STRUCTURAL EQUATION MODELING, *11*(3), 320–341

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

Herbert W. Marsh
SELF Research Centre
University of Western Sydney

Kit-Tai Hau
Department of Educational Psychology
The Chinese University of Hong Kong

Zhonglin Wen
Department of Psychology
South China Normal University

Goodness-of-fit (GOF) indexes provide "rules of thumb"—recommended cutoff values for assessing fit in structural equation modeling. Hu and Bentler (1999) proposed a more rigorous approach to evaluating decision rules based on GOF indexes and, on this basis, proposed new and more stringent cutoff values for many indexes. This article discusses potential problems underlying the hypothesis-testing rationale of their research, which is more appropriate to testing statistical significance than evaluating GOF. Many of their misspecified models resulted in a fit that should have been deemed acceptable according to even their new, more demanding criteria. Hence, rejection of these acceptable-misspecified models should have constituted a Type 1 error (incorrect rejection of an "acceptable" model), leading to the seemingly paradoxical results whereby the probability of correctly rejecting misspecified models decreased substantially with increasing *N*. In contrast to the application of cutoff values to evaluate each solution in isolation, all the GOF indexes were more effective at identifying differences in misspecification based on nested models. Whereas Hu and Bentler (1999) offered cautions about the use of GOF indexes, current practice seems to have incorporated their new guidelines without sufficient attention to the limitations noted by Hu and Bentler (1999).

Requests for reprints should be sent to Herbert W. Marsh, SELF Research Centre, University of Western Sydney, Bankstown Campus, Locked Bag Penrith South DC, NSW 1797. E-Mail: h.marsh@uws.edu.au

Quantitative social scientists, particularly psychologists, have long been engaged in the elusive search for universal "golden rules"—guidelines that allow applied researchers to make objective interpretations of their data rather than being forced to defend subjective interpretations on the basis of substantive and methodological issues. Their appeal—like the mythical Golden Fleece, the search for the fountain of youth, and the quest for absolute truth and beauty—is seductive, but unlikely to be realized. Whereas we do not argue that psychologists must desist in their elusive quest for golden rules, we do, however, contend that they should avoid the temptation to treat currently available "rules of thumb" as if they were golden rules. Instead, applied researchers must contend with rules of thumb that provide, at best, preliminary interpretations that must be pursued in relation to the specific details of their research. Ultimately, as emphasized by Marsh, Balla, and McDonald (1988), Hu and Bentler (1998, 1999), and many others (e.g., Bentler & Bonett, 1980; Browne & Cudeck, 1993; Byrne, 2001; Jöreskog & Sörbom, 1993; Steiger & Lind, 1980) data interpretations and their defense is a subjective undertaking that requires researchers to immerse themselves in their data.

An important reason for the popularity of goodness-of-fit (GOF) indexes to assess the fit of models in covariance structure analyses is their elusive promise of golden rules—absolute cutoff values that allow researchers to decide whether or not a model adequately fits the data—that have broad generality across different conditions and sample sizes. Starting with the widely cited classic research by Bentler and Bonnett (1980), the basic goal of particularly the family of incremental fit indexes was to provide an index of GOF that varied along an interpretable metric with absolute cutoff values indicating acceptable levels of fit. The incremental fit indexes offered what appeared to be a straightforward evaluation of the ability of a model to fit observed data that varied along an easily understood, 0 to 1 scale. Based on intuition and a limited amount of empirical research, they proposed that incremental fit indexes of .90 or higher reflected acceptable levels of fit. Other research stemming from early work by Steiger and Lind (1980) argued for the usefulness of indexes based on a standardized measure of empirical discrepancy (the difference between $\hat{\Sigma}_k$ and S, where $\hat{\Sigma}_k$ is the fitted covariance matrix derived from the sample covariance matrix S based on k parameter estimates) such as root mean square error of approximation (RMSEA). Whereas such indexes have no clearly defined upper limit, the lower limit is zero in the population. A combination of intuition and experience led researchers (e.g., Browne & Cudeck, 1993; see also Jöreskog & Sörbom, 1993) to suggest that RMSEA less than 0.05 is indicative of a "close fit," and that values up to 0.08 represent reasonable errors of approximation. However, as emphasized by McDonald and Marsh (1990; Marsh, Hau, & Grayson, in press) and many others, the traditional cutoff values (e.g., incremental fit indexes > .90) amount to little more than rules of thumb based largely on intuition, and have little statistical justification.

Hu and Bentler (1998, 1999) addressed this issue in a highly influential study that continues to have a substantial effect on current practice. They used the ability of in-

dexes to discriminate between correct and misspecified models as a basis for recommending new, more stringent cutoff values for many GOF indexes. The impact of their research was twofold. First, they provided an apparently much stronger empirical basis for evaluating the validity of decisions based on cutoff values. Second, their research is apparently leading to the routine use of much more stringent cutoff values for what constituted an acceptable fit. We argue, however, that there are important logical issues and inconsistencies in their research that undermine the appropriateness of having a single cutoff value for each index that generalizes across different sample sizes (*N*) and different situations. Moreover, if their logic was appropriate, we show that for normally distributed data, a traditional maximum likelihood (ML) chi-square test would have outperformed all of their GOF indexes in relation to their stated purpose of optimally identifying a misspecified model.

More important, we do not address many of the issues in the extensive scope of Hu and Bentler' (1998, 1999) research (e.g., robustness of rules in relation to nonnormality, alternative estimation methods, combination rules involving multiple indexes of fit, empirical correlations among values for different indexes) and other potentially relevant issues that were not emphasized by Hu and Bentler (1998, 1999; e.g., confidence intervals for GOF indexes and notions of close fit and "not-close fit"; see MacCallum, Browne, & Sugawara, 1996) that are beyond the scope of our brief comment. It is even more important to emphasize that Hu and Bentler (1998, 1999) certainly never suggested that their new guidelines should be interpreted as universal golden rules, absolute cutoff values, or highly rigid criteria that were universally appropriate. Quite the contrary, they stressed potential limitations in the application of their guidelines, the need for researchers to select guidelines most appropriate to their area of research, concerns about the generalizability of their results, and limitations in their study that require further consideration and research (e.g., the nature of misspecified models, the role of parsimony, concerns about local misspecification even when satisfactory levels of global fit are achieved). Indeed, Hu and Bentler (1998) specifically noted that "the performance of fit indices is complex and additional research with a wider class of models and conditions is needed" (p. 446); "it is difficult to designate a specific cutoff value for each fit index because it does not work equally well with various types of fit indices, sample sizes, estimators, or distributions" (p. 449); and "Hu and Bentler (1997) found that a designated cutoff value may not work equally well with various types of fit indexes, sample sizes, estimators, or distributions" (1999, p. 16). Hence, much of the blame for any inappropriate use of Hu and Bentler's (1998, 1999) research clearly lies with those who have misinterpreted Hu and Bentler (1998, 1999) and the field as a whole that has been searching for golden rules, not Hu and Bentler (1998, 1999) themselves.

