MULTIVARIATE BEHAVIORAL RESEARCH, *41*(3), 227–259

Quantifying Parsimony in Structural Equation Modeling

Kristopher J. Preacher
University of North Carolina at Chapel Hill

Fitting propensity (FP) is defined as a model's average ability to fit diverse data patterns, all else being equal. The relevance of FP to model selection is examined in the context of structural equation modeling (SEM). In SEM it is well known that the number of free model parameters influences FP, but other facets of FP are routinely excluded from consideration. It is shown that models possessing the same number of free parameters but different structures may exhibit different FPs. The consequences of this fact are demonstrated using illustrative examples and models culled from published research. The case is made that further attention should be given to quantifying FP in SEM and considering it in model selection. Practical approaches are suggested.

Models are commonly constructed in an attempt to approximate or explain some process of scientific interest that cannot be directly observed. The ability to predict other (or future) data arising from the same latent process is often seen as a mark of a model's usefulness or quality, and it is commonly assumed that a model's fit to a given sample provides a good clue to this predictive ability.[1] But it is also recognized that some models are simply better able to fit data than other, more parsimonious models; that is, competing models often differ in terms of their *fitting propensity* (*FP*), or average ability to fit data. Consequently, model fit adjusted for FP is often used as a way to distinguish between competing models, taking into account differences in model parsimony. Adjusted fit is traditionally quantified by combining two properties of a model: *parsimony* and *goodness of fit*. In this article,

[1]This article limits attention to research scenarios comparing rival theories in a deductive, top-down mode. This is only one of many ways to conducting legitimate scientific research. Many of the ideas presented in this article may also be applied to exploratory model development conducted in an inductive, data-driven mode.

Correspondence concerning this article should be addressed to Kristopher J. Preacher, University of Kansas, Department of Psychology, 1415 Jayhawk Boulevard, Room 426, Lawrence, KS 66045–7556. E-mail: preacher@ku.edu

I address the balance between parsimony and fit in structural equation modeling (SEM) and how this balance affects the practice of model selection. I argue that the traditional approach of adjusting fit indices for the number of free model parameters can yield misleading judgments of relative model quality.

To illustrate the problem introduced by FP, consider Figure 1. Models A and B might represent two competing theories about how V_1, V_2, and V_3 are related in the population. Using methods discussed in more detail later, I generated 10,000 random 3 × 3 correlation matrices and fit both Models A and B to all of the data. In terms of absolute fit, the average root mean squared residual (RMSR) associated with Model A was .246 (poor by most standards), whereas the average RMSR associated with Model B was .082 (much better). Thus, Model B fit random data better than did Model A overall (note that sample size was irrelevant in this example). Figure 2 contains cumulative distribution plots of RMSR for both models.

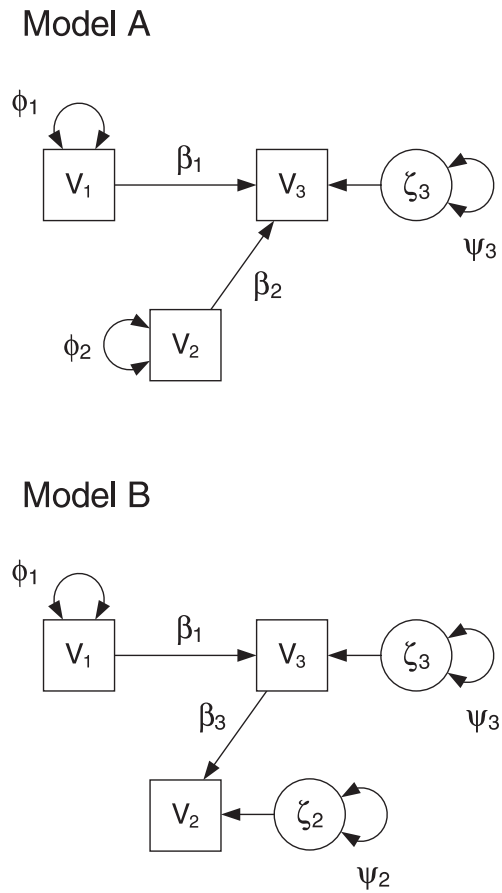


FIGURE 1 Two path models, each with five free parameters.

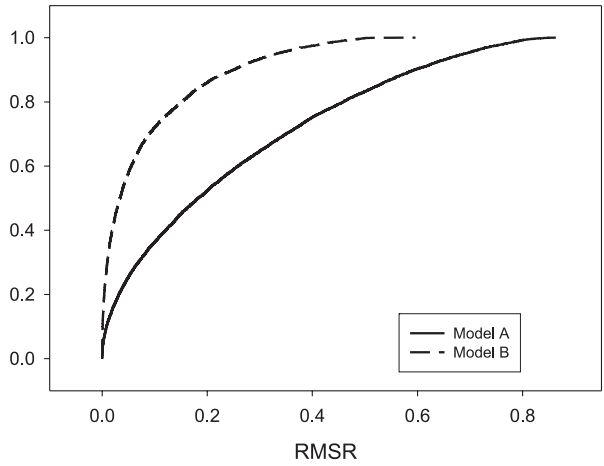


FIGURE 2 Cumulative frequency distributions of root mean squared residual for Models A and B fit to the same 10,000 random correlation matrices.

Both models are able to fit some data sets well, but Model B possesses a clear advantage. If the researcher's goal is to select the model that most accurately predicts data generated by the underlying process, Model B might be chosen for reasons that have little to do with the quality of the theory it represents and more to do with the model's intrinsic ability to fit data arising from other processes. Note that Models A and B each have five free parameters and, in that limited sense, might be considered equally parsimonious.

GOALS AND ORIENTATION

My motivations are to (a) explicate the concept of FP to researchers who employ SEM, (b) show that adjusting fit indices for FP by traditional means is not an optimal approach, and (c) suggest means by which FP can be routinely (and more thoroughly) considered in applications of model selection in SEM. To limit confusion, I use terminology that does not conflict with existing terms (e.g., I use the term FP rather than *scope*, *complexity*, or *flexibility*, which are commonly used in the mathematical psychology literature but would be potentially ambiguous in the SEM context).

