Available online at www.sciencedirect.com
ScienceDirect
Personality and Individual Differences 42 (2007) 815–824
PERSONALITY AND INDIVIDUAL DIFFERENCES
www.elsevier.com/locate/paid

Structural equation modelling: Adjudging model fit

Paul Barrett *

University of Auckland, Management and Employment Relations Department, Commerce Building C, 18 Symonds Street, Auckland, New Zealand

Available online 7 November 2006

Abstract

For journal editors, reviewers, and readers of research articles, structural equation model (SEM) fit has recently become a confusing and contentious area of evaluative methodology. Proponents of two kinds of approaches to model fit can be identified: those who adhere strictly to the result from a null hypothesis significance test, and those who ignore this and instead index model fit as an approximation function. Both have principled reasons for their respective course of action. This paper argues that the chi-square exact-fit test is the only substantive test of fit for SEM, but, its sensitivity to discrepancies from expected values at increasing sample sizes can be highly problematic if those discrepancies are considered trivial from an explanatory-theory perspective. On the other hand, suitably scaled indices of approximate fit do not possess this sensitivity to sample size, but neither are they ''tests'' of model fit. The proposed solution to this dilemma is to consider the substantive ''consequences'' of accepting one explanatory model over another in terms of the predictive accuracy of *theory-relevant-criteria*. If there are none to be evaluated, then it is proposed that no scientifically worthwhile distinction between ''competing'' models can thus be made, which of course begs the question as to why such a SEM application was undertaken in the first place.

Keywords: SEM; Structural equation modelling; Model fit; Model evaluation

* Corresponding author. Tel.: +64 9 373 7599x82143.
E-mail address: paul.barrett@auckland.ac.nz

0191-8869/$ - see front matter
doi:10.1016/j.paid.2006.09.018

1. Structural equation model fit

When modelling data using structural equation modelling (SEM), one or more models may be fit to the same covariance matrix. Statistically appraising the fit of a model to the covariance matrix is accomplished using a "goodness of fit" test referenced against the χ^2 distribution, which takes as its argument, the discrepancy between the model-implied population covariances and the actual observed sample covariances. Given the degrees of freedom associated with a particular model, model fit is a matter of testing whether the discrepancies (or residuals) are greater than would be expected by chance alone. Put another way, the χ^2 test is a simultaneous test that all residuals (calculated by taking the difference between all model implied covariances and the observed sample covariances) are zero. Bollen (1989), provides the actual fit equations.

2. The χ^2 test

This is a conventional null hypothesis significance test (NHST) for the goodness of fit test, albeit with the "hoped for" decision reversed so that the aim is now to "accept" the null hypothesis, and not reject it. If the discrepancy (expressed as a χ^2 variate) between the model implied covariances and the observed sample covariances is larger than the expected distribution value by a probability usually adjudged at a 0.05 threshold (as per convention in NHST), then the model is rejected as "not-fitting". Conversely, if the fit statistic is less than the value expected, with a probability of occurrence >0.05, then the model is accepted as "fitting"; that is, the null hypothesis of "no difference" between the model-implied population covariances and the actual observed sample covariances is not rejected. This test has become known amongst SEM users as the χ^2 "exact-fit" test. In line with all NHST procedures, as the number of sample cases increases, so does the sensitivity of the test increase such that at very large sample size, tiny discrepancies between the observed test statistic and its expected value under a null hypothesis are likely to be adjudged as evidence of "misfit". Mathematically, this occurs as the sample size is a "multiplier" of the discrepancy function in the exact-fit test. In general, the larger the sample size, the more likely a model will fail to fit via using the χ^2 goodness of fit test.

