economic terminology, does SEM offer decreasing, constant, or increasing returns to scale in its processing of models of different size. Fig. 1 would suggest that constant returns to scale occur if the estimation–chi square, SS power, or Blue Chip curves are followed. In other words, past a minimum size, the size of the model does not matter. In contrast, Estimation–Parameter SD and MBS Power suggest increasing returns to scale. They imply a SEM analysis advantage from expanding the boundaries and context of a research problem. Accordingly, there are wider implications than just sample size determination to the way in which Fig. 1 is simplified.

3. Three statistical objectives

The problem can be more accurately portrayed by individually considering the three statistical objectives. Fig. 3 makes this presentation with uncontrollables and controllables separated into their six constituent elements.

Panel 3A of Fig. 3 concerns the objective of estimating model parameters and related standard errors with an appropriate bias and variance; of estimating model chi square statistics with appropriate bias, variance, rejection rate, and distributional behavior; and of estimating models with an appropriate chance of avoiding non-convergence and improper solutions.

Attaining this objective is a more complex task than attaining the other two objectives. This is illustrated by all six controllables/uncontrollables being present in Panel 3A while Panel 3B has only four elements and Panel 3C has three elements. By far, the largest body of literature has been directed at Panel 3A in comparison with the other two panels. As mentioned in the Introduction, this literature sought simultaneous determination of estimation method and sample size.

Panel 3A research typically produces eight different sample size standards, i.e., standards for: parameter estimator bias/variation, standard error estimator bias, chi-square statistic rejection rates/bias/variation, and non-convergent/improper solution frequency. Each of the eight can be multiplied by four giving 32 different sample size standards since separate consideration has been given to the overall model and individual ϕ, λ, and θ parameter sets (this factor would be six and not four if structured instead of unstructured models had been typically investigated). The number of such sample size standards is next expanded by multiplying 32 by the number of different estimation methods under consideration. Which of these sample size standards should be followed and satisfied is part of the question of this paper.

The objectives for Panels 3B and 3C of Fig. 3 concern the attainment of appropriate power for the SEM analysis. Panel 3B relates to the assessment of overall model fit using the RMSEA statistic in the tests of close and not-close fit with the MBS power procedure. Panel 3C involves tests of specific latent variable(s) using the two-step SS power procedure.

For Panel 3B, MacCallum et al. have specified a RMSEA value corresponding to the null hypothesis, a RMSEA value for the close test's alternative hypothesis, and a RMSEA value for the not-close test's alternative hypothesis. These three RMSEA values differ solely because of specification error (that includes deviation from normality). MacCallum et al. have therefore set the standard for specification error by determining these three RMSEA values. Given the MacCallum standards for specification error, the additional standard of 80% power, and the size of the researcher's model, the MBS power procedure allows sample size to be determined.

Turning to the SS power procedure in Panel 3C, the present paper makes only a specific application of this general tool. There are many SS power calculations that can be made. In this paper, only one such calculation is considered. Different ranges of saturation are taken from the literature and combined with

Cohen's effect size standards in order to specify phenomena size. Therefore power is calculated for relationships between two latent variables under a special set of constraints on the related measured variables.

Comparing Panels 3B and 3C reveal phenomena size is missing from Panel 3B while specification error (including non-normality) is missing from Panel 3C. The different way in which discrepancies, between the sample covariance matrix and different population covariance matrices, are defined results in this difference. This subtle difference carries into the different purposes for which the two approaches can be used.

In all three panels of Fig. 3, model size represents the central controllable which the researcher can manipulate to accommodate a small sample size. Marsh et al. (1997) investigated two small sample strategies involving the sub-variables within model size. First, measured variables per latent variable were manipulated. Second, concerning parameters per measured variable, loadings were constrained to be equal across measured variables for a given latent variable. The Marsh work was conducted within Panel 3A. That work is extended in this paper to Panels 3B and 3C.

4. A logical chain

The strategy of this paper is to link, so far as possible, the three statistical objectives to facilitate the application of Occam's razor. Relationships between the three have not been emphasized in the past and are not necessarily strong. Nevertheless some connections can be made.

Neither Panel 3B or 3C of Fig. 3 contain estimation method as a variable. Since each estimation method has a unique fit function that is used in power calculations, estimation method affects both pictures. However, its role is that of an exogenous variable or "uncontrollable" instead of being a dependent variable as in Panel 3A where it is simultaneously determined with a sample size requirement. Since estimation method is determined in Panel 3A and then used in Panels 3B and 3C, Panel 3A determinations must precede those of Panels 3B and 3C.

While the sample size requirement produced by the SS two-step procedure is in no way dependent upon any determination from the MBS procedure, the object of the SS procedure, namely individual hypothesis tests, is dependent upon the object of the MBS procedure, attaining an acceptable RMSEA measure of overall fit. In other words, good overall fit supports and lends validity to the conclusions of individual hypothesis tests. Therefore the three statistical objectives (and three sample size requirements) can be viewed in a logical order of calculation with Panel 3A prior to Panel 3B which is prior to Panel 3C.

Table 1
Logical ordering of standards and objectives.

Logical order	Literature	Standard	Objective	Object of objective
1	Analysis and simulation studies on SEM statistics' behavior	Increase sample size until statistic's behavior within specified tolerance	Estimate parameter–standard error–chi square with desirable statistical properties and with avoidance of convergence/improper solution problems	Estimators of parameters, parameter standard errors, chi square statistics, and performance statistics
2	MacCallum et al. (1996); Browne and Cudeck (1993); Steiger (2000)	Set sample size to provide 80% power in the RMSEA tests of close and not-close fit	Test overall fit with adequate power	RMSEA measure of overall fit
3	Saris and Satorra (1993); Satorra and Saris (1985)	Set sample size to provide 80% power in tests of individual effects	Test hypotheses about individual effects with adequate power	Research hypotheses of study
4	General methodology and philosophy of science literature	Perform empirical study(s)	Choose between competing theories	Competing theories

A fourth element needs to be added to this chain. Individual hypothesis tests are the basis for choices between competing theories, i.e., the ultimate research objective. Combining these four elements in this logical ordering suggests Table 1.