Throughout what follows, my emphasis is on model selection rather than on the evaluation of models in isolation or on hypothesis testing. Model selection is an approach to science that does not derive directly from the hypothetico-deductive tradition currently prevalent in the social sciences. Although model evaluation and hypothesis testing are important, my emphasis on model selection is in accord with the current zeitgeist in the philosophy of science, which holds that relatively little

information of scientific value is gained by evaluating models against arbitrary benchmarks. In model selection, at least two theories are compared in light of observed data to determine which one is preferable. Over several replications, evidence will tend to accrue in favor of the theory that fits data well and in the most parsimonious manner (Lakatos, 1970; Meehl, 1990). When rival theories are to be compared using fit to real data, sampling variability and uncertainty about the models' relative difference in FP can complicate the selection process.

First, I present an overview of the literature to orient the reader to basic ideas and associated concepts. Second, I explain how FP is currently quantified in the SEM paradigm. Third, by adapting an approach devised by Botha, Shapiro, and Steiger (1988), I demonstrate how quantifying FP by traditional means can lead to problems of interpretation and inference in SEM. Finally, I suggest that research should be devoted to developing ways to quantify FP routinely in the evaluation and comparison of structural equation models. Specifically, attention will be devoted to extending two selection criteria—the *uniform index-of-fit* (UIF) of Botha et al. and the *minimum description length* (MDL) criterion of Rissanen (1989)—for practical application in SEM. The general concepts presented here can be understood without reference to any particular fit index or estimation method, although I make use of specific methods for the sake of illustrating important concepts.

FP AND RELATED CONCEPTS

In this section I discuss several concepts crucial to understanding the issues at stake, including FP, goodness of fit, parsimony, generalizability, and overfitting. Afterward I discuss how these concepts facilitate a greater understanding of the importance of FP.

FP

FP is the ability of a model to fit a diverse array of data patterns well by some criterion of fit (Dunn, 2000; Myung & Pitt, 1997, 2004; Pitt, Myung, & Zhang, 2002). FP is also commonly called *model complexity*, *scope*, or *flexibility*. Closely allied concepts are Meehl's (1990) *tolerance* and Bamber and van Santen's (2000) *prediction range*. It is useful to think of FP as the average fit of a model to regions of the *data space*, or the space containing all empirically obtainable data patterns relevant to a particular modeling domain (the data space is what Meehl, 1990, termed *Spielraum* and what Bamber and van Santen, 2000, termed *outcome space*). FP can be understood as the complement to parsimony. Models with greater FP than their competitors enjoy an advantage in terms of goodness of fit for reasons potentially unrelated to the model's approximation to the data-generating process. FP per se is

not an undesirable property of a model. All models have some ability to fit data, otherwise their usefulness would be very limited. But without devoting explicit attention to FP, it is impossible to know how much of a model's good fit is due to a theory's genuine predictive ability and how much is due to the model's inherent ability to fit data arising from unrelated processes and random error. FP is the subject of increasing attention in mathematical psychology and allied fields (e.g., Collyer, 1985; Cutting, Bruno, Brady, & Moore, 1992; Dunn, 2000; Myung, Balasubramanian, & Pitt, 2000; Myung, Forster, & Browne, 2000) but has yet to attract serious attention in the SEM literature.

Goodness of Fit

Goodness of fit is the empirical correspondence between a model's predictions and observed data. If the match between the model's predictions and observed data is deemed adequate (by reaching or exceeding some benchmark), the model is said to show *good fit*, an indication that the theory represented by the model has received support. When a fit index is used to evaluate a model in opposition to at least one other theoretical model, the index is termed a *model selection criterion* because the object is to select the model that is optimal in some sense, given the data.

Parsimony and Degree of Falsifiability

A model's parsimony[2] can be cast as its ability to constrain possible outcomes (Popper, 1959; Roberts & Pashler, 2000) or restrict the proportion of data sets consistent with the model. Parsimony is closely related to *degree of falsifiability*, the capacity of a model to be empirically disconfirmed. It is important that a model have the potential to be disconfirmed by data inconsistent with theory (Popper, 1959), otherwise a theory could not be realistically subjected to scientific scrutiny in the form of risky tests. But falsifiability alone is often not enough to permit truly risky tests. Some models, even if they are technically falsifiable, possess the ability to easily fit a wide array of data. Such models often possess an advantage in model selection when compared to more parsimonious rival models.

Generalizability

Generalizability is a model's ability to fit regularity (reliable, theoretically meaningful variability that is liable to remain stable across repeated sampling from the

[2]The term *parsimony* has come to be equated with having relatively few free parameters or relatively many degrees of freedom in the factor analysis and SEM literature (Mulaik, 2001; Mulaik et al., 1989). The terms are used here in their broader sense to encompass all factors contributing to a model's degree of falsifiability.

same reference population) in data (Myung & Pitt, 2004; Myung, Pitt, & Kim, 2004). Ultimately, researchers are interested not only in how well a model can describe the data in hand but also in how well a model (or, more precisely, a model with its parameters constrained to a particular set of estimated values) can describe other data generated by the same underlying process (Forster & Sober, 1994; Linhart & Zucchini, 1986; Pitt et al., 2002). Thus, generalizability can also be understood as predictive validity, or the potential to cross-validate well using Bentler's (1980) *tight replication strategy*, in which all of a model's free parameters are fixed to values estimated in one sample before fitting the model to a validation sample drawn from the same population. Good fit in the validation sample reflects high generalizability.

