Available online at www.sciencedirect.com
ScienceDirect
Personality and Individual Differences 42 (2007) 831–839
PERSONALITY AND INDIVIDUAL DIFFERENCES
www.elsevier.com/locate/paid

Assessing the adequacy of structural equation models: Golden rules and editorial policies

Richard D. Goffin *

Department of Psychology, Social Science Centre, University of Western Ontario, London, Ont., Canada N6A5C2

Available online 7 November 2006

Abstract

Recommendations for enhancing the present journal's editorial policy with regards to the publication of Structural Equation Modeling (SEMing) studies have been provided in Barrett (2007). Wisely, these recommendations oppose strict reliance on "cut-off" values that have been proposed for well-known Approximate Fit Indices (AFIs). However, the present article critically evaluates other aspects of the recommended editorial policy in light of recent literature and finds a number of weaknesses: AFIs are ignored altogether, predictive accuracy suggestions do not take advantage of the SEMing literature, the chi-square test is overemphasized, power in SEMing analyses is underemphasized, and a number of important aspects of model assessment are not considered. Suggestions for a more comprehensive editorial policy regarding SEMing articles are provided.

Keywords: Structural Equation Modeling; Goodness of fit indices; Assessment of fit; Chi-square; Partial Least Squares

1. Introduction

An abundance of literature emphasizes the important advantages and enormous potential of structural equation modeling (SEMing) in psychological research (e.g., MacCallum & Austin, 2000; Tomarken & Waller, 2005). Nonetheless, the literature has also articulated a list of concerns

* Tel.: +1 519 661 2111x84641.
E-mail address: goffin@uwo.ca

0191-8869/$ - see front matter
doi:10.1016/j.paid.2006.09.019

regarding typical applications of SEMing, and many of these concerns focus on difficulties inherent in assessing the fit of SEMs (e.g., Tomarken & Waller, 2003). After considering a number of recent articles on the assessment of fit in SEMing, Barrett has proposed a series of recommendations for SEMing research published by *PAID*.

In the next section I will present the basic terminology and symbols that I use throughout this article. Following this, I will critically evaluate the main concerns and specific recommendations put forth in Barrett (2007) regarding appropriate editorial policies for SEMing articles.

2. Terminology and symbols

Much like Barrett (2007), I will use the term "Approximate Fit Indices" (AFIs) to refer to the multitude of indices that have been developed to assess the extent to which a SEM is consistent with empirical data. Note, however, that Barrett (2007, p. 816), erroneously implies that *all* AFIs have "... variously adjusted the chi-square test statistic ... " Browne, MacCallum, Kim, Andersen, and Glaser (2002), and Bentler (1995), provide examples of popular AFIs that are not derived from the χ^2 statistic.

The χ^2 null hypothesis significance test (χ^2NHST) is commonly available in SEM programs to test the null hypothesis that the residual matrix (described below) is zero in the population, which implies that the SEM being tested fits perfectly in the population.

Based on Kline (2005), "exogenous" refers to latent variables (constructs) that are not specified to be impacted or "caused" by any of the other latent variables in the model. "Exogenous" can also be used to describe directly-measured (i.e., observed) variables that serve as the indicators of the respective exogenous latent variables. "Endogenous" latent variables are constructs that are specified to be caused by at least one of the exogenous latent variables in the model. Similarly, endogenous directly-measured variables serve as the indicators of endogenous latent variables. "Measurement model" refers to the relations of the latent variables with their indicators whereas "structural model" refers to relations among the latent variables (Tomarken & Waller, 2003).

$\mathbf{S}$ symbolizes the covariance matrix among all the directly-measured variables. $\widehat{\Sigma}$ is the covariance matrix that is implied by a given structural equation model, whereas $(\mathbf{S} - \widehat{\Sigma})$ represents the residual matrix (Bollen, 1989). Some function of the difference between $\mathbf{S}$ and $\widehat{\Sigma}$ is successively minimized in most SEM programs as the freed parameters are estimated. This "discrepancy function" is most often that of maximum likelihood, in which case it is represented as $\boldsymbol{F}_{\mathrm{ML}}$.

3. Barrett's (2007) concerns about SEMing

3.1. Inappropriate reliance on AFI cutoffs in the assessment of SEMs

After conducting Monte Carlo simulations, Hu and Bentler (1999) provided a number of suggestions as to possible AFI cutoffs that may be indicative of a well fitting SEM. Despite Hu and Bentler's cautions that their cutoffs are preliminary and should not be overgeneralized, these cutoffs may have achieved the status of "Golden Rules" in the minds of some researchers (Marsh, Hau, & Wen, 2004). Barrett contends that reliance on these cutoffs as "proof positive" that a model is acceptable cannot be justified.