However, despite the many cautions offered by Hu and Bentler (1998, 1999), it is our impression that many consumers of their results such as applied researchers, textbook authors, reviewers, and journal editors have inappropriately promoted

their new, more stringent guidelines for acceptable levels of fit into something approaching the golden rules that are the focus of comment. Therefore, for example, in their review of the appropriate use of factor analysis, Russell (2002) recently noted, "Clearly, the results reported by Hu and Bentler (1999) indicate the need to use more stringent criteria in evaluating model fit" (p. 1640) and that for incremental fit indexes such as Tucker–Lewis Index (TLI), comparative fit index (CFI), and Relative Noncentrality Index (RNI), the "widely used criterion of .90 or greater should be increased to .95 or greater" (p. 1640). In their applied research, King et al. (2000) noted that, "convention has dictated that values of such indices exceeding .90 reflect reasonable model–data fit, and more recent thinking (Hu & Bentler, 1998) has mandated values above .95 as preferred" (p. 628). Hemphill and Howell (2000) stressed the importance of "using contemporary criteria (see Hu & Bentler, 1998, 1999) to interpret [their] statistical results" (p. 374). Whereas Byrne (2001) also advocated the use of Hu and Bentler's (1999) new criteria of a good fit in her structural equation modeling (SEM) textbooks, she further emphasized that global fit indexes were only one component in the evaluation of model adequacy. In this respect, our comments are directed primarily toward potential problems apparently being incorporated into current practice without appropriate caveats rather than the highly influential and heuristic research by Hu and Bentler (1999) and their more cautious recommendations.

RATIONALES FOR ESTABLISHING GOF CUTOFF VALUES

How good is good enough? There is an implicit assumption in GOF research that sufficiently high levels of GOF (higher than a prescribed cutoff value) are necessary to establish the validity of interpretations of a model. Although nobody seriously argues that high GOF is sufficient to conclude that a model is valid, current practice seems to treat high GOF as if it were both necessary and sufficient. Clearly, a high GOF is not a sufficient basis to establish the validity of interpretations based on the theory underlying the posited model. Therefore, for example, Hu and Bentler (1998) concluded that "although our discussion has focused on issues regarding overall fit indices, consideration of other aspects such as adequacy and interpretability of parameter estimates, model complexity, and many other issues remains critical in deciding on the validity of a model" (p. 450). Byrne (2001) similarly emphasized that fit indexes do not reflect the plausibility of a model and "this judgment rests squarely on the shoulders of the researcher" (p. 88). To make this clear, assume a population-generating model in which all measured variables were nearly uncorrelated. Almost any hypothesized model would be able to fit this data because most of the variance is in the measured variable uniqueness terms, and there is almost no covariation to explain. In a nonsensical sense, a priori models positing one, two, three, or more

factors would all be able to explain the data (as, indeed, would a null model with no factors) and produce a very good fit. The problem with the interpretation of this apparently good fit would be obvious in an inspection of the parameter estimates in which all factor loadings and factor correlations were close to zero. Using a less extreme example, if theory predicts that a path coefficient should be positive, whereas the observed results show that it is negative, high levels of GOF are not sufficient to argue for the validity of predictions based on the model. Without belaboring the point further, suffice it to say that high levels of GOF are not a sufficient basis for model evaluation.

A critical question is whether there are absolute criteria—golden rules or even recommended guidelines—of acceptable levels of fit that are a necessary basis for valid interpretations. In exploring this issue, we briefly digress to the somewhat analogous situation of evaluating reliability estimates, as GOF indexes were historically developed as a measure of the reliability of a model (Tucker & Lewis, 1973). There is no universal consensus about what is an acceptable level of reliability, but there are at least three different approaches to pursuing this issue: ambit claims, a criterion reference approach, and a normative reference approach. Although there are ambit suggestions based on intuition and accepted wisdom that reliability should be at least .70 or at least .80, there is general agreement that—all other things being equal—higher reliability is better. In an attempt to provide a more solid basis for establishing acceptable levels of reliability, Helmstadter (1964), for example, described criterion and normed reference approaches to the issue. Based on a criterion of being able to discriminate between scores differing by one fourth of a standard deviation with an 80% probability, acceptable levels of reliability were roughly .50 for the comparison of two group means; .90 for group mean scores based on the same group on two occasions; .94 for two different individuals; and .98 for two scores by the same individual. Using a normed reference approach, Helmstadter reported median reliabilities for tests in different content areas. Therefore, he concluded that reliability coefficients should be evaluated in relation to the purpose of a test, the content area, and the success of other instruments. Of course, it is also important to ensure that increased reliability is not achieved at the expense of construct validity. In summary, applied researchers seemed to have accepted there are no absolute guidelines for what constitutes necessary levels of reliability accepting imprecise rules of thumb that have limited generality rather than applying golden rules that provide absolute guidance for all situations.

A Normed-Reference Approach to Acceptable Levels of GOF

Analogous to the situation in reliability, rules of thumb about acceptable levels of GOF (e.g., incremental fit indexes > .9) have traditionally been ambit claims based

on intuition and accepted wisdom. Using a normed reference approach, Marsh et al. (in press) proposed the following "strawperson" claim[1]:

> Conventional CFA [confirmatory factor analysis] goodness of fit criteria are too restrictive when applied to most multifactor rating instruments. It is my experience that it is almost impossible to get an acceptable fit (e.g., CFI, RNI, TLI > .9; RMSEA < .05) for even "good" multifactor rating instruments when analyses are done at the item level and there are multiple factors (e.g., 5–10), each measured with a reasonable number of items (e.g., at least 5–10/per scale) so that there are at least 50 items overall. If this is the case, then I argue that "conventional" rules of thumb about acceptable fit are too restrictive (even though there has been a recent push for even stricter standards).

Bollen (1989) offered a related argument, suggesting that the evaluation of incremental fit indexes could depend, in part, on values reported in practice such that progress in relation to previous results might be important even if values did not meet some externally imposed standard.

Marsh (2001, SEMNET@UA1vm.ua.edu) placed this claim on SEMNET (an e-mail network on the discussion of various SEM issues) and invited the 1,500 members to provide counter examples. Although a number of interesting points were raised in response to this strawperson claim, no one offered a published counter example. Indeed, many of the large number of responses to Marsh's (2001, SEMNET@UA1vm.ua.edu) e-mail argued that apparent support for the claim merely demonstrated that currently available instruments are mostly unacceptable. The prevailing sentiment was an extreme reluctance to modify existing (new) standards simply because they were not readily obtainable in practice. A substantial tangent to this discussion questioned whether it was even reasonable or appropriate to expect a factor to be represented by more than two or three items (or perhaps more than a single item). Marsh et al. (in press) responded that it may be unreasonable to have more than two or three items per factor if researchers hope to achieve GOF indexes of .95, but that it is highly desirable if researchers want to have measures with good construct validity. As was the case when researchers resorted to counterproductive procedures (e.g., using smaller sample sizes) to achieve nonsignificant chi-square values that represented the accepted standard of a good fit, it appears that researchers are again resorting to dubious practice to achieve in-

[1]The claim by Marsh et al. (in press) was based on the incremental fit indexes computed with the minimum fit function chi-square, the usual ML chi-square (called C1 in the current version of 8.54 of LISREL), that had been the basis of most research at the time the claim was made. However, in the current version of LISREL, the default is the normal theory weighted least-squares chi-square (called C2). It appears that incremental fit indexes, as provided by the current version of LISREL, are systematically higher when based on the normal theory weighted least-squares chi-square—due in large part to substantial difference in the computation of the values for the null model. Hence, claims be Marsh et al. (in press) and, apparently, guidelines based on the usual ML chi-square (i.e., C1) no longer apply to indexes based on the normal theory weighted least-squares chi-square (i.e., C2).

appropriately high GOF standards. In summary, based on a normative reference approach to establishing cutoff values on the basis of GOF indexes achieved in current practice, there is some evidence to suggest that even the old cutoff values (e.g., RNI and TLI > .90) are overly demanding in relation to a normative criterion of appropriateness based on the best existing psychological instruments. Hence, the new, more demanding cutoff values proposed by Hu and Bentler (1998, 1999) appear to be largely unobtainable in appropriate practice.