If the ultimate objective is to choose between competing theories, then only those tools required to attain this single objective should be exercised. A barebones sample size standard would be based upon only the needs of this restrictive set of tools. For example, all 32 sample size standards of Panel 3A of Fig. 3 would not need to be satisfied if only some of the 32 were used in the task of choosing between competing theories.

The choice between competing theories may be quite local in some cases and more global in other cases. If the models representing competing theories are not nested, the choice has a more global nature. For nested models, a choice between competing theories may involve a single parameter or multiple parameters. The barebones sample size solution proposed in this paper will be constrained to the case of the most simple choice between competing theories — namely the single parameter hypothesis. This will allow focused results from the literature to be used, for example, in selecting only one from the 32 sample size standard of Panel 3A of Fig. 3.

5. Estimation standards

Adequate sample size is needed to obtain well-behaved SEM statistics. Panel 3A of Fig. 3 and Row 1 of Table 1 concern the related group of 32 + sample size standards intended to insure such behavior. These standards are supported with a large body of literature. On the subject of sample size, the most important result found in that literature is presented in Table 2.

When there are more latent variables in a model, when there are more measured variables per latent variable, when there are fewer parameters per measured variable, and when saturation levels are higher, then sample size does not have to be as high in order to avoid non-convergence (speaking in terms of probability). The same is true for improper solution problems and for the bias/variation in parameter and standard error estimators. This opportunity to substitute saturation and especially model size for sample size was labeled as “compensation” by Marsh et al. (1997).

Unfortunately, this substitution opportunity does not occur with chi-square statistics as shown in Table 2. The above factors, that formerly compensated for having a small sample size, have the opposite effect upon chi-square statistics. These four factors, as well as the additional factor of non-normality, increase the need for sample size in order to hold chi-square rejection rates constant. The same is true for other chi-square criteria. These factors can be said to “tax” sample size instead of substituting for it.

Table 2
Behavior of SEM statistics.

	Factors that compensate (substitute) for sample size	Factors that tax (increase the need for) sample size
Nonconvergence and improper solutions	# of latent variables[2], # of measured variables per latent variable [1,2], # of parameters per measured variable[1], saturation level[12]	Non-normality[2]
Parameter and Standard error estimators	# of latent variables[2], # of measured variables per latent variable [1,2], # of parameters per measured variable[1], saturation level [1,2]	Non-normality[2]
Chi-square statistics		Degrees of freedom (# of latent variables, # of measured variables per latent variable, # of parameters per measured variable) [1,2], saturation level [1,2], non-normality[2]

[1] Marsh and Hau (1999).
[2] Hoogland (1999).

This compensation-taxation difference in behavior that is shown in Table 2 presents a problem. Returning to Fig. 1, the problem is represented by the different requirements shown by the Hoogland–chi square curve and the Marsh–parameter SD curve. The Marsh–parameter SD curve limits SEM research with small models to having over 200 observations. The Hoogland–chi square curve requires large models to have an even larger sample.

The chi square statistic of the entire model is a measure of overall fit. To choose between theories however, a different chi square statistic, the sequential chi square statistic, is used in the case of nested models. The sequential chi square has degrees of freedom equal to only the specific number of constraints being tested where as the chi square statistic of the entire model has degrees of freedom that is a function of the size of the model. For the most simple nested hypothesis, namely that of a single parameter, the number of degrees of freedom of the sequential chi square would be only one.

The behavior of the sequential chi square statistic (as opposed to the chi square statistic for the entire model) has not been investigated in the simulation literature. That research can therefore not verify the taxing versus compensating behavior of the sequential chi square statistic when model size changes. While the sequential chi square statistic has not been explicitly investigated, the literature has investigated the bias/variation in parameter estimators and in standard error estimators. If the selection of a sample size standard were restricted to only choices between theories that involve a single parameter, recommendations could be based upon this literature.

Therefore, the barebones recommendation of this paper is given this restriction. Without using a sequential chi square statistic, the assessment of a parameter estimate's importance could still be made.

Table 3
Sample size for 80% power in RMSEA tests of close and not close fit.

# of latent	# of Eta	# of Ksi	Meas/ lat	Parameters/ measured	DF	N (max)	Parameters/ measured	DF	N (max)	Parameters/ measured	DF	N (max)
1	0	1	4	2	2	3493	1	4	1802	0	7	1074
1	0	1	5	2	5	1463	1	8	952	0	12	666
2	1	1	3	2	8	953	1	10	782	0	14	599
2	1	1	4	2	19	490	1	23	434	0	29	374
2	1	1	5	2	34	339	1	40	307	0	48	274
3	1	2	2	2	6	1237	1	6	1237	0	9	857
3	1	2	3	2	24	421	1	27	391	0	33	344
3	1	2	4	2	51	264	1	57	246	0	66	226
3	1	2	5	2	87	192	1	96	181	0	108	169
3	2	1	2	2	7	1074	1	7	1074	0	10	781
3	2	1	3	2	25	410	1	28	382	0	34	338
3	2	1	4	2	52	260	1	58	244	0	67	224
3	2	1	5	2	88	190	1	97	180	0	109	169
4	0	4	2	2	14	598	1	14	598	0	18	507
4	0	4	3	2	48	273	1	52	260	0	60	239
4	0	4	4	2	98	179	1	106	171	0	118	162
5	1	4	2	2	25	410	1	25	410	0	30	366
5	1	4	3	2	80	201	1	85	194	0	95	182
5	1	4	4	2	160	136	1	170	132	0	185	125
6	2	4	2	2	40	306	1	40	306	0	46	281
6	2	4	3	2	48	159	1	127	155	0	139	147
6	2	4	4	2	64	109	1	250	106	0	268	102

N (max) is the larger of the two sample sizes associated with the two tests.
# of Eta and # of Ksi are the number of latent dependent and independent variables respectively.
Meas/lat is the ratio of measured to latent variables.
Parameters/measured is the number of unique parameters assigned to each individual measured variables.
The value “2” refers to a unique lambda coefficient and a unique theta variance parameter for each measured variable. The value “1” refers to the lambda coefficients being the same across measured variables (that are associated with one latent variable) while the theta parameters differ. Alternatively, “1” refers to the thetas being the same for the variables measuring one latent variable while each measured variable has a unique lambda. The value “0” refers to all lambdas of a given latent variable being the same and all thetas being the same for the variables that measure that latent variable.

