
Volume 20, Number 1, January 2010 ISSN 1057-7408

ELSEVIER

Journal of
CONSUMER
PSYCHOLOGY

The Official Journal of
The Society for Consumer Psychology

Available online at www.sciencedirect.com
ScienceDirect
Journal of Consumer Psychology 20 (2010) 90–98
Journal of CONSUMER PSYCHOLOGY

Structural equations modeling: Fit Indices, sample size, and advanced topics

Dawn Iacobucci [1]

Department of Marketing, Owen Graduate School of Management, Vanderbilt University, 401 21st Avenue South, Nashville, TN 37202, USA

Available online 3 October 2009

Abstract

This article is the second of two parts intended to serve as a primer for structural equations models for the behavioral researcher. The first article introduced the basics: the measurement model, the structural model, and the combined, full structural equations model. In this second article, advanced issues are addressed, including fit indices and sample size, moderators, longitudinal data, mediation, and so forth.

Keywords: Structural equations models; Factor analysis; Path models; Statistical methods; Fit indices

Structural equations modeling (SEM) is an important tool for consumer psychology researchers. The *Journal of Consumer Psychology* presents this article (to accompany the previous Part I) to encourage more frequent and knowledgeable use of SEMs. The first article introduced the SEM modeling approach. This article illustrates several advanced uses of SEM, and closes with some comments about the limitations of SEM.

Advanced SEM issues will be discussed, including how to incorporate moderators, how to think about modeling longitudinal data and so forth. We begin with what are perhaps the two most important and pervasive technical issues—the perplexing choice among fit statistics and the question of requisite sample size.

Fit indices

One input when assessing a model is the examination of some fit statistic. When modeling with regression, a researcher reports an R^2. The R^2 is a descriptive index, and the evaluation of goodness-of-fit is somewhat subjective: Is $R^2=0.50$ good? Is $R^2=0.30$ good? Is $R^2=0.10$ good? There are no fixed guidelines for R^2, thus it is desirable to supplement it with the F-test. The F statistic can test a null hypothesis because it comes with a corresponding statistical distribution. Thus, the F-test tells us whether the model is capturing a significant amount of variance.

The issue of model evaluation explodes in SEM because of the plethora of fit indices. In the "Goodness of Fit Statistics" section of the output, Lisrel prints 38 indices. Each of these serves to optimize a slightly different objective function—the indices vary whether they are related to sample size or not, whether they assess absolute fit or fit relative to a benchmark model, whether they value parsimony or not (i.e., if it does, a function is built in which penalizes more complex models, those that estimate more parameters and use up more degrees of freedom). Together, these different indices provide complementary information. Gerbing and Anderson (1992) describe the situation as being analogous to the difficulty in answering the question, "What's the best car on the market?" The answer is that there is no one best car. The definition of "best" car depends on the objective: do you wish to drive a fast car, a stylish car, or a safe car?

So what's a good SEM modeler to do? This section offers guidance through the quagmire of fit statistics.

First, there is some agreement that researchers should report the following profile of indices: the χ^2 (and its degrees of freedom and p-value), the standardized root mean square residual (SRMR), and the comparative fit index (CFI). Ideally, for a model that fits the data, the χ^2 would not be significant ($p>0.05$), the SRMR would be "close to" 0.09 (or lower; Hu & Bentler 1999, p.27), and the CFI would be "close to" 0.95 (or higher; Hu & Bentler 1999, p.27). Let us examine these indices (the gory details are provided in Appendix I).

E-mail address: dawn.iacobucci@owen.vanderbilt.edu.

[1] I am grateful to friends, colleagues, and the SEM giants for their feedback on this research and manuscript: James C. Anderson, Bill Bearden, Richard Bagozzi, Hans Baumgartner, Peter Bentler, Bill Dillon, Jennifer Escalas, Claes Fornell, Steve Hoeffler, John Lynch, Robert MacCallum, Steve Posavac, Joseph Priester, and J. B. Steenkamp.

1057-7408/$ - see front matter © 2009 Society for Consumer Psychology. Published by Elsevier Inc. All rights reserved.
doi:10.1016/j.jcps.2009.09.003

Among the SEM fit indices, the χ^2 is the only inferential statistic; all the others are descriptive. That is, only for the χ^2 may we make statements regarding significance or hypothesis testing, and for the others, there exist only "rules-of-thumb" to assess goodness-of-fit. This quality may make it seem like χ^2 should be the only statistic to report. However, the χ^2 has its own problems. The most important of these is that the χ^2 is sensitive to sample size (Gerbing & Anderson 1985). While it is important to have a large sample to enhance the precision of parameter estimation, it is the case that as N increases, χ^2 blows up. A χ^2 will almost always be significant (indicating a poor fit) even with only modest sample sizes. As a result, it has been suggested, with some consensus in the psychometric literature, that a model demonstrates reasonable fit if the statistic adjusted by its degrees of freedom does not exceed 3.0 (Kline, 2004): $\chi^2/df \leq 3$.

SRMR stands for "standardized root mean square residual." Differences between data and model predictions comprise the residuals, their average is computed, and the square root taken. SRMR is a badness-of-fit index (larger values signal worse fit), and it ranges from 0.0 to 1.0. SRMR is zero when the model predictions match the data perfectly. SRMR is enhanced (lowered) when the measurement model is clean (high factor loadings; Anderson & Gerbing 1984, p.171). The index is a pretty good indicator of whether the researcher's model captures the data, because it is relatively less sensitive to other issues such as violations of distributional assumptions.