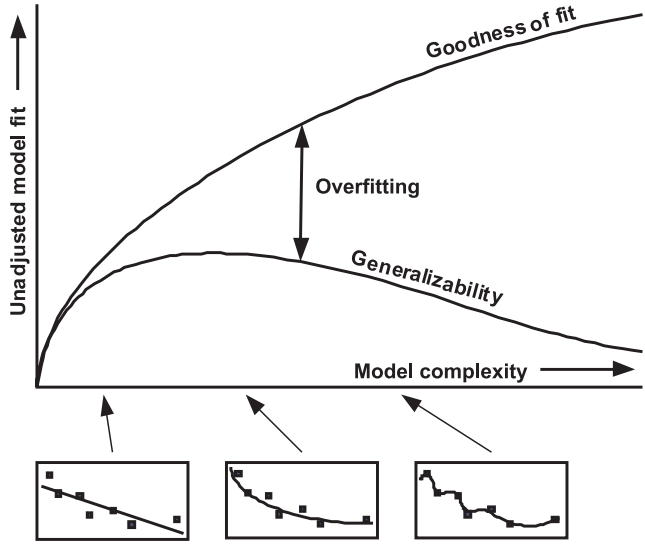
Generalizability represents a balance between goodness of fit and parsimony and is often quantified by combining a measure of model fit with some measure of FP by penalizing fit for FP. Unfortunately, generalizability and goodness of fit, although both desirable characteristics of a model, are not always positively related (Myung, Balasubramanian, & Pitt, 2000).

Overfitting

Overfitting is the tendency for a model to show good fit by capturing error (noise) as well as regularity. Overfitting is a common danger when FP is not adequately considered. A model with relatively high FP may fit a given data set very well but may not generalize to other samples easily (Forster & Sober, 1994; Roberts & Pashler, 2000) or may not cross-validate well using Bentler's (1980) tight replication strategy.

Summary

The relationships among FP, goodness of fit, parsimony, generalizability, and overfitting are illustrated in Figure 3. The *y*-axis in Figure 3 loosely represents goodness of fit, unadjusted for FP. Models with higher FP (lower parsimony) tend to exhibit better fit to data relative to models with lower FP. On the other hand, generalizability reaches a maximum and then decreases as FP increases (Pitt & Myung, 2002). As FP progresses beyond the point of maximum generalizability, overfitting occurs. The implication is that selecting one model from a set of competing alternatives solely on the basis of superior fit may favor the retention of models with higher FP yet lower generalizability (see, e.g., Browne & Cudeck, 1992; Collyer, 1985; Cudeck & Henly, 1991; Forster & Sober, 1994). In addition to generalizability, assessment of the overall quality of a model is (ideally) also based on some combination of parameter estimates, substantive interpretability, faithfulness to theory, explanatory adequacy, ability to generate new research, and the model's historical performance relative to other models intended to account for the



(from Pitt & Myung, 2002)

FIGURE 3 The relationship between generalizability, goodness of fit, and overfitting. As fitting propensity (FP) increases, goodness of fit increases. However, generalizability reaches a maximum. If a model is too complex, overfitting is a danger. The three models below the plot are arranged in increasing order of FP. The model on the left does not fit the data well. The model on the right fits very well, but overfits the data and is not likely to fit well in future samples. Only the model in the middle maximizes generalizability. Adapted from *Trends in Cognitive Science, 6*(10), M. A. Pitt and I. J. Myung, *When a good fit can be bad*, p. 424, Copyright 2002, with permission from Elsevier.

same phenomena (Cutting et al., 1992; Marsh & Balla, 1994; Myung & Pitt, 2004). However, among these aspects of a model's quality, generalizability is the only aspect that has the potential to be quantified and objectively assessed, specifically by trading off fit and parsimony. Consequently, it is desirable that FP and goodness of fit be gauged as accurately as possible to enable fair model comparison. The following is a brief discussion of some of the factors that affect FP and an explanation of how these factors are manifested in SEM.

MODEL FIT AND FP IN SEM

Overview of SEM

SEM involves specifying theory-implied, usually linear relationships among a set of latent and observed variables. The formal representation of the model is $\Sigma = \Sigma(\theta)$.

The parameters of these models (θ) are traditionally specified as either freely estimated or fixed to specific values. The fitting process involves finding a set of parameter estimates $\hat{\theta}$ which minimize a loss function, *F*, usually the maximum likelihood (ML) or ordinary least squares (OLS) discrepancy function. The estimates $\hat{\theta}$ minimizing *F* yield an implied covariance matrix $(\hat{\Sigma})$ as similar as possible to the observed covariance matrix (**S**) of measured variables. To the degree that $\hat{\Sigma}$ resembles **S**, $\hat{F}$ (the minimized *F*) will tend to be small, reflecting good fit. When $\hat{\Sigma} = \mathbf{S}$, $\hat{F} = 0$, denoting perfect fit. Good fit is typically interpreted as support for the model and therefore as support for the theory the model represents.

Factors Affecting FP in SEM

The number of free parameters. A number of model characteristics affect FP. Chief among these is the effective number of free model parameters (*q*), defined as the number of free parameters minus the number of functional constraints placed on otherwise free elements of θ. All else being equal, models with larger *q* are better able to fit data (Forster & Sober, 1994; Jeffreys, 1957; Wrinch & Jeffreys, 1921). In model selection settings in which all competing models are posited to account for relationships among the same *p* variables, the degrees of freedom (*df*) of the models contain information inversely related to *q*. Freeing model parameters reduces the number of dimensions in which observed data are free to differ from model-implied data (Mulaik, 2001, 2004). FP granted by free parameters is termed *parametric complexity* (Markon & Krueger, 2004). The tendency for more free parameters to lead to better fit could create situations in which a model is selected not because it is the best in any meaningful sense but simply because it has greater parametric complexity (James, Mulaik, & Brett, 1982; Myung, 2000; Roberts & Pashler, 2000; Steiger & Lind, 1980).