3. Approximate fit tests

Given the sample size issue referred to above, investigators turned to and/or developed heuristic indices which variously adjusted the χ^2 test statistic for the size of sample, number of variables, or degrees of freedom, as a way of trying to specify the graduated approximate fit of a model to data. That is, the exact fit test might fail, but these new indices would indicate the degree to which a model might be "discrepant", rather than a binary NHST fit/no-fit decision. One problem with this approach was that by definition, these indices would yield values akin to correlation coefficients or distance metrics, which required some interpretation as to whether a model was an "acceptable" or "unacceptable" approximate fit. Until 2000 or so, Bentler and his colleagues and others had published several significant simulated dataset "fit-indicator-behaviour" papers which were taken by the majority of SEM investigators to indicate "Golden rules" which could

be applied to many of these fit indices so as to permit the use of a threshold-value for an approximate fit index to indicate "acceptable fit". The most recent paper by Hu et al. (1999) has essentially become the "bible" for the threshold cutoffs by most SEM investigators.

4. New evidence concerning the appropriate size of approximate fit indices and model fit

However, four recent papers have cast doubt upon the continued utility of using indicative thresholds for approximate fit indices, essentially removing the notion that a single *threshold-value* can be applied to any particular approximate fit index under all measurement and data conditions (Beauducel & Wittmann (2005); Fan & Sivo (2005); Marsh, Hau, & Wen (2004); Yuan (2005)). Each paper demonstrated empirically that, under varying data conditions using known a priori model structures, single-valued indicative thresholds for approximate fit indices were impossible to set without some models being incorrectly identified as fitting "acceptably" when in fact they were misspecified to some degree. Indeed, the main theme running through these papers was that fixed thresholds for approximate fit indices were simply not plausible. Indeed Marsh et al. (2004) went further by noting that whereas approximate fit indices were originally created to specify the degree of fit to data, their current use had evolved into a kind of hypothesis test whereby each index was assigned a threshold value and exactly the same binary decision as might be made with the χ^2 test (fit/no-fit) was now being made with these indices (acceptable approximate fit/unacceptable approximate fit). Some proponents of these indices also suggested using confidence intervals around them, copying the logic of formal statistical inference.

5. What happened to the logic of model testing?

Let us step back for a moment and consider what is happening here. Joreskog developed the structural equations modelling technology, which allows models to be fit to covariance data. As in any area of science which fits models to data, some systematic method or methods of determining which model best fits the data was required. Joreskog showed that model discrepancy (i.e. fit) could be adjudged using a conventional null-hypothesis goodness of fit χ^2 significance test. He also noted that this test, like all statistical tests, would become increasingly sensitive to tiny model-fit discrepancies as the sample size increased. So, he developed the notion of an approximate fit index, like the goodness-of-fit index (GFI), which would index the degree of discrepancy. In some respects, this was an eminently sensible thing to do. For, if it is admitted that the χ^2 test was after all likely to reject a model where the sample size was large, even though the same model might fit if the sample size was reduced (artificially or by sub-sampling the main dataset), then something which might at least index the degree of discrepancy would provide an investigator a second source of information which might be used to adjudge the worth or otherwise of a model. In other respects, the development of an "approximate fit" index was simply wrong in principle and in practice.

Specifically, in other areas of science, model fit is adjudged according to how well a model predicts or explains that which it is designed to predict or explain. In these other areas of science, that to be predicted or explained usually involves occurrences, discrete classes of "things", or

measured outcomes. The task involved in assessing fit is to determine the cross-validated (replicable) predictive accuracy of the model in accounting for the outcomes. Such models invariably rely upon some kind of theory in their construction, and thus predictive accuracy is not just an empirical matter but will also have implications for the theory upon which the model and its parameters have been constructed. In assessing fit, an investigator might take a multi-faceted approach using various types of cross-validation strategy including bootstrapping, V-fold cross-validation, and the use of training/initial model-fit and holdout samples (Hastie, Tibshirani, & Friedman, 2001). Some derivative of an information-theory based test such as the Akaike or Bayesian Information Criterion (AIC, BIC) is also likely to be used. For some regression-type models, some form of inferential statistical test of the model parameters and/or residuals might be made (Harrell, 2001). Within SEM, the χ^2 NHST is employed so as to test for the statistical insignificance or otherwise of a model in terms of its accounting for the discrepancies between model implied and observed covariances. The test is blind to whether the model actually predicts or explains anything to some substantive degree. That requires a separate form of analysis akin to that used by those who aspire to predictive accuracy as the most important (but not sole) arbiter of any model (Breiman, 2001; Mosteller & Tukey, 1977). Indeed, Gigerenzer (2004) calls the application of NHST statistical reasoning "mindless"; so in a certain respect is the application of the χ^2 test in SEM research. However, whereas Gigerenzer is speaking more of the state of mind of investigators who adopt strong positions on the application of NHST methods in research, I am using this term as the descriptor of a statistical test which is only concerned with the overall discrepancy of model implied and sample covariances.