CFI is the "comparative fit index" and unlike the χ^2, which compares a model to data, the CFI takes the fit of one model to the data and compares it to the fit of another model to the same data. Hence, this kind of statistic captures the relative goodness-of-fit, or the fit of one's hypothesized model as an empirical increment above a simpler model (in particular, one in which no paths are estimated). Unlike the χ^2 and SRMR, the CFI is a goodness-of-fit index. It ranges from 0.0 to 1.0, and larger numbers are better. Also unlike the previous two indices, the CFI attempts to adjust for model complexity or parsimony. It does so by including the degrees of freedom used in the model directly into the computation (see details in Appendix I).[2]

Monte Carlo study

SEM scholars frequently use simulations to test certain relationships. The factor that concerns SEM modelers most about fit indices is sample size, so let us see an illustration of its effect on the three fit statistics just described. In this demonstration, a simulation study was run in which the design varied sample size from "probably too small" to "far larger than we typically see in *JCP*": N=30, 50, 100, 200, 500, to 1000. It is customary to test population models for confirmatory factor analysis, thus, a population covariance matrix was created based on two underlying factors, with three items loading on each factor, with loadings of 0.70, and a modest factor intercorrelation, φ=0.30. Then, for a given sample size, six normal random deviates were generated, and transformed by the population correlation matrix. The resulting data were analyzed via SEM, and the fit statistics noted. In each cell, 2000 such replications were created.[3] The general linear model, ANOVA specifically, was used to analyze the effect of sample size on the fit indices.

Results

The mean fits are presented in Fig. 1. As sample size increases, χ^2 increases ($F_{5,11994}$=251.39, p<.0001), and its corresponding p-value decreases ($F_{5,11994}$=745.89, p<.0001). The SRMR declines ($F_{5,11994}$=17,794.90, p<.0001) and the CFI is enhanced ($F_{5,11994}$=6.73, p<.0001, but the effect is negligible after 50).

Fig. 1 illustrates that the effect of sample size on χ^2 is nonmonotonic, exploding for large N (500 or 1000). The effect for SRMR is nearly linear—every new data point contributes to helping SRMR. The effect on CFI is nonlinear and the data suggest that a minimal sample of 50 may be beneficial, after which the boost tails off.

Sample size

In this section, we examine the question of sample size from the other angle, to answer the question, "How many observations are necessary for me to have a good SEM model?" Many potential users shy away from SEM because of the impression that sample sizes must be in the hundreds. It is true that "bigger is always better" when it comes to sample size. This truism holds particularly when the anticipated effects are subtle, the measures not especially clean or reliable, the structural model does not distinguish very clearly among constructs, etc. Notice what that statement implies—if the variables are reliable and the effects are strong and the model not overly complex, smaller samples will suffice (Bearden, Sharma & Teel 1982; Bollen, 1990).

To get a flavor of these interrelationships, consider the following. There was some thinking that strong, clean measures (as defined by the number of variables loading on each factor and the factors' reliabilities) would be somewhat compensatory for sample size, but while the number of variables per factor has an effect on improving fit statistics, its effect is modest compared to that of sample size (Jackson, 2003). Further, the effect may be nonmonotonic: Anderson and Gerbing (1991) found fit indices generally worsened as the number of factors in

[2] It might be overzealous; CFI tends to worsen as the number of variables increases (Kenny and McCoach 2003).

[3] Specifically, the Lisrel model was specified: $\Lambda_x' = \begin{bmatrix} 0 & 0 & 0 & 1 & 0.7 & 0.7 \\ 1 & 0.7 & 0.7 & 0 & 0 & 0 \end{bmatrix}$, $\Phi = \begin{bmatrix} 1.0 & 0.3 \\ 0.3 & 1.0 \end{bmatrix}$, $\Theta_\delta = diag[3,3,3,2,2,2]$. All other matrices involving endogenous variables were set to zero. This modelproduced the population covariance matrix, Σ. In the 6 experimental conditions, 2000 samples were generated with N observations and p=6 standard normal deviates. The population covariance matrix was factored via a standard eigen decomposition and used to transform the p-variate independent normals, i.e., $MVN_p(0,I)$ to create the proper intercorrelations, i.e., $MVN_p(0,\Sigma)$. In each sample, the SEM model was run to obtain the fit statistics. Empirical distributions were thus built with 2000 observations for estimates of fit indices.