Functional form. Most fit indices and selection criteria represent trade-offs between fit and parsimony, with parsimony defined strictly in terms of *q*. However, FP is not governed completely by *q* (Keuzenkamp & McAleer, 1997). *Functional form* also contributes to FP (Jeffreys, 1931; Pitt & Myung, 2002; Roberts & Pashler, 2000; Wrinch & Jeffreys, 1921). Functional form refers to the specific means by which relationships among variables are expressed as functions of free model parameters. In SEM, functional form refers to the set of simultaneous equations relating observed variances and covariances to free parameters. Models with different functional forms usually have different FPs. For example, Model B in Figure 1 has a greater FP than Model A, even though the two models have the same *q*. However, models that are chi-square equivalent (i.e., those models that are indistinguishable on the basis of fit for any **S**; see MacCallum, Wegener, Uchino, & Fabrigar, 1993) have equal FP even though their structures may appear at first to be distinct. FP granted by a model's unique functional form is termed *structural complexity* (Markon & Krueger, 2004).

Quantification of Model Fit in SEM

Many indices have been suggested to help researchers judge the match between model and data. Various indices adjust for FP in different ways. Some include no adjustment. An example of an index that does not adjust for FP is the RMSR (Jöreskog & Sörbom, 1996), used earlier in the introductory example. RMSR is defined as

$$RMSR = \sqrt{\frac{tr(\mathbf{S} - \widehat{\Sigma})^2}{p(p+1)}}. \quad (1)$$

For correlation matrices, RMSR is equivalent to the standardized root mean squared residual (Bentler, 1995).

In recognition of the fact that models with more free parameters tend to yield better fit, many indices penalize models for *q*. For example, the root mean square error of approximation (RMSEA, or ε; Browne & Cudeck, 1992; Steiger & Lind, 1980) is an index reflecting the difference between the population covariance matrix and the implied (reproduced) covariance matrix (Cudeck & Henly, 1991). A sample estimate of RMSEA is $\hat{\varepsilon} = (\widehat{F}_0 / df)^{1/2}$, where $\widehat{F}_0$ is an estimate of the population ML discrepancy function value. Because models with fewer *df* have more free parameters, RMSEA favors models with fewer free parameters, all else being equal (Browne & Cudeck, 1992; Steiger, 2000). However, for models with the same *q* but different functional forms, RMSEA is likely to favor models with greater structural complexity. Another fit index that includes an adjustment for *q* is the Tucker-Lewis Index (Tucker & Lewis, 1973). In addition, James et al. (1982) and Mulaik et al. (1989) suggest that a *parsimony ratio*, the ratio of *df* in the tested model to *df* in an appropriately specified null model, can be multiplied by most fit indices to yield parsimony-adjusted indices. For indices employing *q* or *df* in their formulae, the inclusion of *q* or *df* was not always included as an adjustment for FP per se.

Model selection criteria are indices intended select the simplest model that still adequately explains the observed data. Examples include the Akaike Information Criterion (Akaike, 1973) and the expected cross-validation index (Browne & Cudeck, 1989). Model selection criteria can be seen as formalizations of Occam's razor. These measures tend to be sensitive to sample size, selecting models with higher FP as the sample size increases (Cudeck & Browne, 1983; Cudeck & Henly, 1991; McDonald & Marsh, 1990). This is to be expected; more information accrues with larger samples, and models with higher FP can be selected with greater confidence. At small sample sizes, these criteria are more conservative. Thus, the FP adjustment employed in such criteria is moderated by sample size (Marsh & Hau, 1996).

When multiple models are to be compared, it is assumed that p is constant; that is, all of the competing models are intended to explain relationships observed among the same p variables. For a fixed p, df is a simple function of q. For a given set of competing models, it follows that using df as a penalty for FP reduces to adjusting for q for purposes of model selection. Even when the express intent was to adjust for FP, no indices were ever claimed to account for FP completely. Because functional form is also known to contribute to FP, suboptimal models may be selected when FP is equated with q alone (Raykov & Marcoulides, 1999). In the next section I present a graphical simulation approach to assessing FP in SEM.

INVESTIGATING FP IN SEM

Goals for Illustrative Examples

A straightforward and intuitive way to gauge relative FP is to generate data uniformly representative of the theoretical domain (the data space described earlier), apply the competing, theory-derived models to the generated data, and assess how well the models fit the data relative to one another. Fitting models to representative data has been suggested as an accurate way to discover what, exactly, a theory predicts (Roberts & Pashler, 2000) and is how FP is operationalized in advanced model selection criteria such as the UIF (Botha et al., 1988) and MDL (Rissanen, 1989) criteria to be described in a later section. Roberts and Pashler suggested limiting the data space to consist not of all *possible* data patterns but to all *plausible* ones in particular. Criteria for deciding what might constitute *plausible* data are addressed in the Discussion section. In the study presented here, the plausible data space was limited to uniform coverage of correlation matrices lying in the *positive manifold*.[3] The requirement of uniform coverage was considered important because it was desirable to test models using data representative of every part of the plausible data space. The data generation method employed here owes much to the logic and approach of Botha et al. (1988).

[3]A positive manifold is a (locally) Euclidean space in which all coordinates are positive. Thurstone (1935, 1947) used this phrase to describe the space consisting of coordinates defined by factor loadings λ_{ij} such that all $\lambda_{ij} > 0$. A consequence of this condition is that population correlations implied by a set of loadings will all be positive. The term *positive manifold* is used informally to refer to the tendency for a set of variables to be positively intercorrelated. For the models considered in this study, positive manifold correlation matrices were judged to be an adequate representation of the data space for most SEMs, at least for the sake of illustration.

Data Generation

Random correlations. A correlation matrix (**R**) is defined as a standardized covariance matrix—any symmetric, positive semidefinite matrix with unit diagonal elements and with off-diagonal elements r_{ij}, $i \neq j$, such that $\{-1.0 \leq r_{ij} \leq 1.0\}$. For our purposes here, **R**s are defined to be positive definite (i.e., all eigenvalues must be greater than zero) and are restricted to contain no negative elements. To simulate the data space, I generated random[4] **R**s such that every possible positive-manifold **R** had an equal probability of being generated. I chose this criterion for randomness because it does not rely on preconceived ideas about data patterns that are more or less likely to appear in practice. Matrices generated according to this criterion can be considered uniformly distributed across the data space of interest (Botha et al., 1988).