If the estimate's difference from zero or difference from a second estimate was twice the size of the parameter's standard error estimate, then the estimate in question could be considered as “significantly different”. The choice between competing theories would thus be made upon such significant differences.

To make this comparison, the parameter estimator and the standard error estimator are required. The literature has found the parameter estimator to be essentially unbiased. The focus therefore narrows to only the standard error estimator. Hoogland (1999, p. 108) has found the ϕ standard error estimator to be much less biased than the λ standard error estimator that in turn is less biased than the θ standard error estimator. Of the 32 sample size standards previously mentioned for estimation statistical objective, the sample size standard corresponding to the ϕ standard error estimator is the one that should be satisfied in the case of measurement models. When causality is explicitly introduced into the model, the γ and β standard error estimators of the structural model would join the ϕ standard error estimator in setting the sample size standard.

There is research for which the primary objective is not the choice between theories as has been emphasized in this paper. This second category of research seeks to produce valid scales for use by other researchers. Such research will have to satisfy the higher sample size requirements necessitated by the bias present in standard error estimators of λ and θ when data contains non-normality.

Finally, the Hoogland chi square sample size standard is treated as a quality standard that would allow confidence to be placed in the use of the chi square statistic for the entire model.

6. Power for RMSEA tests of close and not-close fit

Unlike the other three areas of sample size standards, the problem of multiple standards does not occur in the case of MBS power. A researcher would not compute a test of close fit without also computing a test of not-close fit. Therefore the maximum sample size requirement over both tests is the lone requirement considered here.

Table 3 of this paper reproduces MacCallum et al.'s (1996) Table 4 with additional information attached. The size of the model is manipulated to show the impact which this policy variable has on sample size requirements. A brief review of Panel 3B of Fig. 3 shows model size as the only other controllable besides sample size. The three RMSEA standard values established by MacCallum et al. (1996) and Browne and Cudeck (1993) effectively freeze the degree of normality and specification error from a MBS power calculation viewpoint. Table 3 therefore shows the impact of manipulation of the three model size policy variables: number of latent variables, measured variables per latent variable, and parameters per measured variable.

Viewed from a MBS power perspective, SEM is definitely a data analysis tool for analysis of large sized models. A model with one latent variable would require 3493 observations while a model with seven latent variables would require only 109 observations. Some research settings, of course, have fixed boundaries for the model and do not permit the number of latent variables to be treated as a policy variable. In other venues, the SEM researcher should consider expanding internal and external boundaries until the cost of such expansion exceeds the cost of obtaining additional observations. For example, the response rate to a mail survey will decline with the length of the questionnaire. The validity issues related to lower response rates would be weighed against the MBS power advantages of increasing the number of latent variables.

Increasing measured variables per latent variable has an equally potent effect. Adding an item per construct results in a rough decline in required sample size by between one third and one half.

Concerning the strategy of reducing parameters per measured variable, Marsh et al. (1997) found a dramatic improvement in the sample size requirements for estimation benefits. Specifically, the λs associating certain measured variables with one latent variable were constrained to be equal. Table 3 shows the corresponding effect upon MBS power. Decreasing parameters per measured variable is not as potent a MBS power strategy as increasing the number of latent variables or as increasing measured variables per latent variable. Nevertheless, in research situations where these other two policy variables are fixed at low values, constraining the λs, θs, or both to be the same for all measured variables associated with a given latent variable reduces the MBS sample size requirement.

The MBS power standard is surprisingly similar to the sample size requirement of ϕ's SD. They support one another. The ϕ standard would not require revision except when model size was small. Therefore, choosing the higher of the two standards over the degrees of freedom range of Fig. 1 provides the barebones sample size standard recommended in this paper.

7. Power for tests of individual effects

Before the authors turned to SEM as a data analysis tool, they primarily-used regression analysis in the past. A depressing step in the regression analysis process was to make Cohen power calculations to determine the sample size to be collected. The Cohen standards were quite high. The author's first MBS power calculation was a pleasant surprise given the Cohen expectations.

Saris and Sattora (1993) augment their SS power calculations with the isopower contours of Andrews (1989). In applying SS power however, Muthen and Curran (1997) returned to the more traditional Cohen approach. We follow Muthen and Curran.

An immediate question in this application is whether the Cohen standards can be translated into the SEM environment without modification. Cohen (1988) modified his more univariate standards for the environment of complex multivariate hypotheses — compare his Chapter 9 with Chapters 2 through 8 (Cohen, 1988). A similar modification may be warranted for SEM.

We have not made such a modification but instead have simply translated the traditional Cohen univariate standards into the SEM setting. The Cohen standards related to the effect of one measured variable upon a second measured variable. Latent variables were not involved. Multiple measurements of each measured variable were not considered. The Cohen standard certainly took account of measurement error in a broad, multi-discipline setting (Cohen, 1988, p. 13).

Therefore, we apply the Cohen standards as follows. One measured variable impacts a second measured variable. That effect is then divided between three links. The first measured variable is caused by a latent variable. That latent variable has an influence on a second latent variable. The second latent variable causes the second measured variable. Given this causal chain, an overall Cohen effect of the first measured variable on the second measured variable can be separated into a level of saturation in the measurement of the two latent variables and an effect of the first latent variable on the second latent variable.