A Criterion Reference Rationale for Cutoff Values Based on Hu and Bentler (1998, 1999)

There has been some systematic research attempting to validate GOF indexes from a criterion reference approach. Marsh and Balla (1994; Marsh, Balla, & Hau, 1996) constructed a nested series of alternative models that systematically varied in terms of misspecification (over and under parameterization) and evaluated GOF indexes in terms of how sensitive they were in reflecting misspecification. Marsh et al. (1996) also emphasized that most GOF indexes were better able to distinguish between the fit of alternative models of the same data than to establish criteria of acceptable fit. They did not, however, systematically evaluate different cutoff values. Hu and Bentler (1998, 1999) also argued that GOF testing has focused too much on the evaluation of true population models. They, however, evaluated the performance of cutoff values for GOF indexes in relation to an objective criterion of sensitivity to model misspecification.

HU AND BENTLER'S STUDIES ON NEW CUTOFF CRITERIA FOR FIT INDEXES

In addressing the cutoff problem, Hu and Bentler (1999) argued that "an adequate cutoff criterion should result in a minimum Type I error rate (i.e., the probability of rejecting the null hypothesis when it is true) and Type II error rate (the probability of accepting the null hypothesis when it is false)" (p. 5). In pursuing this issue, the seven best GOF indexes recommended in earlier research (Hu & Bentler, 1998, 1999) were evaluated in relation to traditional and new cutoff criteria. They considered two different population-generating CFA models—a simple structure (5 indicators for each of 3 correlated factors) and a complex structure that differed only in that there were a total of three additional cross loadings in addition to the 15 factor loadings in the simple structure. Hence, the simple population-generating structure had 33 nonzero parameter estimates, whereas the complex model had 36 nonzero parameter estimates. For each model type, one "true" and two underparameterized, misspecified models were tested. For the misspecified models evaluated in relation to the simple structure, one (Model M1) or two (Model M2) of the factor correlations were fixed to be zero. For the complex structure, the misspecified models had one (Model M1) or two

(Model M2) of the three nonzero cross loadings fixed to be zero. The design also included seven distributional conditions and six sample sizes.

In their more stringent cutoff rules based on the ML method, Hu and Bentler (1998) suggested .95 as a cutoff value for TLI, Bollen's (1989) Fit Index (BL89), CFI (or RNI) and Gamma Hat; .90 for McDonald's Centrality Index (Mc; McDonald & Marsh, 1990); .06 for RMSEA; and .08 for standardized root mean squared residual (SRMR; Hu & Bentler, 1999, also evaluated various combinations of these fit indexes that are not considered in our comment). There is, however, an apparent problem in their application of the traditional hypothesis-testing approach. To illustrate this problem, consider, for example, the results for RNI = .95 with the simple structure for small ($N \leq 250$), medium ($N = 500$), and large ($N \geq 1{,}000$) N values (Hu & Bentler, 1999, Table 2, p. 13). Traditionally, researchers have sought decision rules that maintained Type 1 error rates at acceptably small values that did not vary with N, so that Type 2 error rates systematically decreased with increasing N. However, Hu and Bentler's (1999) results did not satisfy these traditional patterns of behavior. Although one might want a decision rule to be reasonably consistent across N for acceptance of the true model, the true model was incorrectly rejected 25%, 3.6%, and 0.1% of the time at small, medium, and large N, respectively. Perhaps higher Type 1 error rates would be acceptable at small N to minimize Type 2 error rates, although this is inconsistent with traditional approaches to hypothesis testing (Hu & Bentler, 1999, emphasized the need for flexibility in establishing the most appropriate trade-off between Type 1 and Type 2 errors).

Clearly, however, it is desirable for misspecified, false models to be rejected more frequently when N increases; but the actual pattern of results reported by Hu and Bentler (1998, 1999) was exactly the opposite. The least misspecified model (M1) was correctly rejected 47.2%, 30.1%, and 4.7% of the time at small, medium, and large N, respectively. For the more misspecified model (M2), the corresponding rejection rates were 59.8%, 45.7%, and 19.4%, respectively. Hence, according to the recommended decision rule for RNI, the likelihood of correctly rejecting a false model was substantial for small N and very small for large N. This apparently paradoxical behavior is clearly undesirable for a decision rule. More important, this is not an isolated example. For all of the recommended cutoff values for each of the seven GOF indexes in Hu and Bentler's (1998, 1999) study, the probability of correctly rejecting a false model was lower for large N than small N for at least one of the models considered.

Apparently Paradoxical Behavior of Hu and Bentler Cutoff Criteria

Apparently, the reason why Hu and Bentler's (1998, 1999) decision rules behave as they do is that they constructed misspecified models that should have been judged as acceptable—even according to their new, more stringent criteria. To illustrate the nature of this problem, we first present a hypothetical example. Con-

sider the hypothesis-testing context analogous to that proposed by Hu and Bentler (1998, 1999) for RNI in relation to their proposed cutoff criterion of RNI = .95. In Figure 1, sample size and degree of misspecification are systematically varied. When misspecification is large (e.g., RNI < .95), the decision rules behave appropriately: For small *N,* (Figure 1A) there are moderate levels of false rejection of a true model (the area under the probability density function [PDF] for the true model for RNI < .95) and false acceptance of a misspecified model (the area under the PDF for the misspecified model for RNI > .95). For large *N,* the probability of both types of errors approach zero (Figure 1B). Had Hu and Bentler (1998, 1999) selected models with relatively large levels of misspecification, their decision rules would also have behaved appropriately; however, their recommended cutoff criteria would have been much less stringent.