Three algorithms for generating random **R**s were suggested and used by Botha et al. (1988).[5] One strategy, the uniform correlation matrix (UCM) method, involves generating random square, symmetric matrices with ones along the diagonal and off-diagonal elements drawn from a {0, 1} uniform distribution. Only positive semidefinite matrices are retained. This method yields matrices that are distributed uniformly on the data space of interest but becomes increasingly inefficient as matrix order increases. A new strategy was sought to yield matrices with the desirable distributional characteristics of Botha et al.'s UCM method, yet with greater efficiency. Therefore, a Markov Chain Monte Carlo (MCMC; see Gilks, Richardson, & Spiegelhalter, 1996) algorithm was employed. The MCMC approach employs much the same logic as the UCM method; that is, it searches the space containing all matrices with unit diagonals and off-diagonal elements in {0, 1} for matrices meeting the criteria and evaluates each matrix separately on an accept/reject basis. However, the MCMC algorithm narrows the search to a region likely to yield acceptance (see the Appendix for greater detail).

ILLUSTRATIVE EXAMPLES

It is well known that if Model A is nested in Model B, Model B will show equal or better fit to data than will Model A, all else being equal (barring convergence problems). However, it is instructive to illustrate how differential structural complexity has the potential to lead to the selection of an inappropriate model from among a

[4]To many readers, the word *random* may suggest that data are extraneous, unpredictable, or unsystematic (Schafer & Graham, 2002). In this article, random data are those generated by a probabilistic rather than a deterministic process.

[5]It should be emphasized that sample size has no role in the data generation process. The **R**s produced by this algorithm are neither sample nor population matrices.

set of competing alternatives. In Example 1, two models differing only in functional form are compared in terms of FP. Example 2 explores the relative importance of functional form and the number of free parameters in determining FP. Finally, Example 3 presents models drawn from published research, and illustrates some consequences associated with incompletely considering differences in FP.

The relative FP of competing models was assessed by determining how well each model fits representative data relative to its competitors. Because frequent estimation problems (nonconvergence and improper solutions) were encountered with ML estimation in preliminary simulations, the OLS discrepancy function was chosen. Each model was fit to 10,000 random data matrices. I chose RMSR as a fit index because it incorporates no adjustment for q or functional form, thus allowing the relative FPs of alternative models to be illustrated by simply noting differences in fit with respect to the same data. Cumulative distribution functions (CDFs) of RMSR were plotted. Even in the absence of a benchmark criterion value for good fit, nonoverlapping CDFs are sufficient to illustrate that two or more models differ in terms of FP. For fitting models to data, RAMONA 4.0 for DOS (Browne & Mels, 1990) was chosen for its ability to quickly estimate many models sequentially. Sample size was set to 1,000 for all analyses.[6] Most models converged before reaching the iteration limit.[7]

Example 1: FP due to Functional Form

Consider the models in Figure 4 (Models 1A and 1B). Model 1A is an unrestricted simplex model, generally applied in situations in which a more or less band-diagonal correlation matrix[8] is expected. Model 1B is a contrived 1-factor model, with two loadings constrained to equality so that Models 1A and 1B will have the same number of free parameters (q = 11). Model 1B is expected to have greater FP than Model 1A because Model 1A is designed to account well only for **R**s conforming to a band-diagonal pattern, whereas Model 1B is designed to account well for a broader array of potential correlation patterns.

---

[6]Sample size has no effect on model fit results considered in this article. Sample size affects the precision of parameter estimates in OLS but does not alter the point estimates themselves.

[7]In Monte Carlo studies, one can include improper solutions, exclude them, or constrain solutions to be proper (Gerbing & Anderson, 1993). In this and all subsequent analyses, the maximum number of iterations was set to 2,000 to allow a reasonable amount of time for models to reach convergence. Although not all models converged, fit indices were always produced, indicating the best fit obtainable after 2,000 iterations.

[8]Band-diagonal correlation matrices are those in which every diagonal below the main diagonal consists of correlations homogeneous in magnitude, generally decreasing in magnitude with distance from the main diagonal. Matrices approximating this form are common in longitudinal studies characterized by autocorrelation.

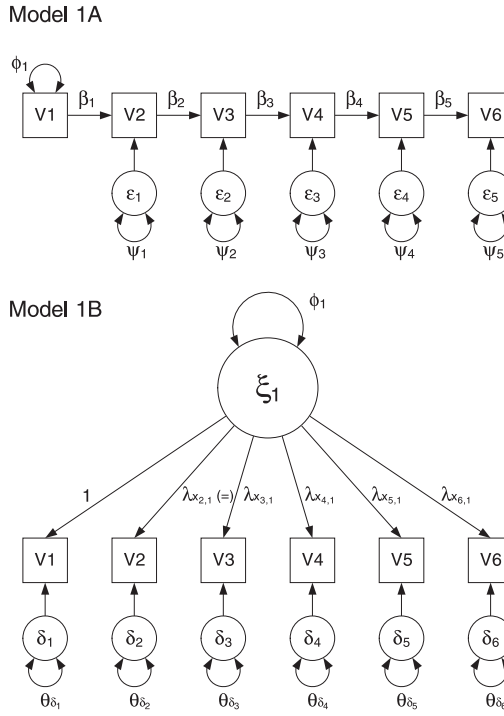


FIGURE 4 A simplex model (Panel A, Model 1A) and a factor model (Panel B, Model 1B).

Models 1A and 1B were each fit to 10,000 randomly generated 6 × 6 matrices, with the resulting RMSR CDFs depicted in Figure 5. The most noteworthy feature of Figure 5 is the distance between the CDFs, which can be interpreted as relative differences in FP. Because the two models were fit to the same random data using the same number of free parameters, the substantial disparity in FP between Models 1A and 1B can be attributed to differences in functional form. The simplex model fails to fit large correlations between distally connected variables, yet fits correlations between adjacent variables very well. The factor model, on the other hand, can fit many more correlation patterns. Thus, even when the same q is involved, models differing in functional form can be quite different in terms of FP.