The preponderance of evidence in the articles Barrett cites and in additional works (e.g., Browne et al., 2002; MacCallum, Browne, & Sugawara, 1996; Sivo, Fan, Witta, & Willse, 2006; Tomarken & Waller, 2005) clearly supports Barrett's contention that AFI cutoffs should *not* have the status of Golden Rules. Collectively, current research convincingly demonstrates that Hu and Bentler's (1999) cutoffs do not have generality beyond the specific conditions and the fairly narrow range of "true" and misspecified models that Hu and Bentler included in their simulations. Moreover, Marsh et al. suggested that many of the presumably misspecified models that were used in Hu and Bentler's simulations, were, in fact, only trivially misspecified and would be considered acceptable on the basis of the cutoffs that Hu and Bentler later proposed. This is likely to have caused Hu and Bentler's cutoff recommendations to be overly conservative, leading to the rejection of many acceptable models in practice (Marsh et al., 2004). Thus, there is much support for Barrett's concern that it would be inappropriate to judge model fit purely on the basis of the AFI cutoffs suggested by Hu and Bentler.

Faced with mounting evidence that AFI cutoffs should not be strictly adhered to, other authors have emphasized a comprehensive approach to the assessment of fit which takes advantage of AFIs in conjunction with a wide range of other information about the SEM (e.g., McDonald & Ho, 2002; Tomarken & Waller, 2003). Similarly, the literature which Barrett cites as raising questions about AFIs still suggests that AFIs (although not necessarily AFI *cutoffs*) have a limited but important role to play in the assessment of SEMs (Beauducel & Wittmann, 2005; Fan & Sivo, 2005; Marsh et al., 2004; Yuan, 2005). Unfortunately, Barrett (2007) goes much further than suggesting that AFIs should be considered as just one part of the "Gestalt" of fit assessment. He states (p. 821): "In fact, I would now recommend banning *ALL* such indices from ever appearing in any paper as indicative of model 'acceptability' or 'degree of misfit'." The above quote, in combination with an apparent lack of mention of AFIs playing any part in Barrett's more specific recommendations (pp. 820–822), implies that he feels AFIs no longer have *any* legitimate role to play in the assessment of fit of SEMs. Regardless of the invaluable role that AFIs can still play in model comparison (including comparisons of *non-nested* models; Yuan, 2005), and the fact that evaluative comparisons of competing a priori models is one of the main contributions of SEMing (MacCallum, 2003), Barrett proposes that we ban AFIs and rely on "predictive accuracy" considerations (see below), or on the flawed χ^2NHST (see further below).

3.2. SEM fit assessment and predictive accuracy

In several places (e.g., pp. 817–818, p. 822) Barrett laments the fact that SEMing has little to do with "predictive accuracy." With respect to AFIs, he explains that a model can fit well regardless of whether it is able to predict outcomes or not – an AFI value that is in the "very acceptable" range says nothing of the model's ability to predict relevant outcomes. I will begin by summarizing the support for this concern and then provide what I feel are the more convincing points which suggest that Barrett's concerns over predictive accuracy in the context of SEMing are excessive.

There are two aspects of SEMing which contribute to Barrett's concern having some merit. First, *if* one were to rely *solely* on AFIs or the overall χ^2 NHST as a means of assessing SEMs, the extent to which the exogenous variables in the model were able to predict the endogenous variables would clearly have no direct bearing on model evaluations. This occurs because, in general, AFIs and the χ^2 NHST tend to support the acceptance of a model to the extent that $\widehat{\Sigma}$ is consistent with $\mathbf{S}$.

Moreover, a high degree of correspondence between $\widehat{\Sigma}$ and **S** is not a direct function of the relations between the exogenous and endogenous variables. It is, therefore, quite possible to achieve impressive AFI values and χ^2NHST results in spite of the fact that the relations between the exogenous and endogenous variables are trivial. Second, the most prevalent approaches to SEMing rely on algorithms that, to a large degree, optimize the overall fit of $\widehat{\Sigma}$ to **S** when estimating model parameters. Thus, the goal of SEM is *typically* to account for the covariance matrix of *all* the manifest variables. Barrett (e.g., pp. 817–818) implies that, by contrast, "regression-type" models are preferable to SEMs because they focus on the prediction of outcomes and can be evaluated strictly according to how well they predict consequential "real-world" outcomes.