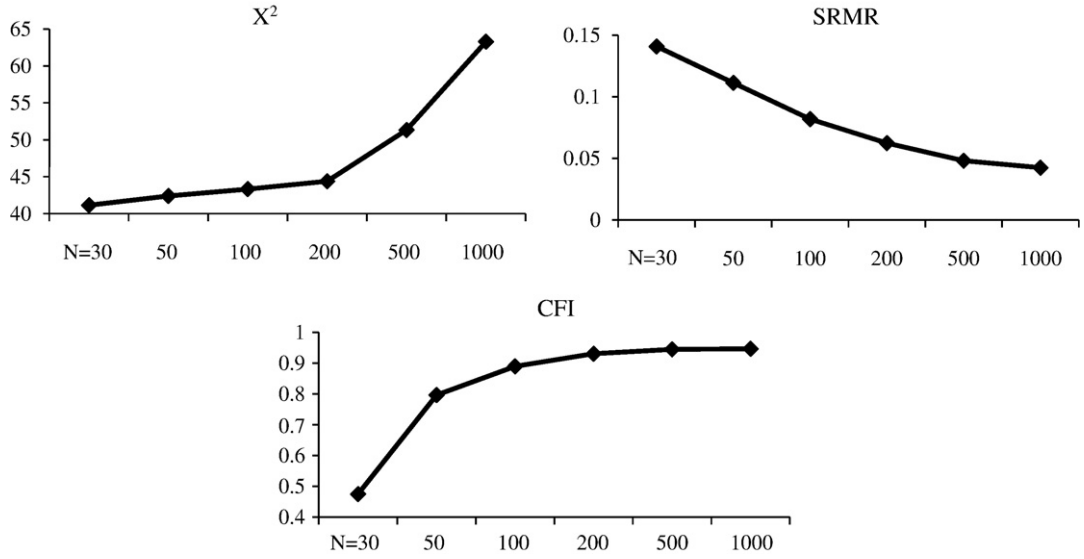

Fig. 1. Effect of sample size on popular fit indices.

the model, or number of variables per factor increased. If only two variables load on a factor, there will likely be bias in the parameter estimates, but "for three or more indicators per factor, this bias nearly vanishes" (Gerbing & Anderson 1985, p.268). In terms of bias reduction and even just getting the model to run, these authors found the added benefit that with "three or more indicators per factor, a sample size of 100 will usually be sufficient for convergence," and a sample size of 150 "will usually be sufficient for a convergent and proper solution" (Anderson & Gerbing 1984, pp.170–171).

Another inquiry along the lines of the interconnections among the characteristics of a model is the study on multicollinearity by Grewal, Cote and Baumgartner (2004). They considered high intercorrelations among exogenous constructs (r values of 0.6 to 0.8), and found there to be many type II errors in conclusions (e.g., paths not being significant when they should be), unless there were compensating strengths in the data, such as strong reliabilities and large samples. Both of these compensating factors serve to reduce overall error, lending more precision and confidence to the parametric estimation.

It is of some comfort that SEM models can perform well, even with small samples (e.g., 50 to 100). The vague, folklore rule of thumb considering requisite sample size, e.g., "$n>200$" can be conservative, and is surely simplistic.

The researcher particularly concerned with sample size can compute the desired N required for a given model (e.g., some determined number of variables, constructs, and therefore degrees of freedom) and desired level of power, or conversely, an estimate of power for a given N (Kim, 2005; MacCallum, Browne, & Cai 2006).

Different data scenarios

In this final section, we briefly cover some advanced topics: moderation, longitudinal data, higher-order factor analyses, mediation, reflective indicators and partial least squares models. SEM is discussed in a tremendous literature, including its own journal, *Structural Equation Models*. Thus, the treatment here of advanced topics is necessarily brief.

Moderators

First, users query how to introduce moderators into a SEM model. A moderator is simply an interaction term, and the approach in SEM is the same as in regression. The main effect variables are mean-centered, their product are computed, and all three are introduced as predictors (see Fig. 2). There may be theoretical interest in the main effects, but often their inclusion is merely as statistical controls to allow for a pure empirical focus on the interaction. This approach is general, allowing moderators that are categorical or continuous. If the moderator is categorical, another option is to run a multigroup analysis, in which a model is fit to one group's data and posited to either be numerically identical or qualitatively the same in the second group.[4]

Longitudinal data

Researchers with longitudinal data, such as repeated measures or within-subjects data can also use SEM. There are two kinds of coefficients that represent effects over time. Some are of theoretical interest, such as the effect in Fig. 3 of cognition at time 1 on affect at time 2. Other effects act as statistical controls, such as the autocorrelation effects between cognition at times 1 and 2 and those between affect measured at times 1 and 2. Autocorrelations are enabled by estimating

[4] The former is run by specifying that parameter values are "invariant," a strong form of cross-validation. The latter is run by specifying the "same pattern" of links—the links are the same in both groups but the parameter estimate values might vary—a weaker form of cross-validation.

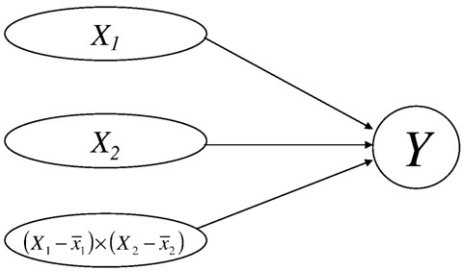

Fig. 2. Interactions to test moderators.

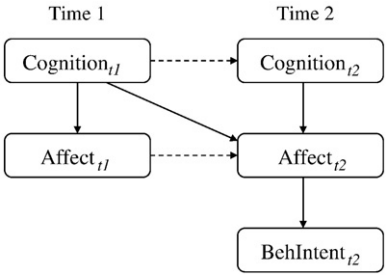

Fig. 3. Repeated measures.

correlations between errors, e.g., in θ_ε or θ_δ, which are normally assumed to be independent (Gerbing & Anderson 1984).