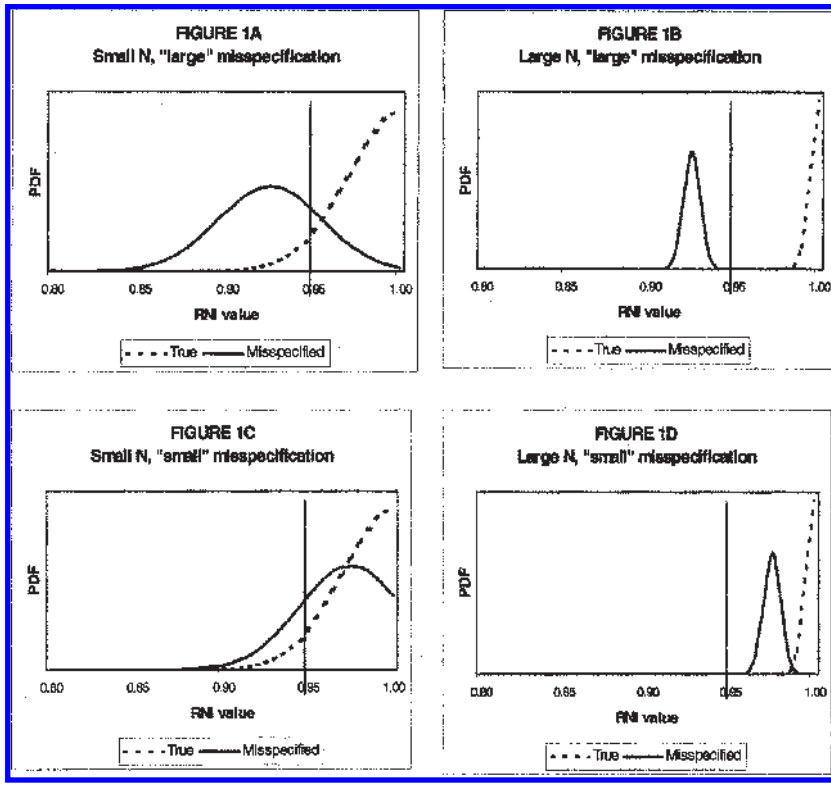

FIGURE 1 Hypothetical data demonstrating the behavior of decision rules based on RNI as a function of the extent of misspecification (true vs. misspecified) and sample size (large *N* vs. small *N*). RNI = Relative Noncentrality Index; PDF = probablility density function.

Now consider the situation when misspecification is small (e.g., RNI > .95) in Figure 1C and 1D. Here, the decision rules behave inappropriately, similar to the behavior observed in the Hu and Bentler (1998, 1999) study. For small *N* (Figure 1C), again there are moderate levels of false rejection of a true model (the area under the PDF for the true model for RNI < .95), but substantial levels of false acceptance of the misspecified model (the area under the PDF for the misspecified model for RNI > .95). For a sufficiently large *N* (Figure 1D), the levels of false rejection of a true model approach zero. In contrast, the false acceptance of the so-called misspecified model approaches 100% when *N* is sufficiently large.

As clearly depicted in Figure 1D, the resolution of this apparent paradox is that the population GOF value for the so-called misspecified model actually falls in the acceptance region—RNI > .95—so that, perhaps, this acceptable-misspecified model should be considered an acceptable model rather than a misspecified, false model. In general, for any acceptable-misspecified model—a misspecified model whose population GOF falls in the acceptance region—the probability of falsely accepting the model will approach 100% for a sufficiently large *N*. The logicial inconsistency in this situation is to define as misspecified a model that should be considered as acceptable according to even the highly stringent cutoff criteria proposed by Hu and Bentler (1998, 1999). In Figure 1C and 1D, we inappropriately attempted to discriminate between a true model and a misspecified model that should have been deemed acceptable. Considering the misspecified model to be acceptable, the behavior of the decision rules now makes sense. Hence, the probability of falsely rejecting an acceptable model is moderate for small *N* and approaches zero for a sufficiently large *N*.

This illustration also demonstrates that the conclusions about recommended cutoff values are highly dependent on the particular misspecified models that are used. It may be inappropriate to consider misspecified models that should really be considered acceptable and results in apparently inappropriately stringent cutoff values for GOF indexes (e.g., RNI = .95). In contrast, if Hu and Bentler (1998, 1999) chose models of sufficiently extreme levels of misspecification, according to their rationales and evaluation methods, even modest cutoff values (e.g., RNI =.80) would have been highly accurate in discriminating between true and misspecified models and would have led to less stringent criteria of acceptable fits.

Acceptable-Misspecified Models: How to Interpret a Misspecified Model That Is Acceptable

To what extent should Hu and Bentler's (1998, 1999) misspecified models have been considered acceptable according to their own standards of acceptability? We evaluated this question by replicating Hu and Bentler's (1998, 1999) simulation study. Here, we considered only ML estimation based on multivariate normal data (i.e., Hu & Bentler's, 1998, 1999, Condition 1) for the ML chi-square and selected

GOF indexes. As in their original design, for both the simple and complex models, one true model and two misspecified models were considered at sample sizes 150, 250, 500, 1,000, 2,500, and 5,000; each with 200 replicates. To estimate the population values for each of the seven GOF indexes, we fitted a model using all 500,000 cases that were used in our simulation, a very large *N*. In many cases, estimated population values for the so-called misspecified models indicated that these models should have been classified as acceptable. For these purposes, we refer to these as acceptable-misspecified models.

For the simple structure solution, the estimated population value for the least misspecified model (M1) was acceptable (i.e., less than the cutoff value) for six of seven indexes, all but SRMR. The estimated population value for the more misspecified model (M2) was also acceptable for five of seven indexes (all but SRMR and Mc). For the complex structure, M1 was acceptable for four of seven indexes (all but TLI, Mc, and RMSEA), whereas M2 was only acceptable according to the SRMR index. Hence, for a majority of the models considered by Hu and Bentler (1998, 1999), the so-called misspecified model actually provided an acceptable GOF in relation to Hu and Bentler's (1998, 1999) new, more stringent criteria. This heavy reliance on acceptable or misspecified models as a basis for establishing criteria of acceptability is a potentially important problem with the Hu and Bentler (1998, 1999) research that may undermine interpretations of their results in relation to acceptable levels of GOF, and it certainly limits the generalizability of their results.

It is interesting to note that the fits for misspecified models M1 and M2 were systematically better for the simple structure than the complex structure for all of the fit indexes (and for chi-square values) except SRMR. In contrast, for the SRMR, the fit was much worse for the misspecified simple model than the complex model. In fact, the misspecified simple structure models were mostly acceptable according to population estimates for all of the fit indexes except for SRMR. The situation was nearly reversed for the complex structure. For the complex structure, the misspecified models were either not acceptable or borderline acceptable for all fit indexes other than SRMR, but clearly acceptable according to population estimates for the SRMR. This complicated interaction between acceptability or unacceptability of the misspecified models, the data structure (simple vs. complex), the particular index, and sample size demonstrates that broad generalizations based on these data may be unwarranted.

Behavior of Decision Rules for Acceptable-Misspecified Models

Of particular interest is the behavior for decision rules based on misspecified models that are acceptable according to Hu and Bentler's (1998, 1999) recommended decision rules (i.e., acceptable-misspecified models). To illustrate this problem with acceptable-misspecified models, we consider results based on RNI and SRMR in detail (but present values for all indexes in Table 1).

RNI. For the simple structure (see Table 1A), the estimated population values for both misspecified models M1 and M2 are acceptable (greater than the recommended cutoff of .95). For small *N* (150 and 250), there is sufficient sampling fluctuation so that a moderate number of these acceptable-misspecified models are rejected (Table 1A). However, as sample size increased and sampling fluctuation decreased, the rejection rate for these acceptable-misspecified models became increasingly small and was zero for large *N* (1,000; 2,500; 5,000). Hence, this pattern of actual results was similar to the hypothetical results in Figure 1 for small misspecification.