Although the above paragraph finds some support for Barrett's concerns about predictive accuracy, the implications he draws are not sufficiently informed by the SEMing literature. First, with respect to the issue that model evaluation fails to incorporate predictive accuracy, Barrett's argument only holds to the extent that researchers rely *solely* on AFIs and/or the χ^2 NHST while ignoring a wealth of other information that is readily available for the evaluation of models. For example, through SEM output one can readily ascertain the extent to which the endogenous variables, which often comprise real-world outcomes such as behaviors, are predicted by the exogenous variables, simply by examining the total effects, direct effects, indirect effects and proportions of variance accounted for (e.g., Kline, 2005; McDonald & Ho, 2002; Tomarken & Waller, 2003). To the extent that exogenous variables fail to predict endogenous variables, or fail to predict them in a manner that is consistent with relevant theory, the model can be deemed less acceptable. Moreover, the fit of the structural model can be assessed separately from the fit of the measurement model to reflect the fact that many SEM investigators may be more interested in testing the predictive relations specified in the structural model than they are in testing the measurement model (McDonald & Ho, 2002; Tomarken & Waller, 2003). Taking this type of information into account has been an important and recommended component of the assessment of fit in SEMing for decades (e.g., Jöreskog & Sörbom, 1982; McDonald & Ho, 2002). Many researchers may fail to report this information in published articles, but this is reflective of a potential deficiency in the peer-review and/or editorial process rather than an inherent weakness of SEMing itself as Barrett suggests. Given his concerns about predictive accuracy, it is hard to understand why Barrett would not recommend that total effects, direct effects, indirect effects, proportions of variance accounted for, and the fit of the structural model (as distinct from the measurement model) be given greater emphasis in *PAID* SEMing articles.

A second problem with Barrett's reasoning regarding the issue of predictiveness is that he does not consider the full range of SEM approaches that are available. Barrett assumes SEMing *necessarily* optimizes the *overall* fit of $\widehat{\Sigma}$ to **S** and therefore fails to place a premium on the prediction of criteria. However, since its inception, the Partial Least Squares (PLS) approach to SEMing has adopted the goal of minimizing the residual variances of the endogenous variables rather than seeking to account for the covariation of the entire set of indicators as in typical SEMing (Haenlein & Kaplan, 2004). Granted, there have been concerns regarding PLS. In particular, PLS does not rely on the consistent minimization of any one criterion (Haenlein & Kaplan, 2004; Hwang & Takane, 2004). However, the Generalized Structured Component Analysis (GSCA) approach to SEM (Hwang & Takane, 2004) is a recently-developed successor to PLS which does consistently minimize a single criterion, namely, *the residual variance of all endogenous variables*. Also, in GSCA model fit is assessed according to how well the model predicts the variance of the endog-

enous variables. Thus, contrary to Barrett's position, SEM is not limited to approaches whereby a model can fit well regardless of whether it is able to predict outcomes.

Third, it appears that Barrett does not appreciate the benefits of minimizing the entire matrix of residuals in typical SEM projects. If the researcher's goal is to assess whether an entire hypothetical process or theory is more consistent with the data than a competing process or theory, it makes sense to consider whether the *entire* matrix of residuals is comparatively smaller (assuming that dfs are also taken into account), rather than just considering the residuals of the criterion variables. The process or theory in its entirety is evaluated in typical SEMs by considering the entire matrix of residuals, whereas approaches such as PLS (or GSCA) can be considered in situations where the focus is limited to the prediction of criteria. Further, as described earlier, decomposition of effects and the separate fit of the structural model can still be readily considered in typical SEMs to capture important elements of predictive accuracy.

Finally, the χ^2NHST, which Barrett supports, says nothing of the ability of a model to predict outcomes. Moreover, a careful read of Barrett's specific recommendations for model evaluation (pp. 820–823) finds that predictive accuracy (e.g., via his recommended BIC/AIC assessment or his "criterion variables" approach, pp. 822–823) is only to be evaluated in cases where the χ^2NHST is ignored. If predictive accuracy is such an important concern, it is hard to understand why Barrett would suggest that it only be considered in the event that the χ^2NHST is ignored. More generally, it seems incongruous to advocate renewed emphasis on the results of the χ^2NHST, which says nothing of the predictive accuracy of a model, while simultaneously mourning the lack of attention to predictive accuracy in SEMs and failing to emphasize existing SEMing avenues for assessing aspects of prediction (described earlier).