Marsh and Hau (1999) considers $\lambda = .6$ as average saturation with $\lambda = .4$ viewed as low saturation and $\lambda = .8$ as high saturation. Given these three saturation levels, Table 4 translates the Cohen standards

Table 4
Effect size, saturation, and structural coefficient.

	Structural coefficient (gamma)	Variance of structural disturbance (Psi)
Low saturation [meas variable loading (lambda) = .4; uniqueness ($s_y = s_x$) = .84]		
Small effect ($f^2 = .01$)	0.25	0.94
Medium effect ($f^2 = .0625$)	NF	NF
Large effect ($f^2 = .16$)	NF	NF
Medium saturation [meas variable loading (lambda) = .6; uniqueness ($s_y = s_x$) = .64]		
Small effect ($f^2 = .01$)	0.17	0.97
Medium effect ($f^2 = .0625$)	0.43	0.82
Large effect ($f^2 = .16$)	NF	NF
High saturation [meas variable loading (lambda) = .8; uniqueness ($s_y = s_x$) = .36]		
Small effect ($f^2 = .01$)	0.13	0.98
Medium effect ($f^2 = .0625$)	0.31	0.90
Large effect ($f^2 = .16$)	0.51	0.74

Table 5
Sample size for 80% power in two step procedure to test an individual effect.

# of Latent	# of Eta	# of Ksi	Meas/lat	Parameters/measured	DF	Saturation	Effect size	N	Saturation	Effect size	N	Saturation	Effect size	N
2	1	1	3	0	15	Low	Small	780	Medium	Small	1264	High	Small	1605
2	1	1	3	0	15	Low	Medium	NF	Medium	Medium	183	High	Medium	248
2	1	1	3	0	15	Low	Large	NF	Medium	Large	NF	High	Large	88
2	1	1	3	2	9	Low	Small	649	Medium	Small	1052	High	Small	1335
2	1	1	3	2	9	Low	Medium	NF	Medium	Medium	153	High	Medium	206
2	1	1	3	2	9	Low	Large	NF	Medium	Large	NF	High	Large	73
2	1	1	4	2	20	Low	Small	797	Medium	Small	1279	High	Small	1682
2	1	1	4	2	20	Low	Medium	NF	Medium	Medium	185	High	Medium	259
2	1	1	4	2	20	Low	Large	NF	Medium	Large	NF	High	Large	92
3	1	2	4	2	53	Low	Small	1069	Medium	Small	1683	High	Small	2170
3	1	2	4	2	53	Low	Medium	NF	Medium	Medium	241	High	Medium	333
3	1	2	4	2	53	Low	Large	NF	Medium	Large	NF	High	Large	116
3	1	4	4	2	167	Low	Small	1405	Medium	Small	2092	High	Small	2538
3	1	4	4	2	167	Low	Medium	NF	Medium	Medium	294	High	Medium	382
3	1	4	4	2	167	Low	Large	NF	Medium	Large	NF	High	Large	128
6	2	4	4	2	245	Low	Small	1583	Medium	Small	2434	High	Small	2999
6	2	4	4	2	245	Low	Medium	NF	Medium	Medium	337	High	Medium	448
6	2	4	4	2	245	Low	Large	NF	Medium	Large	NF	High	Large	147

NF — Non-feasible.

Refer to Table 3 for the numerical values of gamma consistent with specific saturation and effect sizes levels.

Refer to Table 2 for definitions of the first seven columns.

Calculations based upon the following assumptions:

1) Factor loading and factor variances are equal across measured variables associated with a given latent variable.

2) Latent independent variables are independent of one another.

3) Measured dependent and independent variables associated with the effect size in question have the same factor loadings and factor variances (specified in table above). All other measured dependent and independent variables have medium factor loadings and factor variances.

4) See Table 3 for the gammas associated with an effect size tabled above. All other latent relationships were assumed to be medium saturation-medium effect size gammas. Given the gammas, psi's were set to make the latent dependent variable's variance equal to one.

for small, medium, and large effects ($f^2 = .01, .0625$, and $.16$) into values for γ and ψ. These values were obtained by solving the following:

$$f^2 = r_{yx}^2/\left(1-r_{yx}^2\right) = 1/\left[1+\left(\Psi/\gamma^2\right)+\left(\Theta/\lambda^2\cdot\gamma^2\right)\right]\cdot\left[1+\left(\Theta/\lambda^2\right)\right]$$

$$1 = \gamma^2 + \Psi = \lambda^2 + \Theta$$

where:

f^2 is the Cohen effect size measure — the ratio of the variance of means to the variance of disturbance (Cohen, 1988, p 281, 284).
r_{yx}^2 is the Pearson product moment correlation between the two measured variables.
γ is the loading of the dependent latent variable on the independent latent variable.
Ψ is the variance of the disturbance in the latent-to-latent relationship.
λ is the loading of each measured variable on its latent variable.
Θ is the variance of the measurement error in the measured-to-latent relationship.

Within a SEM setting, a Cohen effect of a given size is not represented by a single γ loading of one latent variable on the other. The level of measurement saturation must be taken into account as well. Table 4 shows that only a small Cohen measured effect is consistent with low saturation. When the latent variables are measured with good accuracy, it is possible to have small to large Cohen impacts between the latents. This is not the case with poor measurement. Obviously, the latent-to-latent effect size should be independent of measurement accuracy. This does not occur in Table 4 since a second tradition supplies a constraint. Latent variables are constrained to have a variance equal to one. If Cohen had originally worked in the SEM environment, his standards would have been stated directly in terms of γ and ψ. Translated standards therefore contain a limitation in that latent-to-latent effect size is dependent upon saturation level. For the purposes of the present paper, this limitation is overshadowed in our opinion by the interpretation value of benchmarking SS power sample size standards to older research.

Table 5 presents sample size requirements for SS power when model size varies and phenomena size varies. The high Cohen regression standards re-appear with a vengeance in Table 5. With other data analysis tools, such as regression, a larger sample size is required to have adequate power for measuring a small effect size as opposed to a large effect size. This holds true in the SEM environment.