For the complex structure, the estimated population estimate of RNI (.951) for misspecified Model M1 was nearly equal to the recommended cutoff value. Hence, particularly for the smaller sample sizes, the rejection rates for M1 were close to 50% (see Table 1A; also see Figure 2). For large sample sizes, however, the sampling fluctuations were so small that even the difference between the estimated population value of .951 and the cutoff value of .950 led to rapidly declining rejection rates (49% for *N* = 1,000; 32.5% for *N* = 2,500; 25% for *N* = 5,000) for this marginally acceptable or misspecified model.

For the complex structure, the misspecified Model M2 had an estimated population RNI (.910) that was clearly worse than the recommended cutoff value of .95.

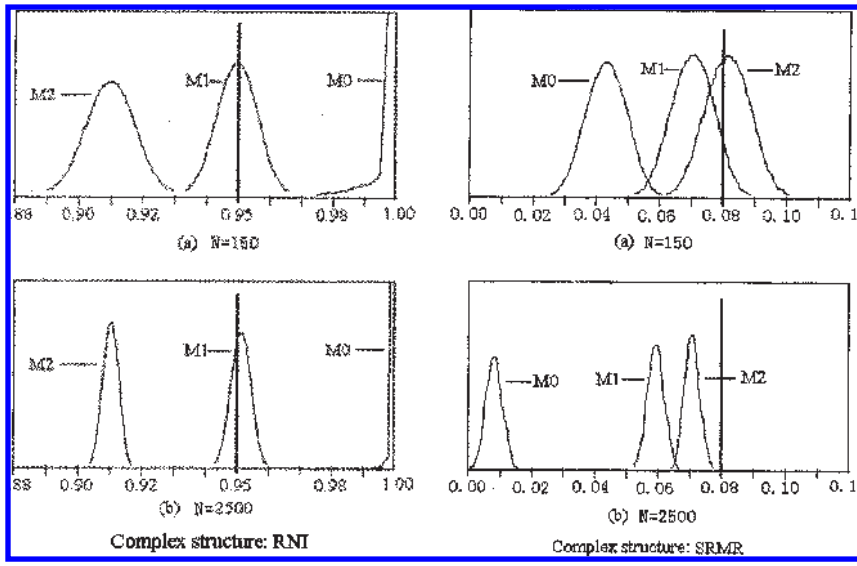

FIGURE 2 Distributions of RNI (left side) and SRMR (right side) for complex structure when (a) *N* = 150 and (b) *N* = 2,500. M0 = the true model; M1 = less misspecified model; M2 = more misspecified model; RNI = Relative Noncentrality Index; SRMR = standardized root mean squared residual.

TABLE 1A
Behavior of Goodness of Fit Indexes for the True Model (M0) and Two Misspecified Models (M1, M2) for the Simple Structure at Different Sample Sizes

		N = 150			*N = 250*			*N = 500*			*N = 1,000*			*N = 2,500*			*N = 5,000*		
GOF (cutoff value)	*Est. Pop.*	*M*	*SD*	*% Reject*	*M*	*SD*	*% Reject*	*M*	*SD*	*% Reject*	*M*	*SD*	*% Reject*	*M*	*SD*	*% Reject*	*M*	*SD*	*% Reject*
TLI (0.95)																			
M0	1.000	0.996	0.018	1.0	0.997	0.010	0.0	0.999	0.005	0.0	1.000	0.002	0.0	1.000	0.001	0.0	1.000	0.000	0.0
M1	0.963	0.960	0.021	28.0	0.962	0.013	15.5	0.963	0.008	4.5	0.963	0.005	0.0	0.963	0.003	0.0	0.963	0.002	0.0
M2	0.951	0.947	0.020	50.0	0.949	0.013	52.5	0.950	0.008	48.5	0.951	0.005	47.5	0.951	0.003	40.0	0.951	0.002	33.5
BL89 (0.95)																			
M0	1.000	0.996	0.014	0.0	0.998	0.008	0.0	0.999	0.004	0.0	1.000	0.002	0.0	1.000	0.001	0.0	1.000	0.000	0.0
M1	0.969	0.967	0.017	17.5	0.968	0.010	5.0	0.969	0.006	0.0	0.969	0.004	0.0	0.969	0.002	0.0	0.969	0.002	0.0
M2	0.958	0.956	0.017	34.0	0.957	0.011	22.5	0.958	0.007	13.0	0.958	0.004	2.0	0.958	0.003	0.0	0.958	0.002	0.0
RNI (0.95)																			
M0	1.000	0.992	0.011	0.0	0.996	0.006	0.0	0.998	0.003	0.0	0.999	0.001	0.0	1.000	0.001	0.0	1.000	0.000	0.0
M1	0.969	0.966	0.017	18.0	0.968	0.011	5.0	0.969	0.006	0.0	0.969	0.004	0.0	0.969	0.002	0.0	0.969	0.002	0.0
M2	0.958	0.955	0.017	36.0	0.957	0.011	23.5	0.958	0.007	15.0	0.958	0.004	2.0	0.958	0.003	0.0	0.958	0.002	0.0
G_ Hat (0.95)																			
M0	1.000	0.997	0.013	0.0	0.998	0.007	0.0	0.999	0.004	0.0	1.000	0.002	0.0	1.000	0.001	0.0	1.000	0.000	0.0
M1	0.973	0.970	0.016	9.5	0.972	0.010	1.5	0.972	0.006	0.0	0.973	0.004	0.0	0.973	0.002	0.0	0.973	0.002	0.0
M2	0.963	0.961	0.015	24.5	0.962	0.010	10.5	0.963	0.007	3.5	0.963	0.004	0.0	0.963	0.002	0.0	0.963	0.002	0.0

Mc (0.90)																			
M0	1.000	0.988	0.05	6.0	0.993	0.027	0.0	0.996	0.014	0.0	0.999	0.007	0.0	1.000	0.003	0.0	1.000	0.001	0.0
M1	0.900	0.892	0.055	50.0	0.896	0.034	53.5	0.898	0.021	56.0	0.900	0.013	54.5	0.900	0.008	49.5	0.900	0.005	48.5
M2	0.867	0.858	0.053	78.5	0.863	0.034	87.5	0.865	0.023	93.5	0.867	0.014	99.0	0.867	0.009	100.0	0.867	0.006	100.0
SRMR (0.08)																			
M0	0.001	0.047	0.005	0.0	0.037	0.004	0.0	0.026	0.003	0.0	0.019	0.002	0.0	0.012	0.001	0.0	0.008	0.001	0.0
M1	0.135	0.144	0.020	100.0	0.140	0.016	100.0	0.137	0.013	100.0	0.136	0.009	100.0	0.136	0.005	100.0	0.136	0.004	100.0
M2	0.165	0.173	0.020	100.0	0.168	0.017	100.0	0.166	0.014	100.0	0.165	0.009	100.0	0.165	0.006	100.0	0.165	0.004	100.0
RMSEA (0.06)																			
M0	0.001	0.012	0.017	0.5	0.011	0.013	0.0	0.009	0.009	0.0	0.006	0.006	0.0	0.003	0.004	0.0	0.002	0.003	0.0
M1	0.046	0.042	0.015	10.5	0.045	0.009	4.5	0.046	0.005	0.0	0.046	0.003	0.0	0.046	0.002	0.0	0.046	0.001	0.0
M2	0.053	0.049	0.013	19.0	0.051	0.008	9.0	0.052	0.005	3.5	0.052	0.003	0.0	0.053	0.002	0.0	0.053	0.001	0.0
χ^2/df ($\alpha = .05$)																			
M0		1.045	0.175	14.5	1.045	0.156	9.0	1.043	0.159	8.0	1.030	0.151	11.5	1.006	0.156	6.5	1.003	0.155	5.0
M1		1.395	0.212	73.5	1.623	0.217	96.0	2.219	0.266	100.0	3.393	0.336	100.0	6.973	0.501	100.0	12.942	0.687	100.0
M2		1.518	0.208	92.0	1.829	0.224	100.0	2.624	0.299	100.0	4.206	0.366	100.0	9.019	0.559	100.0	17.017	0.760	100.0