3.3. Can a "χ^2NHST Golden Rule" be supported?

Barrett strongly recommends that the χ^2NHST be given careful consideration in SEM projects unless there is a particular rationale for ignoring it. Despite his concerns about the Golden Rules that have evolved around Hu and Bentler's (1999) work, Barrett is essentially suggesting a χ^2NHST Golden Rule which could be stated as follows: provided the χ^2NHST assumptions are met, and $N \geqslant 200$, the χ^2NHST results should be taken at face value, unless the author chooses what could be called "Plan B" (discussed below).

There are serious problems with taking the χ^2NHST at face value. First, the null hypothesis upon which the χ^2NHST is predicated is unreasonable because it suggests that the proposed SEM could be expected to fit *perfectly* in the population. Given the state of our collective knowledge in most areas of psychology, and the complexity of many psychological processes, it is unreasonable to expect that most hypothesized SEMs could fit perfectly in the population (Browne & Cudeck, 1993; MacCallum, 2003; MacCallum et al., 1996). Structural equation models are best regarded as potentially useful approximations of reality, not perfect reflections of it. A model that provides some semblance of the truth can have much value in guiding further theoretical and practical developments, but the models we develop in psychology should virtually never be presumed to contain the whole truth and therefore be subjected to a test of perfect fit (MacCallum, 2003; MacCallum et al., 1996).

Second, given the unreasonable null hypothesis of a zero residual matrix in the population, and the fact that the researcher seeks to accept the null hypothesis, *power becomes the sworn enemy of*

the SEM researcher who is forced to rely on the χ^2NHST. The χ^2NHST simply indicates whether the SEM analysis possessed sufficient power to reject a null hypothesis which, in the vast majority of cases, is already known to be untrue (MacCallum, 2003). Barrett (p. 820) is aware that "huge" samples may increase power to the point where reasonable models may be rejected but he fails to appreciate the inevitability of virtually all models failing the χ^2NHST given sufficient power. As a compromise, Barrett (pp. 821–823) offers what I call "Plan B" to authors whereby they can ignore the χ^2NHST provided that an argument pertaining to N being very large is formulated. If Plan B is chosen, Barrett (pp. 821, 823) still recommends that AFIs be "banned" as suggestive of model acceptability, but he supports the use of "cross-validated predictive accuracy and model parsimony via AIC/BIC indices" (p. 822). As discussed in detail above, predictive accuracy is, indeed, one of many things that an author can and should consider in a comprehensive analysis of SEM acceptability. Nonetheless, as also discussed above, a researcher may have good reason to also be concerned with whether an entire hypothetical process or theory is reasonably consistent with the data, in which case overall fit, as indicated via AFIs, may deserve some consideration.

Third, assuming that the χ^2NHST Golden Rule is followed and Plan B is avoided in many cases, there is likely to be a flood of models which pass the χ^2NHST largely because they have an overabundance of freed parameters. Because F_{ML}, and virtually all common discrepancy functions, cannot increase and are almost certain to decrease as more freed parameters are added to a model (Bollen, 1989; Goffin, 1993), one can typically increase the chance of a "not significant" χ^2NHST by freeing more parameters (e.g., see MacCallum, Roznowski, & Necowitz, 1992). Model modification aids included in popular SEM programs contribute to the temptation to add freed parameters that will improve the chance of passing the χ^2NHST by taking advantage of sample-specific fluctuations in the data that result in no necessary improvement in the true validity of the model (MacCallum et al., 1992). Thus, by placing a premium on a not significant χ^2NHST result, the χ^2NHST Golden Rule is likely to result in the proliferation of models that capitalize on sample-specific variation. In fairness, Barrett shows awareness of the importance of cross-validation. Nonetheless, in cases where the χ^2NHST yields a not significant result, his recommendations fail to provide any protection against the upsurge in overparameterized models that is likely to follow as a logical consequence of the χ^2NHST Golden Rule – he does not prescribe cross-validation except in the case of models which "fail" the χ^2NHST. Considering some function of the extent to which $\hat{\Sigma}$ differs from $\mathbf{S}$, in conjunction with df, provides a perspective on whether a model may be capitalizing on sample-specific variation through the inclusion of "wastebasket" parameters (i.e., parameters which have no reasonable substantive meaning; Browne, 1982). However, such considerations form the basis of many of the AFIs that Barrett seeks to ban. In sum, Barrett's suggestion that the χ^2NHST be taken seriously, will lead to models with an excess of freed "wastebasket" parameters ruling the day in *PAID*.