Higher-order factor analysis

When sufficient variables have been measured, a factor analysis can be followed up with another factor analysis where the second is conducted on the correlation matrix among the factors extracted in the first analysis. Those first factors are called "first-order" factors, and the next factors are the "higher-order" factors. Fig. 4 demonstrates an example of an eight-item survey, where four questions tap affect, and four tap cognition. In particular, two facets of affect are posited, and also for cognition. It is highly likely in such data that the two affect factors, AF1 and AF2, would also be correlated, as would be the two cognitive factors, CF1 and CF2. To fit a higher-order factor analysis in SEM, treat the variables as "*y*'s", the first-order factors as endogenous, and the higher-order factors as exogenous.[5]

Mediation analysis

A popular use of SEM is the examination of the process by which an independent variable *X* is thought to affect a dependent variable *Y*, directly, as $X \rightarrow Y$, or indirectly through a mediator, $X \rightarrow M \rightarrow Y$ (see Fig. 5). Traditionally, researchers have fit a series of regressions to estimate these relationships; but more recently, statistical researchers have shown the superiority of SEM in simultaneously and more efficiently estimating these relationships (Iacobucci, 2008). All three paths are fit at once, in a single model. The significance of the path coefficients would be tested, and if desired, compared in magnitude.

Reflective indicators

The factor analytic piece of the SEM model draws on a rich and extensive psychometric literature, dating back to the 1860s when Galton conceived of measuring intelligence. His philosophy, and that of Spearman and those factor modelers who followed, was essentially Platonic; the unobservable was the ideal, pure form, and the observed was a combination of the ideal and imperfections. Translating to our purposes, the unobservable, or latent factor was reflected in the observed, measured variables, and those variables were also affected by noise, in the form of systematic and random errors.

As Fig. 6 depicts, this philosophy is reflected in the direction of the arrows for the hypothetical constructs labeled, C, D, and E. (We will consider A and B shortly.) The construct, C (in the oval), such as intelligence, or attitude toward an ad, gives rise to the measures C1, C2, C3 (in the boxes). Errors also contribute to those measures, C1-C3. People with greater intelligence or more positive attitudes (C) are likely to score higher or more positively on the measures, C1–C3 than others who are less intelligent or less positive. The mapping is not perfect, and those imperfections are noted in the ε's (the error terms are not equal across C1–C3, but subscripts are eliminated for the sake of brevity). Also in the figure, the ζ's capture how well each endogenous variable is predicted—these structural errors are like $1-R^2$, thus, if knowing D helps us predict E very well, then E's ζ will be small.

Fig. 6 shows this reflective philosophy with three variables measuring construct C, and four measuring E. Occasionally, one might create a survey where no scale exists and a rough measure of a single indicator might be pragmatic. This scenario is represented for construct D, where only a single item, D1, is available to tap the construct. In this scenario, the measurement mapping is considered to be one-to-one, that is, the measure is essentially equated with the construct, hence the factor loading is the identity and the measurement error is set to zero. There is no situation when studying human behavior for which this situation actually holds—we will never have zero measurement error. Thus, single items are never optimal; however, they are sometimes used for practical reasons. We know multiple items are desired to tease apart the substantive or ideal part of D1 (in

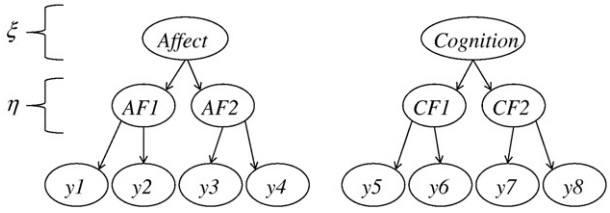

Fig. 4. Higher-order factor analysis.

[5] Gerbing and Anderson (1984) showed conditions under which a higher-order factor model may be equated with a model that incorporates correlated errors between variables loading on the same factor but concluded that typically the higher-order factor model was the proper model specification. In general, the only acceptable practice for allowing correlated errors is in the application to repeated measures and longitudinal data (as discussed previously).

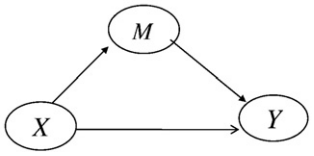

Fig. 5. Mediation.

D) from the noisy part of D1 (in ε). But with a single item, these two parts are confounded.

Similarly, recall that a part of the superiority of SEM over regression is that regression confounds prediction error with measurement error. The lack of fit in $1-R^2$ cannot be unambiguously attributed to a mis-specified model or to poor measures. In contrast, SEM allows for such distinctions: the factor analytical piece takes care of the measurement errors and the structural path modeling piece hosts the model prediction errors.

Some time ago, some researchers not versed in psychometric theory posited that the measurement arrows in Fig. 6 should go the other way. In their approach, variables combine to create a construct, in what is called "formative" measurement. The example that is routinely used (so frequently that one wonders whether another instantiation exists) is socioeconomic status (SES). They claim that education, income, and occupation combine to form SES and that a change in education, income, or occupation changes SES, but that the reverse is not true.[6]

These researchers propose to measure education, income, and occupation each with one item, as we saw construct D measured solely by D1, with the accompanying (invalid) assumption of perfect measurement. Next, a regression is modeled in which the education, income, and occupation items predict socioeconomic status. There is nothing wrong with regression. There is nothing wrong per se with single-item measures, albeit it is theoretically unsophisticated and empirically problematic. But "formative" as a new term or approach is unnecessary and misleading. Let us make clear that this proposed formative measurement is merely incomplete reflective measurement.