Note. Est. Pop. = estimated goodness of fit (GOF) using 500,000 cases that were used in replicated simulation; M0 = the true model (simple structure); M1 = misspecified model 1 (PH21 = 0); M2 = misspecified model 2 (PH21 = PH31 = 0); % Reject = % of rejection rate; TLI = Tucker–Lewis Index (1973); BL89 = Bollen Fit Index (1989); RNI = Relative Noncentrality Index; G_Hat = Gamma hat; Mc = McDonald Centrality Index (McDonald & Marsh, 1990); SRMR = standardized root mean squared residual. RMSEA = root mean squared error of approximation. χ^2/df ($\alpha = .05$) = ML χ^2/df ratio and rejection rates based on the traditional $\alpha = .05$ test of statistical significance.

Therefore, for all but the smallest sample size, the rejection rate for this unacceptable-misspecified model was close to 100%. Hence, the pattern of results for this unacceptable-misspecified model was similar to the hypothetical results in Figure 1 for the large misspecification.

SRMR. Now consider the results based on the SRMR (Table 1). Because SRMR is somewhat biased by sample size, the estimated value of particularly the true model was negatively related to *N*. For the simple structure misspecified models (M1 and M2 in Table 1A), the SRMRs were substantial and unacceptable according to the recommended cutoff values for SRMR. These unacceptable-misspecified models (M1 and M2) were correctly rejected 100% of the time. Hence, this pattern of actual results based on this highly unacceptable-misspecified model was most analogous to the hypothetical results in Figure 1 for the large misspecification (keeping in mind that higher values of RNI and lower values of SRMR reflect better fits so that SRMR is actually a measure of badness of fit rather than GOF).

The behavior of the SRMR for the complex structure was quite different (see Figure 2). The misspecified models (M1 and M2) were both acceptable in that their estimated population values were better (smaller) than the recommended cutoff value of .08. Therefore, for large *N,* the rejection rates for both the acceptable or misspecified models (M1 and M2) were 0%. However, the situation was more complicated for small *N,* because the estimated population value for SRMR was biased by *N*. Therefore, the mean value for M1 varied from .072 to .057 (for *N*s that vary from 150 to 5,000), although all of these values were still acceptable (less than .08). For small *N,* there was sufficient sampling fluctuation so that the acceptable-misspecified M1 was sometimes correctly rejected (rejection rate = 19% for $N = 150$). For large *N,* the distribution was less variable so that the misspecified M1 was never rejected. Hence, this pattern of actual results based on this acceptable-misspecified model is most analogous to the hypothetical results in Figure 1 for the small misspecification.

For M2, however, the mean values of SRMR varied from .083 to .070 for different *N*s. Therefore, the values for solutions based on small *N*s were unacceptable (greater than .08), whereas those based on large *N* were acceptable. Consequently, the rejection rates were large for small *N* (62% for $N = 150$), dropped off quickly, and were zero for large *N* ($N \geq 1{,}000$). Because the acceptability of the M2 based on SRMR shifts depending on the sample size, this pattern of results is not analogous to any of those presented in Figure 1.

Summary. In summary, whereas the behavior of decision rules based on the SRMR is logical in relation to a detailed evaluation of the simulation results, the pattern of results is complicated and apparently different from the RNI. The most dramatic difference between the two is that RNIs were more acceptable for

misspecified models based on the simple structure, whereas the SRMRs were more acceptable for misspecified models based on the complex structure. Interpretations were also complicated in that RNI was relatively unbiased by *N,* whereas there was a clear sample size bias in SRMR values. Nevertheless, there was an important consistency in the two sets of results that is the central feature of our comment illustrated in Figure 1. For unacceptable-misspecified models, the behavior of decision rules was reasonable in that rejection rates increased systematically with increasing *N* and were 100% for sufficiently large *N.* However, for acceptable-misspecified models, the behavior of decision rules was apparently inappropriate for a misspecified model in that rejections decreased systematically with increasing *N* and were 0% for sufficiently large *N.*

Comparison of Hu and Bentler Decision Rules With Those Based on the ML Chi-Square Test Statistic

The specific paradigm proposed by Hu and Bentler (1999) was a traditional hypothesis-testing scenario in which values of the best GOF indexes were used to accept or reject models known a priori to be true or false in relation to the population-generating model. Most GOF indexes considered by Hu and Bentler (1998, 1999) were specifically designed to reflect approximation discrepancy at the population level and to be relatively independent of sample size. However, Hu and Bentler's (1998, 1999) approach to evaluating these indexes used a hypothesis-testing approach based substantially on estimation discrepancy that was primarily a function of sampling fluctuations. Hence, there was an apparent inconsistency between the intended purposes of the GOF indexes and the approach used to establish cutoff values. To illustrate this point, we have replicated and extended their simulation to include traditional ML chi-square tests (Table 1). Briefly, these results show the following:

1. Mean values are relatively stable across *N* for all GOF indexes (only SRMR showed a moderate sample size bias), but the ML chi-square:degree of freedom ratio for misspecified models increased dramatically with increasing *N.*
2. Standard deviations systematically decreased with increasing *N* for all GOF indexes.
3. Type 1 errors (i.e., the probability of rejecting the null hypothesis when it is true) decreased with increasing *N* for all GOF indexes, but were relatively stable for the ML chi-square test ($\alpha = .05$).
4. Type 2 errors (i.e., the probability of accepting misspecified models) increased with increasing *N* for all GOF indexes when the population value was better than the cutoff criterion (i.e., acceptable-misspecified models), but decreased with increasing *N* for all GOF indexes when the population

value was worse than cutoff (i.e., unacceptable-misspecified model) and for ML chi-square tests ($\alpha = .05$).

5. ML chi-square tests ($\alpha = .05$) consistently outperformed all of the seven GOF indexes in terms of correctly rejecting misspecified models.

In summary, these results show that the ML chi-square test ($\alpha = .05$) did very well at least in the normal situation, and generally outperformed all of the seven GOF indexes in relation to rejecting misspecified models; one of the main criteria proposed by Hu and Bentler (1998, 1999). This implies that either we should discard all GOF indexes and focus on chi-square test statistics or that the hypothesis-testing paradigm used to evaluate GOF indexes was inappropriate. In fact, the reason why the chi-square test statistic is not a good GOF index is precisely the reason why it does perform so well in the hypothesis-testing situation. The expected value of the ML chi-square varies directly with *N* for misspecified models. Although this may be undesirable in terms of evaluating approximation discrepancy, it is precisely the type of behavior that is appropriate for optimization of decision rules in a traditional hypothesis-testing situation. Whereas our actual results are based on multivariate normal data that may not generalize to other situations, appropriate test statistics will generally be more effective in a traditional hypothesis-testing situation than GOF indexes.