Finally, Barrett rejects the use of AFIs as indicators of model fit partially on the grounds that the several papers he cites have found problems with these indices. Admittedly, research has found AFIs to be sensitive to certain conditions (Yuan, 2005). This casts doubt on the viability of rigid AFI cutoffs, but not on the viability of AFIs in general for purposes such as model comparison (Yuan, 2005). Moreover, Barrett fails to reject the χ^2NHST as a viable test and asserts that it be taken at face value (if assumptions are met) even though, in addition to the above problems, it has also been persuasively shown to be sensitive to certain conditions. Browne et al. (2002) found that

the χ^2NHST, when used in combination with the most popular estimation method (maximum likelihood), is more likely to lead to the rejection of models with measured variables that have small unique variances and highly reliable indicators, than to the rejection of models with measured variables that have large unique variances and highly unreliable indicators. This is a serious problem which contributes to the χ^2NHST not being interpretable at face value because it essentially disadvantages researchers to the extent that they have gone to the trouble of using more reliable indicators. On the other hand, Browne et al. found that a well-known AFI called the Root Mean Square Residual did not suffer at all from this form of bias – it reflected model fit regardless of whether the indicators were very reliable or not. Similarly, two other well-known AFIs, the Normed Fit Index and the Relative Non-centrality Index, showed drastically less tendency to indicate poor fit in the case of models with more reliable indicators than did the χ^2NHST. It is difficult to staunchly support the use of the χ^2NHST and the banning of AFIs in the face of evidence such as this – particularly when conditions such as the reliability of the indicators are not taken into account in Barrett's recommendations.

3.4. Evaluation of power and the $N \geqslant 200$ rule in SEM fit assessment

Barrett (2007, p. 821) suggests that power estimation in SEMing is not currently feasible for most researchers, but he recommends the rejection of SEM submissions with $N < 200$ (unless the respective population is "small or restricted") and considers $N \geqslant 200$ to be reasonable. In contrast to Barrett's recommendations, although there is not total unanimity, recent literature shows a trend toward embracing the assessment of power in SEMing (e.g., MacCallum & Austin, 2000; McQuitty & Bishop, 2006; Tomarken & Waller, 2003, 2005) and is generally sanguine about the recent groundbreaking work that has occurred on this front (e.g., MacCallum, Browne, & Cai, 2006; MacCallum et al., 1996; MacCallum & Hong, 1997; Muthen & Muthen, 2002; Paxton, Curran, Bollen, Kirby, & Chen, 2001). Moreover, collectively, this research contributes to the perspective that rules of thumb such as "$N \geqslant 200$ is sufficient" are far too simplistic because many factors other than N have a potent effect on power in SEMs. For example, if N were held constant at somewhere in the 200–300 range, power could easily range anywhere from .3 to .9 depending on factors such as df and number of indicators per latent variable (e.g., see MacCallum et al., 1996; Tomarken & Waller, 2003). What purpose is served by a $N \geqslant 200$ cutoff that will likely result in models with $N = 200$ but power less than .3 being published, and models with $N = 175$ but power greater than .8 being rejected out of hand? Also, as already alluded to above, Browne et al. (2002) have shown that models with more reliable indicators tend to engender greater sensitivity for detecting non-zero residuals and therefore have a higher probability of failing the χ^2NHST. Thus, by recommending that power analysis is intractable and that $N \geqslant 200$ is mostly what matters, Barrett will disadvantage researchers who adopt more reliable measurement procedures.

I do not mean to imply that power determination in SEMing has been perfected. Nonetheless, the research cited above has provided evidence that not only N, but also df, the reliability of the indicators, the number of indicators per latent variable, and other aspects of the model contribute to considerations of power. Thus, it is hard to appreciate how a simple $N \geqslant 200$ rule-of-thumb could find reasonable support in recent empirical and statistical SEMing literature.