8. Blue Chip standards

Detailed guidelines for empirical research are largely a product of the twentieth century. Their purpose has been to motivate the participant and not just to set boundaries for acceptability.

Fig. 1 contains two curves labeled Blue Chip 5 and Blue Chip 10. The Blue Chip 5 curve can be viewed as a candidate for the barebones sample size standard. The Blue Chip 10 curve would then represent a standard that quality research should be expected to satisfy in comparison to minimally acceptable research satisfying only barebones standards.

The elements going into Blue Chip 5 and Blue Chip 10 of Fig. 1 are actually four different standards concerning: 1) observations per measured variable, 2) minimum absolute sample size, 3) measured variables per latent variable, and 4) the absence of sample size saving strategies. These elements were jointly applied in Fig. 1. However, they have not been joined in the literature, and their merits are discussed independently below — beginning in reverse order.

Marsh et al. (1997) considered an equality constraint as a sample size saving strategy. Such a strategy is compatible with the idea of a barebones sample size. However, it would have a potentially biasing effect on the other estimated parameters of a model. Therefore it is inconsistent with a quality standard that certifies that adequate sample size has been attained. Therefore a barebones strategy should allow sample size saving strategies, and a quality standard should prohibit them.

Marsh et al. (1997) explicitly recommends four items per construct. There are local (specific latent variable) and global (overall model) aspects of this standard. Identification can be a local issue. Two items per construct is the low level where identification problems begin. In the other direction, over-identification is considered an empirical virtue (Bollen, 1989). Turning to the global side, three items per construct is the level where a construct breaks-even in terms of being a net source for or net user of degrees of freedom with respect to the model as a whole. Four items per construct is the lowest level where a construct produces net degrees of freedom for the other areas of the model.

The four-items-per-construct standard is not a direct sample size standard. However, it impacts sample size through the sometimes-compensating and sometimes-taxing impact that degrees of freedom have on sample size. Most of these compensating and taxing impacts have firm support in the literature. The impact of the four-items-per-construct standard therefore depends upon the degrees of freedom to sample size link with which it is paired.

From a barebones standpoint, the four-items-per-construct standard is unnecessary. In sample size constrained or high observations cost situations, four or five items per construct may be a useful small sample strategy for compensation. However, being a useful strategy to overcome a problem is not the same as being a minimum requirement. Also, four items per construct may be a valid standard for scale development. Again however, a barebones sample size standard is a minimum requirement to be met by all SEM applications — not just those involving scale development.

There is much to recommend the four-items-per-construct standard from Marsh et al. (1997), other independent literature, and earlier sections of this paper. The standard should serve as a general indicator of quality research. Accordingly, we believe it should be part of a quality standard but excluded from a barebones requirement.

Turning to the second of four Blue Chip elements, a minimum absolute sample size is a practical suggestion since underlying estimation theory is asymptotic. Such a standard would be a barebones candidate. Specifically regarding this minimum, Lomax (1989) and Hoogland (1999) recommend 100 observations. Baldwin (1989) recommends 200.

Since an absolute standard originates in the estimation standard area, the work of Marsh et al. (1997) and the major review by Hoogland (1999) is given weight here in suggesting a 200 observation standard is too high for barebones. Further detailed simulation research may raise or lower the Hoogland 100 observation minimum standard. If it is raised, however, that rise would not be to the 200 observation level. Finally, a 100 observation standard is never critical when viewing Fig. 1, but may be important when model size rises above the seven latent variable level.

The first of the four Blue Chip elements, five/ten observations per measured variable have been informal standards for over 50 years in regression and exploratory factor analysis. The five observations per measured variable was a barebones standard. The ten observations per measured variable was a quality standard.

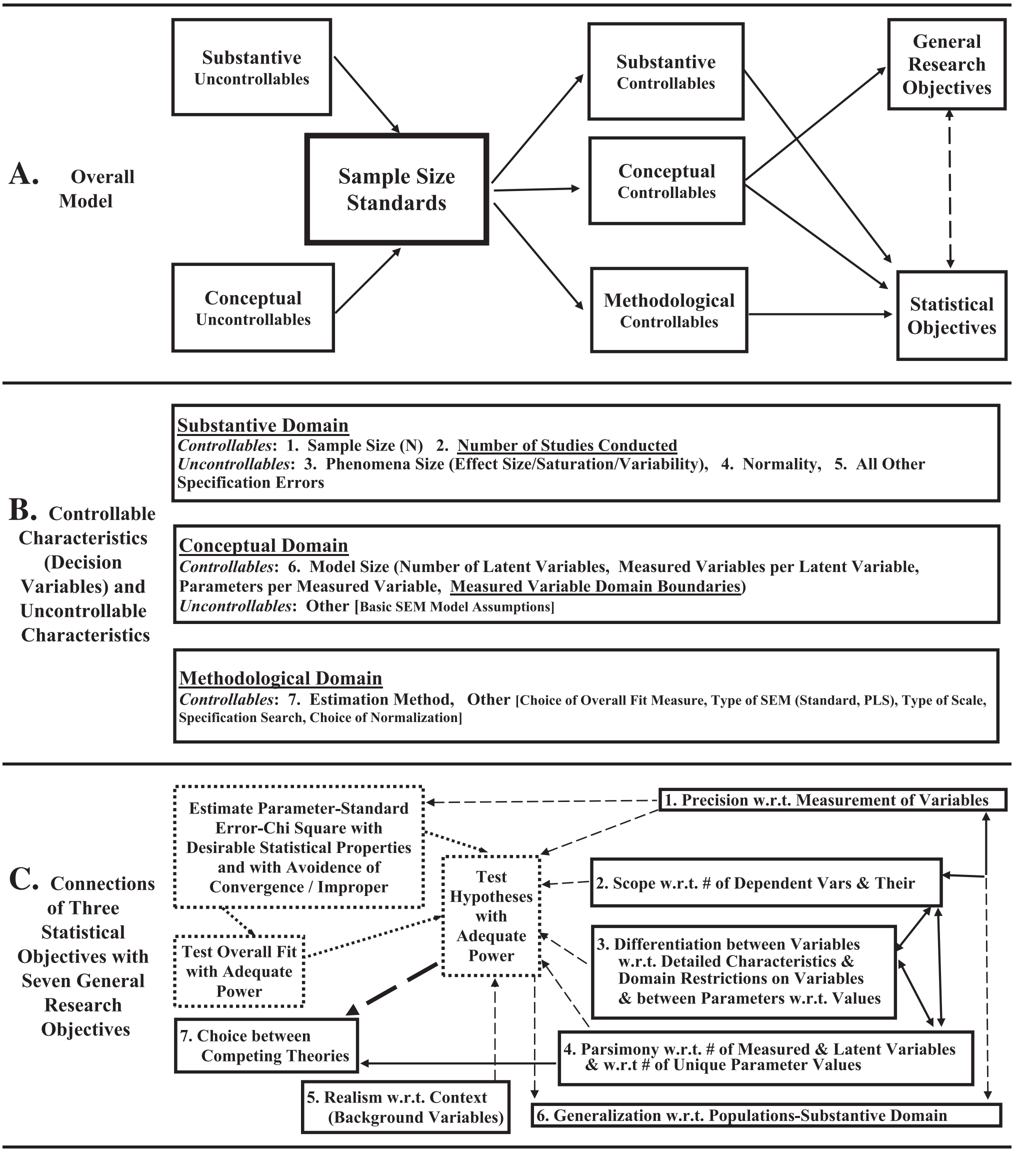
The view of this paper is that tradition is a poor justification for barebones guidance. A barebones standard should be a minimum that is feasible but still accompanied by risk. Such a standard should never be traditional but constantly revised when new knowledge is learned that allows a risk re-assessment. In this regard, the MBS power curve and the Estimation Parameter SD curve both suggest the five-observations-per-measured-variable standard should be revised downward.