Generalizability of Hu and Bentler's Decision Rules

The same cutoff value of .95 was proposed by Hu and Bentler (1999) for all four incremental fit indexes that they recommended. However, extensive evaluation of these indexes (e.g., Marsh et al., 1988; Marsh et al., in press), particularly the comparison of TLI (that penalizes for model complexity) and the RNI (that does not penalize for model complexity), indicates that these indexes do not vary along the same underlying 0 to 1 continuum. For the TLI cutoff value of .95, both simple misspecified models were acceptable (population estimate of TLI > .95, Table 1A), whereas both complex models were unacceptable (population estimate of TLI < .95, Table 1B). For the RNI cutoff value of .95, only the most misspecified complex model was unacceptable. For each of the five cutoff values evaluated by Hu and Bentler (1998), the percentage rejection for misspecified models was consistently higher for TLI than RNI. Therefore, for example, for large $N = 5{,}000$, misspecified Model M1 was correctly rejected 100% of the time with TLI but only 25% of the time for RNI. Indeed, for all specified and misspecified models, TLI = .95 led to a higher percentage of rejections (larger Type 1 errors and lower Type 2 errors) than did BL89 = .95, CFI (or RNI) = .95, or Gamma Hat = .95. These results were consistent with other findings (see Marsh et al., in press) indicating that misspecified models typically had systematically lower TLIs than RNIs (or CFIs).

The comparison of the behavior of each of the seven indexes was quantitatively different in relation to the percentage of rejection of true and misspecified models, whereas the SRMR was qualitatively different from the others. In particular, for six of the seven indexes, the two complex misspecified models led to higher rejection percentages than the two simple misspecified models. In contrast, the SRMR led to higher rejections of the simple misspecified model than the complex misspecified model. For example, for large *N,* SRMR led to 100% rejection of both simple misspecified models and 0% rejection of the both complex misspecified models, whereas TLI resulted in 100% rejection for both complex misspecified models and 0% and 33.5% rejection of the simple misspecified models. As concluded by Hu and Bentler (1999), this implies that SRMR is clearly sensitive to different sorts of misspecification than the other GOF indexes.

Although Hu and Bentler (1999) evaluated different types of misspecification, their data were simulated such that most parameters were known to be equal to zero in the population-generating model, and all estimated parameters perfectly reflected the population-generating model with the exclusion of a few key parameter estimates that led to the models being misspecified. Although relevant for purposes of evaluating the behavior of GOF indexes in relation to varying degrees of misspecification, these misspecified models were clearly not representative of typical application in which true population values are rarely, if ever, exactly equal to an a priori value (e.g., factor loadings of zero). Indeed, many of their misspecified models were not sufficiently misspecified to be rejected by even their more stringent guidelines of acceptable levels of fit (also see earlier discussion of conclusions by Marsh et al., in press). Recognizing this problem, Hu and Bentler (1998) noted that the rationale for the selection of misspecified models was a limitation to their study. Despite their warning, however, current practice seems to have incorporated their new recommended guidelines of acceptable fit without the appropriate cautions, although these guidelines were based on simulation studies in which only a very limited range of misspecification was considered. Apparently consistent with limitations recognized by Hu and Bentler (1998, 1999), it seems that much more research with a wide variety of different types of misspecified models and under different conditions is needed before we adequately understand the behavior of rules of thumb or decision rules based on existing fit indexes.

Hu and Bentler (1998) specifically noted that they were only concerned about the evaluation of the GOF of a single model in isolation, not the comparative fit of alternative, nested models. However, on average, all their GOF indexes were apparently very good at distinguishing between the more and less misspecified models within each structure (simple and complex) type. Whereas there might be minor differences in the performances of different indexes at small *N* and under different distributional conditions, the indexes were apparently better at distinguishing between degrees of misspecification than providing absolute guidelines about the acceptability of a particular model or whether the extent of difference in

TABLE 1B
Behavior of Goodness-of-Fit Indexes of the True Model (M0) and Two Misspecified Models (M1, M2) for the Complex Structure at Different Sample Sizes

		N = 150			*N = 250*			*N = 500*			*N = 1,000*			*N = 2,500*			*N = 5,000*		
GOF (cutoff value)	*Est. Pop.*	*M*	*SD*	*% Reject*	*M*	*SD*	*% Reject*	*M*	*SD*	*% Reject*	*M*	*SD*	*% Reject*	*M*	*SD*	*% Reject*	*M*	*SD*	*% Reject*
TLI (0.95)																			
M0	1.000	0.996	0.014	0.0	0.998	0.008	0.0	0.999	0.004	0.0	1.000	0.002	0.0	1.000	0.001	0.0	1.000	0.000	0.0
M1	0.940	0.937	0.022	71.5	0.938	0.014	83.0	0.938	0.009	91.0	0.939	0.006	98.5	0.940	0.004	100.0	0.940	0.003	100.0
M2	0.890	0.889	0.025	99.5	0.887	0.017	100.0	0.888	0.012	100.0	0.889	0.008	100.0	0.890	0.005	100.0	0.890	0.003	100.0
BL89 (0.95)																			
M0	1.000	0.997	0.011	0.0	0.998	0.007	0.0	0.999	0.003	0.0	1.000	0.002	0.0	1.000	0.001	0.0	1.000	0.000	0.0
M1	0.951	0.950	0.017	49.5	0.950	0.012	48.0	0.950	0.007	47.5	0.951	0.005	39.0	0.951	0.003	32.0	0.951	0.002	25.0
M2	0.910	0.910	0.020	97.5	0.908	0.014	100.0	0.909	0.010	100.0	0.909	0.007	100.0	0.910	0.004	100.0	0.910	0.003	100.0
RNI (0.95)																			
M0	1.000	0.994	0.008	0.0	0.996	0.005	0.0	0.998	0.002	0.0	0.999	0.001	0.0	1.000	0.000	0.0	1.000	0.000	0.0
M1	0.951	0.949	0.017	52.5	0.949	0.012	50.0	0.950	0.007	49.0	0.951	0.005	39.0	0.951	0.003	32.5	0.951	0.002	25.0
M2	0.910	0.909	0.021	98.0	0.908	0.014	100.0	0.908	0.010	100.0	0.909	0.007	100.0	0.910	0.004	100.0	0.910	0.003	100.0
G_Hat (0.95)																			
M0	1.000	0.997	0.013	0.0	0.998	0.007	0.0	0.999	0.004	0.0	1.000	0.002	0.0	1.000	0.001	0.0	1.000	0.000	0.0
M1	0.950	0.946	0.018	57.5	0.946	0.012	64.0	0.947	0.007	68.0	0.948	0.005	67.0	0.948	0.003	73.5	0.948	0.002	79.5
M2	0.910	0.907	0.019	99.0	0.906	0.014	100.0	0.907	0.009	100.0	0.907	0.006	100.0	0.908	0.004	100.0	0.908	0.003	100.0