4. Summary and conclusions

Although Barrett's concern that suggested AFI cutoffs should not be rigidly adhered to is justifiable, there are many ways in which his specific set of recommendations are incongruous with the SEMing literature. The issues covered above reflect serious core problems with his recommendations, however, there are at least three additional issues that space restrictions disallowed me from covering in detail. First, Barrett (2007, p. 821), suggests successively modifying models based on consideration of the residuals, but it is this type of post hoc model-fitting that is likely to lead to capitalization on chance and the proliferation of meaningless models (e.g., MacCallum et al., 1992). Second, Barrett (2007, p. 822), seems to generally reject the worth of CFA and EFA in assessing the internal structure of psychometric instruments even though there is obvious value in such endeavors from the viewpoint of establishing construct validity (e.g., see Kline, 2005). Third, whereas Barrett (e.g., p. 818) refers to a "multifaceted" approach to the assessment of SEMs, his specific recommendations overstress the χ^2NHST and predictive accuracy aspects of a model at the expense of other important elements of model evaluation (e.g., see Tomarken & Waller's, 2003, principles described below).

This article was intended to provide a critical analysis of Barrett (2007). After having done so, there is insufficient space to comprehensively explicate an alternative to his recommendations. Fortunately, Tomarken and Waller (2003) have presented an informed and thoughtful articulation of a comprehensive approach to SEM assessment. While still emphasizing that AFIs should not provide the sole basis for decisions regarding model adequacy, their article provides a set of guiding principles for editorial decisions regarding SEMs that is both comprehensive and informed by the collective SEMing literature. Issues such as due consideration of equivalent alternative models; omission of important variables; decomposition of effects and the separate fit of the structural and measurement models (as discussed above); sensitivity of one's design and analysis to detecting model misspecification (i.e., power); and a priori versus post hoc model specification, are given appropriate voice in Tomarken and Waller's recommendations but are not apparent in Barrett's. Consequently, at this point in time Tomarken and Waller (2003) would appear to provide a more defensible foundation for *PAID*'s SEMing editorial policy than Barrett (2007) does. But there is one additional desideratum that makes sense in regards to *PAID*'s editorial policy. Currently, *PAID*'s 5000-word limit may well prevent researchers who use complex statistical methodologies such as SEMing from adequately complying with comprehensive but important guidelines such as Tomarken and Waller's. Perhaps the editorial board could consider loosening the strict 5000-word limit in the case of studies that employ more complex methodologies such as SEM?

References

Barrett, P. (2007). Structural equation modelling: adjudging model fit. *Personality and Individual Differences, 42*(5), 815–824. doi:10.1016/j.paid.2006.09.018.

Beauducel, A., & Wittmann, W. (2005). Simulation study on fit indices in confirmatory factor analysis based on data with slightly distorted simple structure. *Structural Equation Modeling, 12*, 41–75.

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in multivariate analysis* (pp. 72–141). Cambridge, England: Cambridge University Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage Publications.

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*, 403–421.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indices to misspecified structural or measurement model components: rationale of two-index strategy revisited. *Structural Equation Modeling, 12*(3), 343–367.

Goffin, R. D. (1993). A comparison of two new indices for the assessment of fit of structural equation models. *Multivariate Behavioral Research, 28*, 205–214.

Haenlein, M., & Kaplan, A. M. (2004). A beginner's guide to partial least squares analysis. *Understanding Statistics, 3*, 283–297.

Hu, Li-tze, & 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.

Hwang, H., & Takane, Y. (2004). Generalized structured component analysis. *Psychometrika, 69*, 81–99.

Jöreskog, K. G., & Sörbom, D. (1982). Recent developments in structural equation modeling. *Journal of Marketing Research, 19*, 404–416.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research, 38*, 113–139.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*, 201–226.

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: power analysis and null hypotheses. *Psychological Methods, 11*, 19–35.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130–149.

MacCallum, R. C., & Hong, S. (1997). Power analysis in covariance structure modeling using GFI and AGFI. *Multivariate Behavioral Research, 32*, 193–210.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin, 111*, 490–504.

Marsh, H. W., Hau, Kit-Tai, & Wen, Z. (2004). In search of Golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*, 320–341.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*, 64–82.

McQuitty, S., & Bishop, J. W. (2006). Issues in multi-item scale testing and development using structural equation models. *Journal of Applied Measurement, 7*, 117–128.

Muthen, L. K., & Muthen, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*(4), 599–620.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: design and implementation. *Structural Equation Modeling, 8*, 288–312.

Sivo, S. A., Fan, X., Witta, E. L., & Willse, J. T. (2006). The search for "optimal" cutoff properties: fit index criteria in structural equation modeling. *The Journal of Experimental Education, 74*, 267–288.

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with well fitting models. *Journal of Abnormal Psychology, 112*, 578–598.

Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology, 1*, 31–65.

Yuan, K. H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research, 40*(1), 115–148.