In contrast, tradition is quite important in quality standards. It makes new knowledge more interpretable and old findings (obtained with different methods) more trusted when matched with the new. Fig. 1 shows the ten-observations-per-measured-variable standard to be inadequate when there is only one dependent latent and one independent latent. Otherwise, this long-standing, multiple-research-tool goal remains valid in Fig. 1 and is supported by the Hoogland–chi square and the SS power curves.

9. Expanded framework

The above analysis was largely conducted using elements from the SEM literature. To avoid myopia, Mulaik and James (1995) have previously connected SEM to the philosophy of science and general methodology literatures.

A similar effort is made here. The general methodological presentation of Brinberg and McGrath (1985) emphasizes a longer-term perspective. They analyze the limitations of a single study and show that researchers can only overcome these limitations in a multi-study research program. To discuss this prospect, a few elements from the Brinberg and McGrath framework can be added to Fig. 2 — see Fig. 4.


A. Overall Model
Substantive Uncontrollables
Conceptual Uncontrollables
Sample Size Standards
Substantive Controllables
Conceptual Controllables
Methodological Controllables
General Research Objectives
Statistical Objectives
B. Controllable Characteristics (Decision Variables) and Uncontrollable Characteristics
Substantive Domain
Controllables: 1. Sample Size (N) 2. Number of Studies Conducted
Uncontrollables: 3. Phenomena Size (Effect Size/Saturation/Variability), 4. Normality, 5. All Other Specification Errors
Conceptual Domain
Controllables: 6. Model Size (Number of Latent Variables, Measured Variables per Latent Variable, Parameters per Measured Variable, Measured Variable Domain Boundaries)
Uncontrollables: Other [Basic SEM Model Assumptions]
Methodological Domain
Controllables: 7. Estimation Method, Other [Choice of Overall Fit Measure, Type of SEM (Standard, PLS), Type of Scale, Specification Search, Choice of Normalization]
C. Connections of Three Statistical Objectives with Seven General Research Objectives
Estimate Parameter-Standard Error-Chi Square with Desirable Statistical Properties and with Avoidence of Convergence / Improper
Test Hypotheses with Adequate Power
Test Overall Fit with Adequate Power
1. Precision w.r.t. Measurement of Variables
2. Scope w.r.t. # of Dependent Vars & Their
3. Differentiation between Variables w.r.t. Detailed Characteristics & Domain Restrictions on Variables & between Parameters w.r.t. Values
4. Parsimony w.r.t. # of Measured & Latent Variables & w.r.t # of Unique Parameter Values
5. Realism w.r.t. Context (Background Variables)
6. Generalization w.r.t. Populations-Substantive Domain
7. Choice between Competing Theories

Note: w.r.t. abbreviates "with respect to".

Fig. 4. Sample size standards within an expanded context. Note: w.r.t. abbreviates “with respect to”.

Panel 4A of Fig. 4 contains a separate set of general research objectives. Controllables and Uncontrollables are also separated into three groups as belonging to either the substantive, conceptual, or methodological domains. In Panel 4B, one new variable, number of studies conducted, and one sub-variable of model size, measured variable domain boundaries, are added to the list of controllables. In Panel 4C, seven general research objectives are added. Three are single-study, short-term, tactical objectives: precision, scope, and differentiation. Two are intermediate: parsimony and realism. The last two are long-term, strategic objectives: generalization and choice between competing theories. Panel 4C exhibits only some of the many connections (dashed arrows) that relate the seven objectives to the three statistical objectives.

Panel 4C of Fig. 4 shows the sensitivity that the research process might have to the sample size standard selected by the research community. In Panel 4B, the fluid nature of conceptualization in many research areas is emphasized by adding the sub-variable, measured variable domain boundaries, to the list of controllables. Brinberg and McGrath discuss the breadth, degree of segmentation, and the amount of specific detail embodied in the decision process of conceptualization. As labeled in Panel 4C, these decisions involve precision, scope, differentiation, parsimony, realism as well as other issues along the route leading to choice between theories and to generalization of those theories to as large a population of entities as possible.