In Fig. 6, consider the constructs A and B to the left of the model. A formative approach would treat A and B like D (as single items), and the dashed arrows and boxes to the left would not exist. However, there is no reason that A and B cannot be considered as constructs in the traditional sense, latent factors which are reflected in multiple measures, as in the dashed relationships to A1–A3 and B1–B3. If A is the construct of education, A1 might be father's education, A2 mother's, A3 oldest child's. If B is income, B1 might be total household income, B2 might be the income of the major adult earner, B3 might be income based on soft money sources such as book royalties. Perhaps the measures of A1–A3 and B1–B3 are excessive, and typically a single rough estimate of A and B will suffice. However, the choice to use a single item for A and B for expediency purposes should not be confused with a theoretically unsubstantiated modeling choice that distorts a 150-year-old psychometric theory-laden tradition.

Thus, we see in Fig. 6 that the supposed new model may be subsumed in the more general, traditional reflective model. Scholars who defend the superiority of the reflective approach over the formative new comer decry the lack of theory supporting the new approach—there is no psychometric theory to support it. They point to a number of technical issues, such as problems in model identification. They also point to the fact that for the formative formulation, the measurement error and prediction error are once again confounded (cf., Bagozzi, 2007; Franke, Preacher, & Rigdon, 2008; Howell, Breivik, & Wilcox, 2007). In the formative approach, the observed variables are all thought to be measured without error, and the measurement error contributes instead to the factor itself, along with the factor's prediction error. In terms of measurement model development, it is a step backward.

Thus, let us write off the handful of papers in the literature which took us down an amusing little foray into pretending that formative indicators have more substance than the emperor's new clothes. Let us proceed as scientists would, drawing on theory, building models of data on a combination of ideal factors and impure errors—the reflective indicators model of factor analysis, the only defensible measurement model.

Partial least squares

As SEM is a combination of factor analysis and path modeling, partial least squares (PLS) is essentially a combination of principal components and path models (Fornell & Bookstein, 1982). Thus one of the distinctions is the measurement model—factor analysis is concerned with measurement theory, reliability and validity, etc. Principal components analysis creates linear combinations of variables, but not to model measures, instead only to predict the dependent variable(s) as best as possible.

Another distinction between SEM and PLS is in the computational method—SEM estimates are usually obtained via maximum likelihood, and PLS via least squares. This difference leads some statisticians to characterize PLS as being robust. For example, in theory, it can be used when the number of variables in the model exceeds the number of observations. However, note that this scenario poses less of a statistical limitation than a logical one. PLS pays a price, in that loadings tend to be overestimated and path coefficients underestimated (Dijkstra, 1983), but recall the goal of PLS is not the model coefficients per se, but the prediction and capturing the variance of the dependent variable. Hence PLS is useful, e.g., in predicting when consulting, but not for theoretical development or testing (McDonald, 1996).

Limitations of SEM models and other issues

Perhaps the first concern that potential users cite is, "Don't I have to have a huge sample?" If the measurement is strong (3 or 4 indicators per factor, and good reliabilities), and the structural

[6] This argument is already problematic because the question can be reformulated as a perfectly suitable reflective question; namely, if one were to profile those high in SES in one's data, generally they would have higher education, income, and such, compared to changing one's focus to profile those low in SES, where those people would largely have lower education, income, and such.

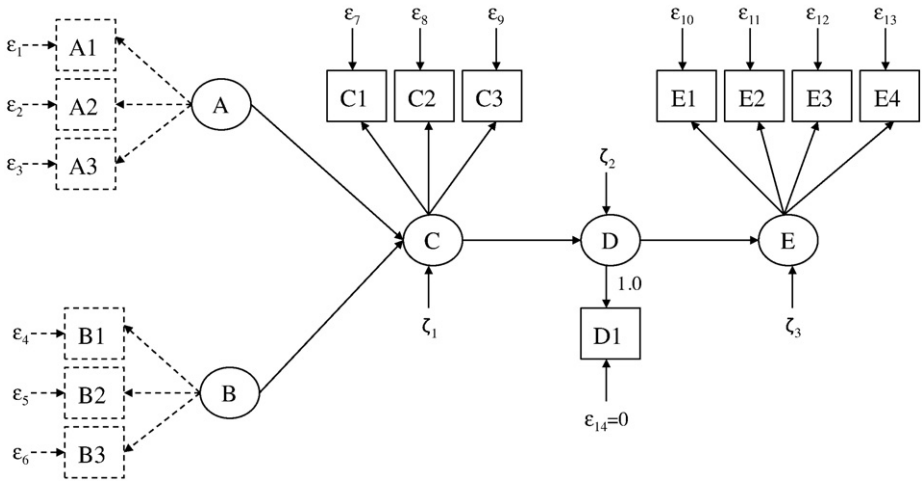

Fig. 6. Reflective measurement.

path model not overly complex (we cannot expect SEM to perform miracles), then samples of size 50 or 100 can be plenty.