But, mindless or not, the χ^2 test is the ONLY statistical test for SEM models at the moment. It is not totally "exact" as its proponents maintain, as it relies upon setting an arbitrary alpha level to adjudge whether the χ^2 test statistic is to be considered significant or not. As Fisher (1956) stated (reported in Gigerenzer, 2004) ... "*no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas*". However, once that alpha level is set subjectively, the logic of the test is crystal clear, and it becomes "exact". At that alpha level, given we clearly specify the various assumptions under which we make the test are to be considered valid, then either the residuals exceed levels expected by chance, or they do not. To avoid using this test as the test of the statistical fit of a model can only be warranted if the investigator seeks the alternative multi-faceted model *assessment* approach outlined above (I use that word "assessment" on purpose – because that is how models are evaluated, not by a simple binary test result, but from a multiple-perspective set of assessments which are primarily concerned with predictive accuracy, parsimony, as well as theoretical meaningfulness). However, for this kind of assessment approach to work, there has to be some theory-substantive or real-world substantive criterion which can serve as the standard of predictive accuracy of a model, and that such predictive accuracy can be replicated using at least one cross-validation strategy. The judgment of model-fit is thus subjective, but based upon a powerful criterion of predictive accuracy. When the model is theory-based, cross-validated predictive accuracy is even more compelling as the sole arbiter of model "acceptability".

However, many models in SEM research involve no measurable outcomes and are not linked to any "real-world" criteria; in fact they are simply "statistical" models for explaining covariation between several variables (as in confirmatory factor analysis: CFA) . Unlike the approach how-

ever in exploratory factor analysis (EFA), where latent variables are constructed "blindly" or "mechanically" from the data by the process of eigenvalue decomposition, in SEM applications, some form of theory or prior knowledge is used by an investigator to "create" one or more latent variables which are considered as "causal" for the observed or manifest variation in the variables. One or more such models may be created. Now the task becomes how to decide which model might be considered the best fit to the data, especially since the SEM model is considered to embody causality. Model fit assessment is now a real problem. There is no obvious external variable criterion outcome against which a model might be tested or assessed except the observed covariance matrix itself, and a model-implied covariance matrix.

Where it has all gone badly wrong in this type of SEM model testing is that many investigators have actively avoided the statistical test of fit of their models, in favour of a collection of ad hoc indices which are forced to act like "tests of fit" rather than indices of "degree of approximation of discrepancy". The problem is that no-one actually knows what "approximation" means in terms of "approximation to causality" in the many areas in which SEM modelling is used. The criterion used for "fit" is actually an abstract concept in the majority of SEM models. It is clearly not predictive accuracy. In fact whether models "approximately fit" with an RMSEA of 0.05 or 0.07 is a literally meaningless scientific statement. What actual consequence is imparted by the acceptance of a model with an RMSEA of 0.05 or 0.08? What is the substantive scientific consequence of accepting a model with a CFI of 0.90 rather than one of 0.95? There is none associated with such statements. Standards of correctness using this kind of "degree of approximation" of a "causal" model have been confounded by SEM investigators with the rather more definitive standards of correctness found when using a criterion which is associated with replicable predictive accuracy. That is the illogic which now permeates "approximate" model-fit testing in SEM research. One might also note with envy that in the more demanding measurement-oriented area of latent variable item response theory, the notion of "approximate fit" does not even exist. Models either fit, or they do not fit. When SEM investigators use their methodology for analysis of variance, they use the χ^2 tests as they would *F*-ratio tests within conventional MANOVA and ANOVA. When SEM investigators use the methodology for DIF or bias analysis, adherence to the statistical test is what permits a researcher to identify bias for an item (Raju, Laffitte, & Byrne, 2002; Reise, Widaman, & Pugh, 1993). No-one speaks of approximate bias, or approximate mean difference. Yet, when it comes to the majority of other models in marketing, organizational psychology, individual differences and questionnaire psychometrics, approximate fit suddenly becomes the order of the day. Not only is this inconsistent, it is also illogical.