Brinberg and McGrath emphasize the conflicting desiderata that researchers cannot attain in a single study. They discuss two sets of tradeoffs. The solid arrows in Panel 4C show the connections between the seven desiderata. One conflict involves scope, differentiation, and parsimony. A second is between scope and precision. For this second conflict, we have added one of our concerns to the panel by also connecting generalization in this conflict. When sample collection is expensive, the traditional approach of controlling an experiment within a narrow domain before expanding its boundaries may not be followed. In many research examples, a more radical one-step merger of experiment/generalization occurs. In both of these sets of tradeoffs however, Brinberg and McGrath show that only a multi-study research program can achieve success.

In Panel 4B of Fig. 4, the substantive domain, i.e., the body of raw phenomena, is sampled in two ways. Sample size is a measure of the amount of sampling in a given study. However, the number of separate studies conducted is a new measure that captures the importance of a research program. In some instances, two small-sample studies may substitute for a single study having their combined sample size. The strategy of meta-analysis is built to some degree on this idea.

In other instances, two small-sample studies would actually improve upon a single study possessing a sample equal to the sum of the two smaller studies. Two small-sample studies using measured variables with different domains, having different models, and applying different methods can replicate and triangulate a research finding. This gives robustness by lessening alternative threats to validity of the finding (Brinburg and McGrath).

In still other instances, there is no substitution or complementing aspects. Naturally, two small-sample studies that investigate separate links in a chain do not triangulate one another's findings. Also when researchers publish one small-sample study, they do not sign a contract to triangulate the study in later work. More importantly, multi-study research reports are more difficult to publish than single-study research reports given the current peer review process. There would be more replication if the opposite were true.

Finally, if there were two small-sample studies with each relying upon low-power chi square statistics to test a hypothesis involving a small-effect phenomenon, they would both likely fail to reject the hypothesis. A single study using the combined sample would have greater power. In this case, two small-sample studies do not substitute for one larger study.

For all of these instances, a long-term, multi-study perspective is needed to insure that most research problems are solved successfully. In that long-term view, the sample size of one study may or may not be linked to that of a second study. The implication to be drawn from this discussion is as follows. As shown in Panel 4A, sample size standards are a front-end constraint on the research process. A myopic standard that considers only the application of SEM in a single study will likely restrain the inherent multi-study flexibility needed to attack non-SEM threats to the validity of the final research product. In contrast, a low, barebones sample size standard would allow this flexibility. If a separate quality standard accompanied the barebones standard, it would complement the barebones standard by identifying the relative strength of the various studies that are part of the multi-study effort.

10. Conclusion

This paper considered four different sets of sample size standards. The paper treated the four sets as largely independent. Their degree of dependence still awaits further research.

Actual empirical research was divided into four types: 1) the choice between competing theories where that choice only requires a nested hypothesis that can be individually decided based upon the value of a single parameter; 2) the choice between competing theories where that choice requires a nested

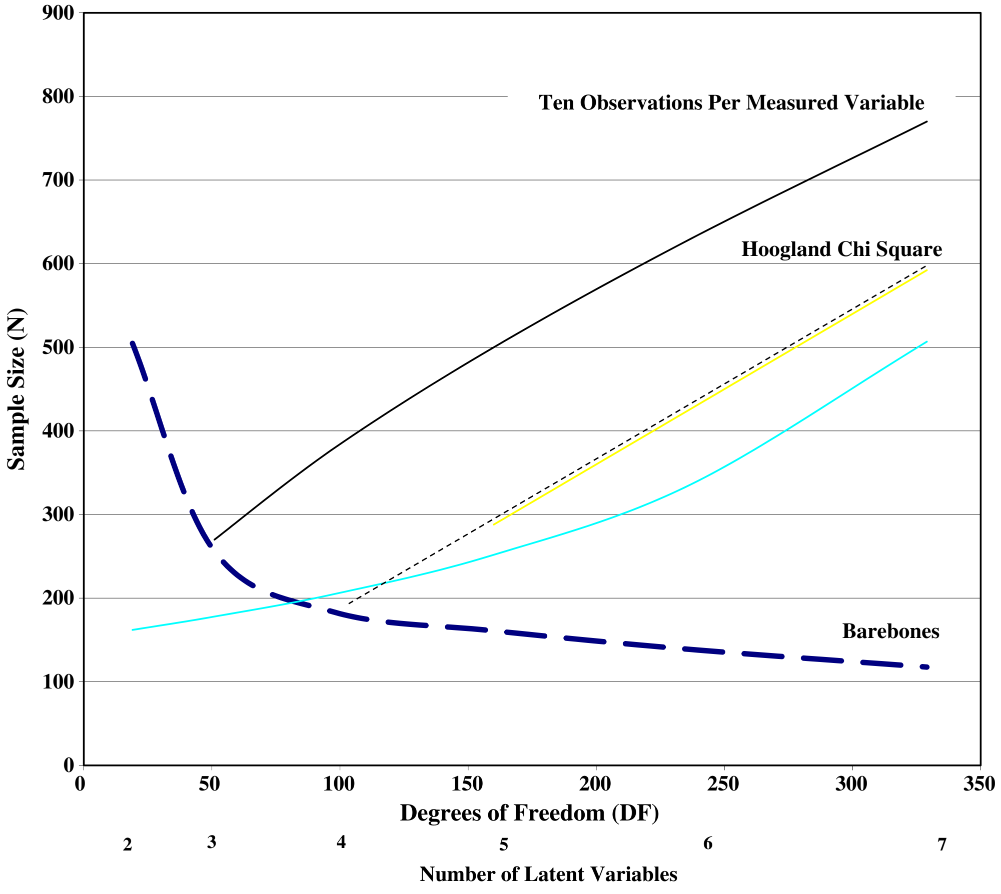


Fig. 5. Barebones & quality sample size standards. Barebones Standard: MBS power is effective for the first two data points (2 and 3 latent variables) and Estimation–Parameter SD is effective for the last four data points (4 through 7 latent variables). For this particular graph, see the note associated with Fig. 1 for assumptions that are in effect.

hypothesis that involves multiple parameter values; 3) the choice between competing theories where that choice requires comparisons of non-nested models; and 4) research where the scales development is the objective instead of the choice between competing latent variable models. The first type of choice can be made by comparing parameter estimate with standard error estimate — importantly using the most reliably estimated statistics of the analysis. The second type of choice requires the use of a sequential chi square statistic. The bias and variability of this statistic has not been determined and therefore awaits further research. The third type of choice requires comparisons of measures of overall fit such as the chi square for the entire model. Such statistics are taxing. Therefore the Hoogland standard is required for this category. The fourth type of choice requires parameter estimates and standard error estimates that are more sensitive to violation of the model assumptions. The multiple sample size standards for the measured variable loading and the variance of the measured variable disturbance should be considered.