Another concern is how to handle categorical data. It is well known that binary variables truncate the magnitudes of correlations (or covariances), the data which serve as the basis for SEMs. Alternatives exist. If all the variables in the model are discrete, the researcher can fit a log linear model. If some variables are discrete and the sample size is large, Muthén (1984) models polychoric correlations (for ordered variables). Research consistently says the correlations (and corresponding parameter estimates) are attenuated (i.e., underestimated), and standard errors and χ^2 values overestimated (Schumacker & Beyerlein, 2000), which is good news because all these results err in the statistically conservative direction. Indeed studies suggest that results based on categorical variables approximate those of their continuous counterparts, except in the extreme case where dichotomous variables were skewed in opposite directions (Ethington, 1987).

Still another concern that has been tracked in the psychometric literature is the method of estimation and the seemingly strict requirements of multivariate normality. Comparisons of estimation methods show maximum likelihood (ML) generally performs best, better than generalized least squares (GLS), and especially better than weighted least squares (WLS) (Ding, Velicer & Harlow, 1995; Olsson et al., 2000). ML has been found to be relatively robust (e.g., to violations of the multivariate normality assumption) and is generally endorsed for most uses (Hu & Bentler, 1998; Olsson et al., 2000). There is mixed evidence on the effect of non-normality: some say there is no effect on standard errors of parameter estimates (Lei & Lomax, 2005) but that there can be an effect on the parameter estimates themselves, if sample size falls below 100 (Lei & Lomax, 2005). Others say parameter estimates are fairly robust to non-normality (Finch, West, MacKinnon, 1997). Let us hope for continued evidence of robustness, because alternative modeling approaches, the asymptotically distribution-free methods, bootstrapping, other nonparametric methods need very large sample sizes (at least 4000–5000; Finch, West, MacKinnon, 1997), especially as data are more nonnormal. In general, distribution-free tests are said to perform "spectacularly badly" (Hu, Bentler & Kano, 1992). The bottom line is: stick to ML.

Conclusion

We close with a few suggestions regarding SEM. These comments are equally relevant to the researcher building and testing models as to the reviewer assessing a paper in which the authors had used SEM.

1. SEMs are not scary—they are natural progressions from factor analysis and regression.
2. As such, be careful not to over interpret path coefficients as if they were causal, any more so than if the results had been obtained via regression.
3. Shoot for a sample size of at least 50.
4. Ideally each construct would be measured by at least three indicator variables. If a few constructs are single items, that is probably okay. Constructs measured with four or more variables is probably excessive.
5. Use maximum likelihood estimation. (It is usually the default anyway.)
6. Check the fit statistics, but as Marsh et al. (2004) say: Do not take the rules-of-thumb too seriously. Do not be overly concerned with χ^2—it simply will not fit if the sample size is 50 or more. Instead, see if χ^2/df is about 3 or under. Do not be overly critical if the CFI is not quite .95, or the SRMR not quite .09.
7. On the other hand, ask good theoretical questions: Is every hypothesized link logically supported, and is there a sound, comprehensive yet parsimonious theoretical story for the entire model?
8. Fit at least one nontrivial competing model, presumably representing the extant literature on which the focal model is building, to see a demonstrable improvement.

SEM could be used more frequently among academics, and in industry wherever practitioners espouse conceptual models.

For more information, read Kline (2004) and Marcoulides and Schumacker (1996).

Appendix I: Fit Indices

SEM scholars distinguish two classes of fit indices: those that reflect "absolute" fit, and those that reflect a model's "incremental" fit, or the fit of one model relative to another. Absolute indicators of model fit include χ^2 and SRMR, among others. Incremental fit statistics include CFI, among others. Here are their definitions and basic behavioral properties.

Chi-square: χ^2

As Gerbing and Anderson (1992) describe it, what users refer to as the χ^2 is based on the "likelihood test statistic," the traditional statistical inferential measure of fit of a model on data, which when multiplied sample size (for large samples), is an index distributed χ^2 with degrees of freedom $=[k(k+1)/2]-t$ (where $k=p+q=$the number of observed endogenous variables plus the number of observed exogenous variables (note: variables, not constructs), and $t=$number of parameters estimated). This statistic tests the null hypothesis H_0: $\hat{\Sigma}=\hat{\Sigma}=S$, thereby reflecting the extent to which the residuals are zero. Specifically, the equation for χ^2 is:

$$X^2=\left\{N\left[tr\left(S\Sigma^{-1}\right)+\log|\Sigma|-\log|S|-(p+q)\right]\right\},$$

where S is the sample covariance matrix (computed on the data) and is the predicted covariance matrix (based on the model), and other terms as defined previously. Note that if the model produces predictions that closely mimic the data, then the and the will cancel each other. Similarly, would equal the identity matrix. The trace ("tr") of a matrix is the sum of the diagonal (the diagonal elements would equal 1.0 in an identity matrix, and there would be $p+q$ of them), so that term would cancel with the $-(p+q)$ term. Thus, a model that fits well would produce a χ^2 close to N. Hence the conclusion that χ^2 is sensitive to N.