For researchers interested in evaluating the fit of each competing model against external benchmarks, it could be of interest to explore the regions of the data space fit well by each model relative to a fixed criterion of good fit. Hu and Bentler (1999) recommend that a criterion close to .08 be chosen when only RMSR is used, so .08 was chosen as a convenient (but arbitrary) benchmark for good fit. Table 1 shows the number of random data sets, out of 10,000, fit well and poorly by

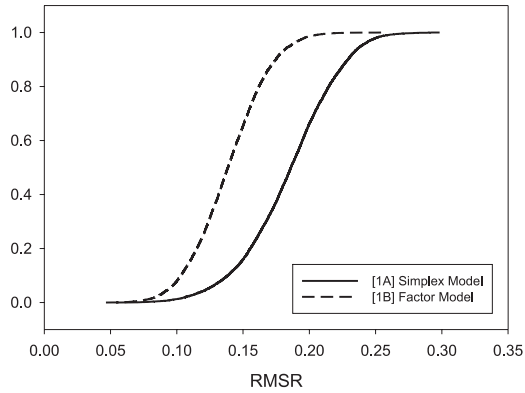


FIGURE 5 Cumulative frequency distributions of RMSR for a simplex model and factor model fit to the same 10,000 random correlation matrices.

the two models jointly and in isolation. Even though Model 1B fit 136 data sets well by the .08 criterion, Model 1A fit 23 data sets well that Model 1B fit poorly. Similarly, even though Model 1A fit 27 data sets well, Model 1B fit 132 data sets well that Model 1A fit poorly. Thus, the competing models can fit data from different parts of the data space differentially well.

This example shows one way in which two models differing only in functional form can have different FPs. There are other ways in which two models may differ in structural complexity. For example, alternative models may imply different numbers of zero correlations. The greater the number of *implied zeroes*, the less FP a given model will have because the model likely will be notably misspecified for that subgroup of correlations (Example 2 contains a clear illustration of this effect). Another situation in which models may differ in structural complexity occurs with the imposition of certain constraints. For example, equality constraints represent the requirement that two or more otherwise free parameters must be equal to each other. It is also possible for two models to dif-

TABLE 1
Frequencies of Data Sets Fit Well and Poorly
by Models 1A (Simplex) and 1B (Factor)

	Model 1B	
Model 1A	*RMSR < .08*	*RMSR ≥ .08*
RMSR < .08	4	23
RMSR ≥ .08	132	9,841

Note. RMSR = root mean squared residual.

fer only in that one has an equality constraint where the other has a fixed-value constraint. The two models would have the same q but different functional forms, and thus potentially different FPs.

Example 2: The Relative Importance of Free Parameters and Functional Form

It could be that the impact of functional form on FP is negligible relative to the influence of the number of free parameters. If that is true, then adjusting fit for q may be a sufficient penalty for FP, and further consideration of structural complexity would supply little additional information. To illustrate that functional form potentially can determine a model's FP to a greater extent than can q alone, two models were specified. Model 2A is a simple structural equation model in which the unique variances of indicators for particular latent variables have been constrained to equality (see Figure 6). Model 2B is a simple confirmatory factor model with the factor correlation constrained to zero. Note that Model 2B has more free parameters ($q = 12$) than Model 2A ($q = 9$) and thus would ordinarily be expected to have greater FP. However, the functional form of Model 2B is such that there will be nine implied zero correlations, which should seriously curtail Model 2B's ability to fit data.

In fact, as can be seen from the RMSR CDFs in Figure 7, the model with more free parameters (2B) fit substantially worse than the model with fewer free parameters (2A). Neither model fit particularly well overall by standard criteria, but the large difference in the proportion of data patterns fit by the two models at any value for RMSR is telling. The number of free parameters is not always the most important factor in determining FP.

Example 3: Model Selection in a Real Example

Keyes, Shmotkin, and Ryff (2002) investigated the relationship between the constructs *subjective well-being* (SWB) and *psychological well-being* (PWB) by conducting a series of confirmatory factor analyses. Indicators of SWB included an item representing global life satisfaction and scales assessing positive and negative affect. PWB was represented by short forms of Ryff's (1989) six scales of PWB (Self-Acceptance, Environmental Mastery, Positive Relations With Others, Personal Growth, Purpose in Life, and Autonomy). Because substantive concerns are beyond the purview of this investigation, only those details germane to FP and model selection are presented here.

The authors tested a series of six confirmatory models to ascertain the relationship between SWB and PWB. These included (1) an independence model (omitted here), (2) a one-factor model, (3) a two-factor model with uncorrelated factors,

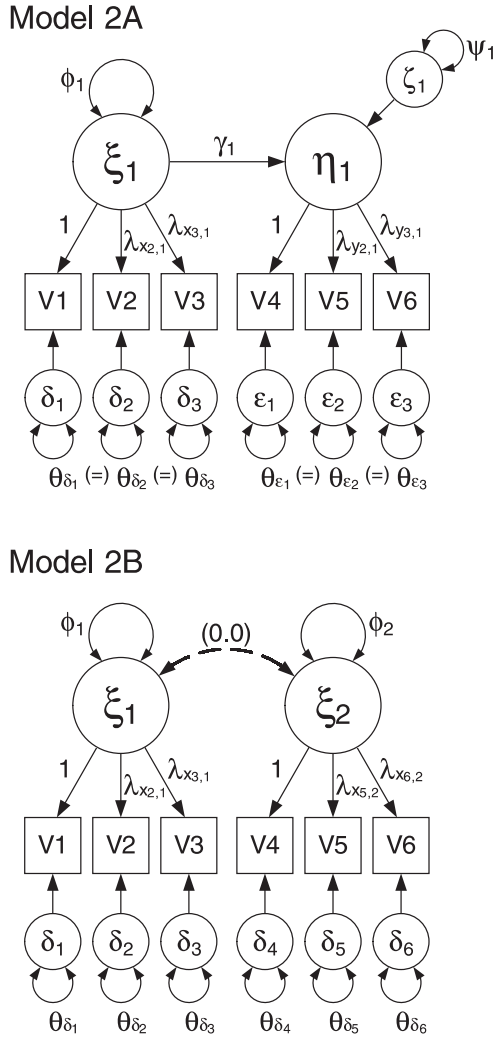


FIGURE 6 Two factor models with different functional forms. The model in Panel A (Model 2A) has 9 free parameters. The model in Panel B (Model 2B) has 12 free parameters.