This paper argues that Fig. 1 be replaced by a single barebones standard and an array of quality sample size standards. The barebones standard should offer a reasonable chance of attaining the objective of choosing between competing theories. Importantly, the term, reasonable chance, is considered here to mean that significant risk would still remain. To formulate this standard, the most current methodological research findings should be assessed and utilized. The general benefit of a barebones standard would be in the flexibility it would allow to the overall multi-study research process.

To offer a specific current version of a barebones standard, this paper suggests:

1) that this standard be applied only when a choice between theories can be made based upon the value of a single model;
2) that theory choice should be made by a simple comparison of parameter estimator with standard error estimator;
3) that sample size requirements for estimation should be satisfied only for the standard error estimators of β, γ, and Φ; (The higher requirements for the standard error estimators of λ and θ would only have to be satisfied when scale development was the objective.)
4) and that additionally the sample size requirement for MBS power be satisfied.

A special benefit of this specific proposal lies in its exploitation of the increasing returns to scale allowed in the SEM models applied. Research generally has more than enough factors favoring small size and narrowness. This proposal would free SEM to be a factor favoring larger-sized studies.

This paper argues as well for an array of quality standards to accompany the barebones sample size standard. Quality standards have different purposes from a barebones standard. They would identify relative strengths that distinguish some studies from others. It would also motivate researchers to obtain larger samples for analysis.

To offer a specific version of a quality standard, this paper suggests:

1) that the current barebones standard be satisfied;
2) and that additionally
    a) either the Hoogland chi square standard be met;
    b) or adequate SS power for medium saturation-medium sized effect be reached;
    c) or ten observations per measured variable standard be attained.

Each of the three quality standard would distinguish a given research study in a specific way. The two specific recommendations for barebones and quality standards would simplify Fig. 1 and in its place would be Fig. 5. This is a final determination of specific barebones.

There is a learning cost for something different, but is that the only downside of SEM? Instead we should ask, “when is SEM better in terms of data analysis and when is it worse?” SEM is definitely more detailed (in terms of more parameters) than traditional. Specifically, SEM gives separate estimates of λ, γ, and λ while traditional analysis only gives estimates of the products, $\lambda \cdot \gamma \cdot \lambda$ and $\lambda \cdot \beta \cdot \lambda$. When is having a more detailed model better as opposed to having a less detailed model? Clearly the more detailed is better (since it disentangles structural from measurement) when you have the information to disentangle the two. Does the SEM literature really give you any guidance as to when you have that information? In other words, the SEM literature does not contain tables suggesting that SEM is better here and traditional is better comparable. The questions are more in line with this presentation objective of introducing SEM to none as a familiar person.

References

Andrews WK. Power in econometric applications. Econometrica 1989;75:1059–90.

Baldwin B. A primer in the use and interpretation of structural equation models. Meas Eval Couns Dev 1989;22:100–12.

Bollen K. Structural equations with latent variables. New York: John Wiley & Sons; 1989.

Brinberg D, McGrath EJ. Validity and the research process. Beverly Hills, CA: Sage; 1985.

Browne MW, Cudeck R. Alternative ways of assessing model fit. In: Bollen KA, Long JS, editors. Testing structural equation models. Newbury Park, CA: Sage; 1993.

Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, New Jersey: Lawrence Erlbaum Associates; 1988.

Hoogland JJ. The robustness of estimation methods for covariance structure analysis. (PhD dissertation)Netherlands: Rijksuniversiteit Groningen; 1999.

Lomax R. Covariance structure analysis: extensions and developments. In: Thompson B, editor. Advances in social science methodology, Vol. 1. Greenwich, CT: JAI Press; 1989. p. 171–204.

MacCallum RC, Browne MW, Sugawara HM. Power analysis and determination of sample size for covariance structure modeling. Psychol Methods 1996;1(2):130–49.

Marsh HW, Hau Kit-Tai. Confirmatory factor analysis: strategies for small sample sizes. In: Hoyle RH, editor. Statistical strategies for small sample research. Thousand Oaks, CA: Sage; 1999.

Marsh HW, Hau K-T, Balla J. Is more ever too much: the number of indicators per factor in confirmatory factor analysis. Paper presented at the annual meeting of the American Eduicational Research Association, Chicago; 1997. (March).

Mulaik SA, James LR. Objectivity and reasoning in science and structural equation modeling. In: Hoyle, editor. Structural Equation Modeling; 1995.

Muthen BO, Curran PJ. General longitudinal modeling of individual differences in experimental designs: a latent variable framework for analysis and power estimation. Psychol Methods 1997;2(4):371–402.

Saris WE, Satorra A. Power evaluations in structural equation models. In: Bollen KA, Long JS, editors. Testing structural equation models. Newbury Park, CA: Sage; 1993.

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

Steiger JH. Point estimation, hypothesis testing, and interval estimation using the RMSEA: some comments and a reply to Hayduk and Glaser. Struct Equ Model 2000;7:49–162.

Fan, X., Thompson, B., and Wang, L., Effects of Sample Size, Estimation Methods, and Model Specification on Structural Equation Modeling Fit Indexes. Struct Equ Model, 6(1), 56-83.