Facts about χ^2 include the following: 1) It increases as function of df, hence the concern for N. The previous paragraph shows that even if the model fit very well, if a sample were say of size 1000, then the χ^2 would be approximately 1000. 2) χ^2 Ranges from zero to very high. It is zero when the saturated model is fit (i.e., all possible paths are in the model to be estimated). It is at its highest on any data set for the model of independence (i.e., no paths are entered into the model). 3) χ^2 Penalizes models with a large number of variables (i.e., it is large when there are many variables). 4) χ^2 Reduces as parameters are added to the model (much like an R^2 would increase as one adds predictors). However, adding parameters means the model is getting more complex, and less parsimonious. 5) χ^2 Can be used to compare the fits of nested competing models. We compute , where model A is a restricted version of B, and the result is distributed χ^2 with degrees of freedom equal to $(df_A)-(df_B)$. To say A is a restricted version of B is to say that model A is nested in model B; i.e., Model B estimates more parameters, whereas in model A, more parameters are fixed (usually to zero) and not estimated. The is also affected by N. If two models are not nested, they may be compared using descriptive goodness-of-fit measures, such as Akaike's Information Criterion (AIC) or an adjusted goodness-of-fit index (AGFI).

SRMR

RMR stands for "root mean square residual." The differences between the data in S and the model in $\hat{\Sigma}$ are called residuals. The average of these residuals is computed—on average, just how far off was the model. The square root of that value is taken—to put the index on a "standard deviation" scale, rather than a "variance" scale. The matrices S and $\hat{\Sigma}$ are typically (should be) covariance matrices, so the index is more easily interpreted if it is standardized (as if it were computed on a correlation matrix where the variances were equal to 1.0), so that it ranges from 0.0 to 1.0. The equations for the RMR and SRMR (the standardized root mean square residual) follow (Browne et al., 2002):

$$RMR=\sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}\left(s_{ij}-\hat{\sigma}_{ij}\right)^2}{k(k+1)/2}}$$

where $k=p+q$, and

$$SRMR=\sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}\left[\left(s_{ij}-\hat{\sigma}_{ij}\right)/\left(s_{ii}s_{jj}\right)\right]^2}{k(k+1)/2}}.$$

Like the χ^2, the RMR and SRMR are badness-of-fit indices—higher values indicate worse fits. If the model predicted the data fairly closely, then the residuals should be close to zero, making the numerator of RMR obviously zero (or zero squared), and the numerator of the numerator of SRMR similarly zero.

Regarding the SRMR: 1) Hu and Bentler (1999, p.27) suggest that an SRMR "close to .09" represents a reasonable fit (meaning in part that the model was not overly likely to have been the result of too many type I or type II errors). 2) In thorough simulation testing, the SRMR has been characterized as more sensitive to model misspecification than to sample size or violations of distributional assumptions. Thus, if SRMR is not as low as would be desired, the inflation is a fairly clear indicator that something is wrong with the (measurement and/or structural) model.

The RMSEA is an index that sounds somewhat like the SRMR but it is computed differently and it behaves differently (Steiger, 2000). The RMSEA is the "root mean square error of approximation": RMSEA $=\sqrt{\left(X^2-df\right)/df\,(N-1)}$. Unfortunately, it does not behave well. In simulation studies, RMSEA over-rejects true models for "small" N ($N<250$), the fit tends to worsen as the number of variables in the model increase, etc. (Fan & Sivo 2005; Hu & Bentler, 1998; Kenny & McCoach, 2003). Thus SRMR is preferred.

CFI and friends

The comparative fit index (CFI) is best understood in the context of its development. More than 30 years ago, the problem of large χ^2's and the seemingly non-informative state of nearly always rejecting the null hypothesis led researchers to develop other model evaluation criteria. In particular, Bentler and Bonett (1980) reasoned that an index should compare a model's fit not against a straw-model (the null) but against an idealized (yet still simple) model. Thus, this class of statistics became known as model comparison, or incremental fit indices (Bentler 1990).

The NFI (the normed fit index) is defined as follows: NFI = $(\chi^2_{\text{null}} - \chi^2_{\text{model}})/\chi^2_{\text{null}}$ and it ranges from 0.0 to 1.0. The χ^2_{model} is the fit of the model of interest, and the χ^2_{null} is the fit of the model of independence which estimates variances, but no covariances (i.e., there are no paths in the model between any constructs, and all the variables are thought to be independent). NFI was quickly trounced (it is influenced by sample size, it underestimates fit in small samples, it is difficult to compare across data sets, etc.; Ding, Velicer & Harlow 1995; Marcoulides & Schumaker 1996; Marsh et al. 1988), thus a new index was created to correct these shortcomings. The CFI ranges from 0.0 to 1.0, and its definition follows:

$$\text{CFI} = 1 - \{\left[\text{Max}\left(\left(\chi^2_{\text{model}} - \text{df}_{\text{model}}\right), 0\right)\right] / \left(\text{Max}\left(\chi^2_{\text{null}} - \text{df}_{\text{null}}\right), 0\right)]\}.$$

The comparison (by subtraction) of a model's χ^2 and its *df* is an adjustment for model parsimony. Models tend to fit worse (χ^2's are larger) when few parameters are estimated (i.e., when there are many *df*). Yet if a model fits well (the χ^2 is small), there is a penalty if that fit is achieved via an overly complex model (one with many parameters, using many *df*). Then, the comparison (by ratio) of the focal model to the null model reflects the extent to which something more interesting than independence is present in the current dataset. Instead, if there were nothing going on in the data, and in fact the independence model were true, the χ^2s (for model and null) would be similar, though the *df* might be different. If the *df* were similar, the entire ratio would be approximately 1.0, hence the CFI = 1 − 1, would be 0.0. Thus, a CFI gets larger as the model and data become more interesting, away from a simplistic model of independence.