(4, 5) two models with two correlated factors, and (6) a model with two correlated factors and two extra loadings. These models, with the exception of the independence model, are depicted in Figure 8. According to the likelihood ratio test, Model 6 was the best-fitting model, followed in order by Model 4, Model 5, Model 2, and Model 3. On the basis of superior fit indices, the authors chose Model 6 as the best model, acknowledging that Model 4 also fit well.

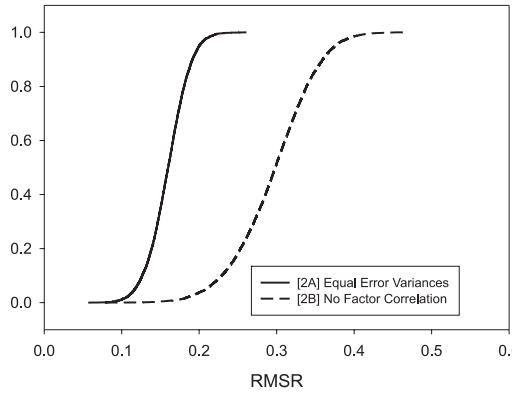


FIGURE 7 Cumulative frequency distributions of root mean squared residual for the two factor models in Figure 6.

Using OLS estimation, the five models in Figure 8 were also each fit to 10,000 random **R**s to assess their relative FPs. Model 6 was the best-fitting model when fit to random data, followed by Model 5, Model 4, Model 2, and Model 3 in that order, although there was virtually no separation between Models 4 and 5. In other words, the rank of the models in terms of fit to real data closely corresponds to the order expected simply by knowing the models' relative FPs. The implication is that little knowledge about the preference that should be assigned to each of the competing models based on observed data was gained above and beyond the models' antecedent propensities to fit random data.

Cumulative plots of RMSR for all five models are shown in Figure 9; some features bear close examination. First, note the wide separation between Model 3 (uncorrelated factors) and the other models. Model 3 fits much worse than, for example, Model 2 even though the two models have the same number of free parameters. The two models differ only in functional form; because the factor correlation in Model 3 is constrained to zero, Model 3 implies 18 zero correlations, whereas Model 2 implies none. Specifically, Model 3 permits no correlations between indicators loading on PWB and SWB, so the model will be misspecified for any nonzero correlations between indicators of the two factors in the population. Second, Models 4 and 5 also differ only in functional form, yet the two models have virtually identical FPs (mean RMSRs for Models 4 and 5 were, respectively, .1385 and .1384). Third, despite the addition of two extra free parameters, Model 6 fit little better than Models 2, 4, and 5. The primary lesson to be learned here is that, whereas *q* certainly influences FP, every free parameter does not contribute equally to a model's ability to fit data. The way in which free parameters are combined in a model (i.e., functional form) can have a large impact.

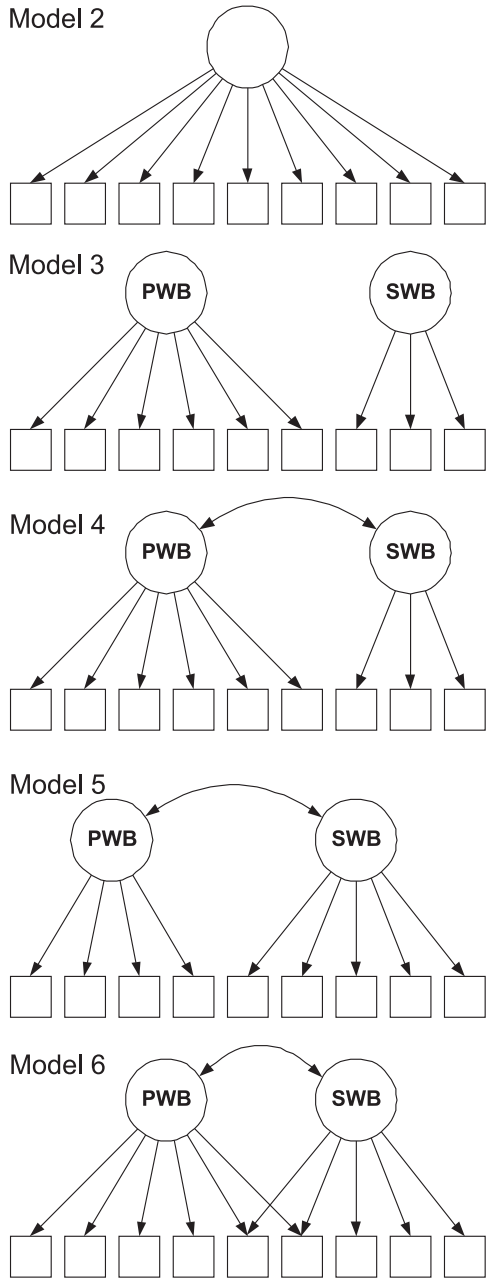


FIGURE 8 Five competing, theory-implied factor models compared by Keyes et al. (2002).