6. Recommendations

Well, now what? This journal (Personality and Individual Differences: PAID), as with many other flagship journals in the behavioural, business, and social sciences publishes many papers which utilize SEM methodology. Few of the models in any of these journals fit via the χ^2 test, indeed some authors of submitted PAID manuscripts fail to even report the test result and probability, needing a reminder by some diligent reviewers. That is the extent to which the test is ignored. Instead, a raft of approximate fit indices and ad hoc thresholds are put forward to justify "acceptable model-fit". Indeed, one gets the feeling that social scientists cannot actually contem-

plate that most of their models do not fit their data, and so invent new ways of making sure that by referencing some kind of ad hoc index, that tired old phrase "acceptable approximate fit" may be rolled out as the required rubber stamp of validity. Look at the many papers published which simply analyse questionnaire data using CFA, fitting multiple models to the covariance data. Invariably (but not always), given a sample size above 200 cases, no models fit via the χ^2 tests. But the authors will then proceed to utilize the nested model χ^2 test to select the model which is the most significantly different from a baseline or other sequenced and nested models. Again, the fact that the model which shows most improvement still fails to fit via the global test is barely ever reported. Further, no attempt is ever made to explain the real-world or substantive theoretical consequences of accepting one model over another. For example, by substantive, I mean that where say a multiple correlated factor vs a Schmid-Leiman hierarchical general factor vs a nested general factor model might be the models in contention, how many authors bother to explain or even empirically demonstrate the consequences of accepting each model in terms of something other than "minimised covariance discrepancy"? The problem is that many SEM investigators are not prepared to understand that model fitting is a time-consuming process, fraught with many kinds of difficulties, and invariably requiring huge amounts of work to produce a well-fitting, parsimonious, cross-validated model which accounts for something more substantive than variation in a "covariance matrix".

However, whilst the SEM practices of many social scientists may be less than optimal, the use of SEM as a powerful data analysis and causal modelling tool is here to stay. It is an extremely useful methodology for a particular kind of model fitting. So, what advice might be given to referees of papers and authors, let alone the associate editors and chief editor of the journal, given the above? I would state the following recommendations:

1. *Test of Fit.* The χ^2 test for any SEM model fit must be reported in the same way one would report the results for any statistical test. A statement to the effect that the model fits or fails to fit via this statistic must be provided. If the model fits via this test, and the sample size is reasonable (above 200 say), then the investigator might reasonably proceed to report and discuss features of the model accordingly. This is the only statistical test for a SEM model fit to the data. A problem occurs when the sample size is "huge", as stated succinctly by Burnham and Anderson (2002). They note that "model goodness-of-fit" based on statistical tests becomes irrelevant when sample size is huge. Instead our concern then is to find an interpretable model that explains the information in the data to a suitable level of approximation, as determined by subject-matter considerations. Unfortunately, in life sciences we probably never have sample sizes anywhere this large; hence statistical considerations of achieving a good fit to the data as important as well as subject-matter considerations of having a fitted model that is interpretable" (p. 220). The numbers being used in examples of "huge" datasets by Burnham and Anderson are of the order of 10,000 cases or more. Not the 200 s or so which seems to be the "trigger" threshold at which many will reject the χ^2 test as being "flawed"!
2. *Sample size.* SEM analyses based upon samples of less than 200 should simply be rejected outright for publication unless the population from which a sample is hypothesised to be drawn is itself small or restricted in size (measured in the hundreds or thousands say, as might be the case in medical research). The issue of interest here, as in all statistical sampling,