The CFI has been said to be somewhat forgiving in exploratory modeling (Rigdon, 1996). Other indices in this class include TLI (Tucker–Lewis index), BL89 (Bollen's fit index), RNI (relative noncentrality index), gamma hat, and Mc (McDonald's centrality index). Overall, Hu and Bentler (1998) have demonstrated strong performance (power and robustness) of the CFI.

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.

Anderson, J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology*, *76*(5), 732–740.

Bagozzi, R. (2007). On the meaning of formative measurement and how it differs from reflective measurement: Commentary on Howell, Breivik, and Wilcox. *Psychological Methods*, *12*(2), 229–237.

Bearden, W. O., Sharma, S., & Teel, J. E. (1982). Sample size effects on chi square and other statistics used in evaluating causal models. *Journal of Marketing Research*, *19*(Nov.), 425–430.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606.

Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin*, *107*(2), 256–259.

Browne, M. W., MacCallum, R. C., Kim, C. T., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods*, *7*(4), 403–421.

Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling*, *2*(2), 119–144.

Dijkstra, T. (1983). Some comments on maximum likelihood and partial least squares methods. *Journal of Econometrics*, *22*, 67–90.

Ethington, C. A. (1987). The robustness of Lisrel estimates in structural equation models with categorical variables. *Journal of Experimental Education*, *55*(2), 80–88.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components. *Structural Equation Modeling*, *12*(3), 343–367.

Finch, J. F., West, S. G., & MacKinnon, D. P. (1997). Effect of sample size and nonnormality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling*, *4*(2), 87–107.

Fornell, C., & Bookstein, F. L. (1982). Two structural equation models: Lisrel and PLS applied to consumer exit-voice theory. *Journal of Marketing Research*, *19*(Nov.), 440–452.

Franke, G. R., Preacher, K. J., & Rigdon, E. E. (2008). The proportional structural effects of formative indicators. *Journal of Business Research*, *61*(12), 1229–1237.

Gerbing, D. W., & Anderson, J. C. (1992). Monte Carlo evaluations of goodness of fit indices for structural equation models. *Sociological Methods and Research*, *21*(2), 132–160.

Gerbing, D. W., & Anderson, J. C. (1985). The effects of sampling error and model characteristics on parameter estimation for maximum likelihood confirmatory factor analysis. *Multivariate Behavioral Research*, *20*, 255–271.

Gerbing, D. W., & Anderson, J. C. (1984). On the meaning of within-factor correlated measurement errors. *Journal of Consumer Research*, *11*(1), 572–580.

Grewal, R., Cote, J. A., & Baumgartner, H. (2004). Multicollinearity and measurement error in structural equation models: Implications for theory testing. *Marketing Science*, *23*(4), 519–529.

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods*, *12*(2), 205–218.

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

Hu, L., & Bentler, P. M. (1998). Fit indexes in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453.

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance Structure analysis be trusted? *Psychological Bulletin*, *112*(2), 351–362.

Iacobucci, D. (2008). *Mediation analysis.* Thousand Oaks, CA: Sage.

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling*, *10*(1), 128–141.

Kenny, D. A., & McCoach, B. D. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling*, *10*(3), 333–351.

Kim, K. (2005). The relation among fit indexes, power, and sample size in structural equation modeling. *Structural Equation Modeling*, *12*(3), 368−390.

Kline, R. B. (2004). *Principles and practice of structural equation modeling 2nd ed.* New York: Guildford.

Lei, M., & Lomax, R. G. (2005). The effect of varying degrees of nonnormality in structural equation modeling. *Structural Equation Modeling*, *12*(1), 1−27.

MacCallum, R. C., Browne, M. W., & Li, C. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, *11*(1), 19−35.

Marcoulides, G. A., & Schumacker, R. E. (Eds.). (1996). *Advanced structural equation modeling: Issues and techniques* Mahwah, NJ: Erlbaum.

Marsh, H. W., Hau, K. -T., & Wen, Z. (2004). In search of golden rules. *Structural Equation Modeling*, *11*(3), 320−341.

Marsh, H. W., Balla, J. R., & McDonald, R. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103*, 391−410.

McDonald, R. P. (1996). Path analysis with composite variables. *Multivariate Behavioral Research*, *31*(2), 239−270.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115−132.

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling*, *7*(4), 557−595.

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling*, *3*(4), 369−379.

Schumacker, R. E., & Beyerlein, S. T. (2000). Confirmatory factor analysis with different correlation types and estimation methods. *Structural Equation Modeling*, *7*(4), 629−636.

Steiger, J. H. (2000). Point estimation, hypothesis testing, and interval estimation using the RMSEA: Some comments and a reply to Hayduk and Glaser. *Structural Equation Modeling*, *7*(2), 149−162.