Mc (0.90)																			
M0	1.000	0.988	0.047	5.0	0.993	0.028	0.0	0.997	0.014	0.0	0.999	0.007	0.0	1.000	0.003	0.0	1.000	0.001	0.0
M1	0.810	0.808	0.060	94.0	0.809	0.041	99.0	0.810	0.025	100.0	0.813	0.017	100.0	0.814	0.010	100.0	0.815	0.008	100.0
M2	0.680	0.682	0.060	100.0	0.679	0.042	100.0	0.680	0.027	100.0	0.682	0.020	100.0	0.684	0.012	100.0	0.684	0.008	100.0
SRMR (0.08)																			
M0	0.001	0.043	0.005	0.0	0.034	0.005	0.0	0.024	0.003	0.0	0.017	0.002	0.0	0.011	0.001	0.0	0.007	0.001	0.0
M1	0.057	0.072	0.010	19.0	0.067	0.008	5.5	0.062	0.005	0.0	0.060	0.004	0.0	0.058	0.003	0.0	0.058	0.002	0.0
M2	0.070	0.083	0.009	62.0	0.079	0.007	41.0	0.075	0.005	14.5	0.072	0.003	0.0	0.071	0.002	0.0	0.070	0.002	0.0
RMSEA (0.06)																			
M0	0.000	0.012	0.016	0.0	0.011	0.013	0.0	0.008	0.009	0.0	0.006	0.006	0.0	0.003	0.004	0.0	0.002	0.003	0.0
M1	0.068	0.065	0.012	66.5	0.067	0.008	82.5	0.068	0.005	95.5	0.067	0.003	98.5	0.067	0.002	100.0	0.067	0.001	100.0
M2	0.092	0.088	0.010	100.0	0.090	0.007	100.0	0.091	0.004	100.0	0.091	0.003	100.0	0.091	0.002	100.0	0.091	0.001	100.0
χ^2/df ($\alpha = .05$)																			
M0		1.047	0.171	11.0	1.045	0.166	10.0	1.040	0.167	9.5	1.028	0.155	8.0	1.004	0.155	5.5	1.000	0.160	4.5
M1		1.758	0.265	98.5	2.250	0.299	100.0	3.473	0.364	100.0	5.872	0.492	100.0	13.085	0.727	100.0	25.086	1.098	100.0
M2		2.341	0.306	100.0	3.255	0.361	100.0	5.487	0.459	100.0	9.898	0.666	100.0	23.104	0.991	100.0	45.080	1.438	100.0

Note. Est. Pop. = estimated GOF using 500,000 cases that were used in replicated simulation. M0 = the true model (complex structure); M1 = misspecified model 1 (LX13 = 0); M2 = misspecified model 2 (LX13 = LX42 = 0); % Reject = % of rejection rate; TLI = Tucker–Lewis Index (1973); BL89 = Bollen Fit Index (1989); RNI = Relative Noncentrality Index; G_Hat = Gamma hat; Mc = McDonald Centrality Index (McDonald & Marsh, 1990); SRMR = standardized root mean squared residual. RMSEA = root mean squared error of approximation. χ^2/df ($\alpha = .05$) = ML χ^2/df ratio and rejection rates based on the traditional $\alpha = .05$ test of statistical significance.

misspecification between two models was substantively meaningful. Although disappointing to researchers in search of a golden rule, interpretations of the degree of misspecification should ultimately have to be evaluated in relation to substantive and theoretical issues that are likely to be idiosyncratic to a particular study. This consideration is often facilitated by comparing the performances of alternative, competing models of the same data.

SUMMARY

There are important logical problems underlying the rationale of a hypothesis-testing approach to setting cutoff values for fit indexes. The intent of the GOF indexes has been to provide an alternative to traditional hypothesis-testing approaches based on traditional test statistics (e.g., ML chi-square). However, Hu and Bentler (1998, 1999) specifically evaluated the GOF indexes in relation to a traditional hypothesis-testing paradigm in which the ML chi-square test outperformed all of the GOF indexes. More important, many of the misspecified models considered by Hu and Bentler (1998, 1999) provided a sufficiently good fit in relation to population approximation discrepancy that they should have been classified as acceptable models even according to the more stringent cutoff values proposed by Hu and Bentler (1998, 1999). Hence, rejection of these acceptable models, perhaps, should have constituted a Type 1 error (incorrectly rejecting an acceptable model), further complicating the interpretations of their results. This led to apparently paradoxical behavior patterns in their decision rules. In particular, for most indexes, the probability of correctly rejecting misspecified models systematically decreased with increasing *N,* whereas an appropriate decision rule should result in higher rejection rates for misspecified models with increasing *N.* Particularly because of this heavy reliance on acceptable-misspecified models, the results by Hu and Bentler (1998, 1999) may have limited generalizability to the levels of misspecification experienced in typical practice. Hence, we strongly encourage researchers, textbook authors, reviewers, and journal editors not to overgeneralize the Hu and Bentler (1998, 1999) results, transforming heuristic findings based on a very limited sample of misspecified models into golden rules of fit that are broadly applied without the cautions recommended by Hu and Bentler (1999). In contrast to decisions based on comparisons with a priori cutoff values, all of the indexes seemed to be more effective at identifying differences in misspecification based on a comparison of nested models (see also Footnote 1).

ACKNOWLEDGMENTS

The research was funded in part by grants from the Australian Research Council and the Chinese National Educational Planning Project (DBA010169).

Professors Hau and Wen were visiting scholars at the SELF Research Centre, University of Western Sydney while working on this study.

REFERENCES

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

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

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.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications and programming.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Helmstadter, G. (1964). *Principles of psychological measurement.* New York: Appleton-Century-Crofts.

Hemphill J. F., & Howell, A. J. (2000). Adolescent offenders and stages of change. *Psychological Assessement, 12,* 371–381.

Hu, L., & Bentler, P. M. (1997). Selecting cutoff criteria for fit indexes for model evaluation: Conventional versus new alternatives (Tech. Rep.). Santa Cruz: University of California.

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

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

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Chicago: Scientific Software International.

King, D. W., King, L. A., Erickson, D. J., Huang, M. T., Sharkansky, E. J., & Wolfe, J. (2000). Posttraumatic stress disorder and retrospectively reported stressor exposure: A longitudinal prediction model. *Journal of Abnormal Psychology, 109,* 624–633.

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.

Marsh, H. W., & Balla, J. R. (1994). Goodness-of-fit indices in confirmatory factor analysis: The effect of sample size and model complexity. *Quality & Quantity, 28,* 185–217.

Marsh, H. W., Balla, J. R., & Hau, K. T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical processes. In G. A. Marcoulides & R. E. Schumacker (Ed.), *Advanced structural equation modeling techniques* (pp. 315–353). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Marsh, H. W., Hau, K. T., & Grayson, D. (in press). Goodness of fit evaluation in structural equation modeling. In A. Maydeu-Olivares & J. McArdle (Eds.), *Psychometrics. A festschrift to Roderick P. McDonald.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness-of-fit. *Psychological Bulletin, 107,* 247–255.

Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in *Personality and Social Psychology Bulletin. Personality and Social Psychology Bulletin, 28,* 1629–1646.

Steiger, J. H., & Lind, J. M. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10.