is how well that sample might be said to contain all likely members of a specified population. Using 200 1st year undergraduate university students might be fine for a model using or invoking variables which are deemed universal amongst all members of the human race, irrespective of age or gender, but useless for a set of clinical variables which might only be appropriate to individuals possessing say psychopathic tendencies. I hesitate to advise authors to test the power of their model as the reality is there is no straightforward test of the kind one might see implemented for almost any other statistical test. The RMSEA-based analysis is practically useless because by using it an investigator is already "buying into" the notion of an RMSEA approximate fit index. The procedure outlined by Muthen and Muthen (2002) is simply beyond the computational stamina of many SEM investigators to implement as a routine method; but at the moment, as a coherent method for power analysis in SEM, it is all there is available to an investigator. It may well be the price that has to be paid for using a modelling procedure and subsequent goodness-of-fit NHST which is so complex that no routine way of assessing its statistical power exists.

3. If an author decides that the χ^2 test result will be considered a valid hypothesis test, and the test indicates model misfit, then three courses of action appear to remain:
   (a) Examine the data for the validity of the assumptions implied by the use of maximum-likelihood estimation (or whatever estimation method was used on the data). If say the data are not multivariate normal, then this might be the reason why the χ^2 test has yielded a significant result. At this point, the data might need to be transformed, re-scaled, or even some variables dropped altogether from the analysis in order to refit a model using data which meets the assumptions of the estimation methodology and hypothesis test.
   (b) If the assumptions appear reasonable from #1 above, then just report the model as failing, and discuss the significance of this to the theory on which the model was based, and to the domain in which the theory plays a role.
   (c) If the assumptions appear reasonable from #1 above, and an author is curious to explore further, then begin examining the residual matrix for clues as to where misfit is occurring, adjust the model accordingly, refit, and proceed in this way to explore minor-adjustment alternative models until either fit is achieved, or where it becomes obvious that something is very wrong with the a priori theory from which this class of models is being generated. Either generate a new class of model which is based upon a new theory, or simply report the evaluative process undertaken and suggest that a new theory might be required for the area.
4. If an author decides that the χ^2 test result will be ignored, then a clear statement of this and a rationale for this action must be provided. It is always likely to revolve around the argument that as a sample increases in size, so will increasingly trivial "magnitude" discrepancies between the actual and model implied covariance matrix assume statistical significance. However, simply harnessing a raft of ad hoc approximate fit indices using "Golden rule"/"credentialed person authorised" cutoffs can no longer be considered acceptable given the recent evidence from the papers quoted above. 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". Model fit can no longer be claimed via recourse to some published "threshold-level recommendation". There are no recommendations anymore which stand serious scrutiny.

5. Given #4 above, then one of two courses of action will need to be chosen:
(a) If the SEM model includes real world or "measurable" criterion classes or outcomes of one form or another, then strategies for determining cross-validated predictive accuracy and model parsimony via AIC/BIC indices might prove most effective. Here the argument is that "good enough/empirically adequate for theory-purposes or pragmatic use" multi-facet cross-validated predictive accuracy might supersede the use of a single global statistical discrepancy test. But, this strategy will only work where an outcome or consequence for a model can be evaluated quantitatively (whether by class composition, frequencies, or continuous-level variable predictions). It is also likely to be computationally intensive. The typical statement about such a model is that although it fails to fit via the χ^2 test, given the huge sample size or the cross-validated predictive accuracy of the model, it might be considered "good enough" for practical purposes. Note, no strong claim is being made about model fit or misfit – rather the claim is to "empirical adequacy".
(b) In a CFA, there are no "outcome variables". Latent variables are hypothetical, considered causal, exogenous, variables. So, the model being fit can only be assessed using the discrepancy between model implied covariances and the observed covariances. Here, the χ^2 test is the obvious statistical test of fit. But, in choosing to ignore its result, an investigator might proceed to show that further exploratory or descriptive analyses of the residual matrix (perhaps standardized/normalized for ease of interpretation) might lead them to conclude that the test result is misleading, or that it appears to have little consequence in terms of the distribution, location, and/or size of residual discrepancy. An alternative strategy might be to include some criterion variables external to the SEM analysis, such that alternate latent variable conceptualisations of say something like "Emotional intelligence" might be examined with regard to their predictive accuracy against some real-world or behavioural criteria for "emotional intelligence". Likewise, if summated item questionnaire scales are being modelled, then the usual psychometric criteria of test score reliability and manifest scale score correlation (as against latent variable correlation, as these are useless in real-world applications as they can never be realised in practice) might be a minor but useful indicator of the effects of accepting one model over another. Personally, I must admit that the problem for CFA is exactly the same problem which EFA has always had; that is, all solutions are entirely "inward facing". That is, there are no consequences for accepting any kind of model or factor structure unless either strong theory with substantive explanatory and measurable outcomes is dictating the process, or alternate factor models are evaluated directly using external criterion variables. It is useful here to look at the statement from Reise et al. (1993)... "*no CFA model should be accepted on statistical grounds alone; theory, judgment, and persuasive argument should play a key role in defending the adequacy of any estimated CFA model*" (p. 554). This is almost all wrong. What if the χ^2 test is actually valid (given the validity of its assumptions about the data) and is based upon a "non-huge" sample size? To me this seems a very dangerous statement to make as it appears to justify the superiority of personal judgment and "credentialed person" argument over an objective test of model fit (Meehl, 1997). Where a sample size is huge, judgment and persuasive argument are still no answer to empirical outcome analysis of the consequences of model selection. Where that is not possible, one wonders why the model is even being spoken of as a model, when it apparently possesses no means of validation outside of the operations employed to attempt to fit it

to some data along with some accompanying subjective judgment and persuasive "argument".

What must be understood by any investigator using SEM is that by rejecting the result of the χ^2 test, which is the most direct and obvious test of model fit, other substantive methods of demonstrating and justifying that a chosen model might still be considered theoretically useful or worthwhile will need to be deployed. Ad hoc "approximate fit" indices fail miserably here. Why? Because there is no obvious consequence as to whether the approximation is good, bad, or indifferent. The χ^2 test itself also carries no "consequences" in this regard, but it is a simultaneous test that all residuals are statistically zero, albeit at a subjectively specified alpha test level. Approximate fit indices do say something about "fit" – much in the same way that a measure of agreement indicates something about similarity. But judging the consequence of accepting a particular value of an index indicative of "approximate fit" is not a matter for abstract "thresholds" but ideally the empirical calculation of the actual consequences of using that value as indicative of "useful or approximate fit". I have to wonder whether an "accepted" model which has no empirical consequences over and above a "rejected" model is worth publication, let alone any kind of scientific consideration.

7. In conclusion

SEM is a modelling tool, and not a tool for "descriptive" analysis. It fits models to data. These models require testing in order to determine the fit of a model to the data. If an investigator refuses to use or acknowledge the results from the one statistical test available to assess model fit, then something other than some ad hoc scaled discrepancy index will be required to justify the worth of the model. I have provided the logic of how to approach assessing model acceptability if a sample size if huge, or where the assumptions of the χ^2 test do not hold in the sample. I anticipate shrieks of indignation and outrage at the hard line adopted here. But, social science, business, marketing, and management journals are littered with the consequences of the use of "approximate fit indices" – a plethora of forgettable, non-replicable, and largely "of no actual measurable consequence" models. Investigators still have a choice to use SEM methodology or something other than SEM to explore phenomena. If an investigator wishes to avoid using NHST or inferential, hypothetical sampling distribution data model statistics, then many other methodologies exist in the statistical toolbox (Breiman, 2001). But, if SEM is used, then model fit testing and assessment is paramount, indeed crucial, and cannot be fudged for the sake of "convenience" or simple intellectual laziness on the part of the investigator.

References

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*(1), 41–75.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley, pp. 263–269.

Breiman, L. (2001). Statistical modeling: the two cultures. *Statistical Science, 16*(3), 199–231.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

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.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh: Oliver and Boyd.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*, 587–606.

Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

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.

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*(3), 320–341.

Meehl, P. (1997). Credentialed persons, credentialed knowledge. *Clinical Psychology: Science and Practice, 4*(2), 91–98.

Mosteller, F., & Tukey, J. (1977). *Data analysis and regression*. Readding, MA: Addison-Wesley.

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.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: a comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517–529.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552–566.

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