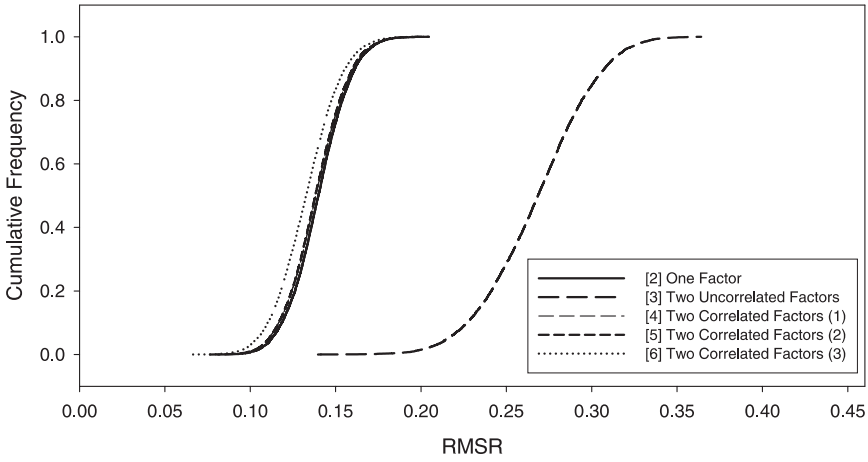


FIGURE 9 Cumulative frequency distributions of RMSR for the five competing factor models compared by Keyes et al. (2002).

Summary

The findings reported here are intended only to illustrate how model selection in SEM might be influenced by quantifying FP in terms of the rival models' abilities to fit random data rather than in terms of the number of parameters. Nowhere do I claim that the conclusions drawn by Keyes et al. (2002) are incorrect. Models are typically not evaluated solely on the basis of fit—other factors also determine judgments of a model's quality, such as the interpretability of parameter estimates and the theoretical plausibility of the entire model. However, it should be emphasized that the way in which FP is quantified can have an impact on substantive conclusions. Fitting competing models to the data space provides additional information that can augment traditional fit indices and selection criteria.

DISCUSSION

Models with different functional forms but the same number of free parameters can, and usually do, have different FPs. However, in fit indices and selection criteria traditionally used in SEM, only q is considered; functional form typically is not. By way of example, I illustrated some consequences associated with relying only on traditional methods of considering FP. The traditional conceptualization of parsimony in SEM is guided by the implicit assumption that parsimony is a linear decreasing function of q or, conversely, that FP is a linear increasing function of q. But FP, as defined here and elsewhere, is a complicated function of q,

functional form, and other factors, although the treatment presented here focused only on q and functional form. Ideally, consideration of FP should incorporate these factors in addition to q in order to quantify falsifiability as accurately as possible. Current methods of model evaluation and selection in SEM do not adequately achieve that goal.

Implications and Recommendations for Practice and Future Research

Ideally, in the context of model selection, researchers would hope to encounter models with equal goodness of fit but different FPs or models with equal FPs but different fit. In the former situation, parsimony would be the obvious deciding factor (Forster & Sober, 1994; Quine, 1966); in the latter, the better fitting model would emerge as the winner (Dunn, 2000; Turney, 1990). Unfortunately, we are rarely presented with so convenient a situation—the models being compared usually vary along both dimensions. Most SEM fit indices adjust for FP in a way that is only partially consistent with the actual level of FP. Practical methods are therefore needed for the direct evaluation of FP in substantive studies. I offer two possibilities for consideration.

Uniform index-of-fit. Botha et al. (1988) developed a method for estimating the appropriate number of factors (m) in exploratory factor analysis when the statistical determination of m is inappropriate, such as when the entire population is available for examination. They defined a distance measure based on the OLS discrepancy function:

$$d_{p,m}(\mathbf{R}) = \{\min_{\Lambda,\Psi} tr(\mathbf{R} - \Lambda\Lambda' - \Psi)^2\}^{1/2} \tag{2}$$

where $\mathbf{R}$ is a correlation matrix, Λ is the estimated factor loading matrix, and Ψ is the matrix of unique factor variances. The distance $d_{p,m}(\mathbf{R})$ reflects the square root of the sum of squared residual correlations. Botha et al. generated 1,000 random $\mathbf{R}$s, fit a series of candidate factor models to every generated matrix, and calculated $d_{p,m}(\mathbf{R})$ for each $\mathbf{R}$. Then, for a given empirical correlation matrix $\mathbf{R}^*$, $d_{p,m}(\mathbf{R}^*)$ is obtained, as well as the probability that $d_{p,m}(\mathbf{R})$ is greater than or equal to $d_{p,m}(\mathbf{R}^*)$:

$$i_{p,m}(\mathbf{R}^*) = \Pr \mid d_{p,m}(\mathbf{R}) \geq d_{p,m}(\mathbf{R}^*) \mid \tag{3}$$

In other words, a given data set $\mathbf{R}^*$ is assigned a percentile based on where its fit index falls in the CDF of fit statistics. This quantity Botha et al. termed the uniform index-of-fit. The UIF reflects the rank of the observed fit relative to the model's fit to all randomly generated data. A small value of $i_{p,m}(\mathbf{R}^*)$ implies that a given factor model is inappropriate for the data, because the m-factor model fits a large propor-

tion of random data sets better than the observed data. Values close to 1.0, however, indicate that the *m*-factor solution might be appropriate.

Botha et al. (1988) did not explicitly consider FP, but their procedure is essentially a model selection routine that considers all factors contributing to FP. The UIF assigns a handicap to each factor model to a degree commensurate with its ability to fit diverse (random) data patterns, placing all competing models on the same footing with regard to FP before they are compared in terms of fit.[9] The index $i_{p,m}(\mathbf{R}^*)$ is defined in terms of least-squares discrepancy, but $\widehat{F}_{ML}$ may be used just as easily as $\widehat{F}_{OLS}$ for purposes of establishing $i_{p,m}(\mathbf{R}^*)$.

The graphical approach to considering FP described earlier clearly owes much to Botha et al.'s (1988) UIF, in the sense that **R**s representative of the relevant data space are generated and the models of interest are fit to those random data. In theory, there is nothing about UIF that should prevent it from being generally applicable in SEM. In practice, there are two challenges to its implementation. First, prohibitively large amounts of random data must be generated and fit by the set of rival models to establish the distribution of $d_{p,m}(\mathbf{R})$ values, which would in turn permit the researcher to draw fine distinctions between models with very small $i_{p,m}(\mathbf{R}^*)$ values. Development of faster and more efficient methods of generating **R**s or reasonable methods for restricting the bounds of the data space may solve this challenge. Second, most applications involve sample data. It would be useful to construct confidence intervals around each observed $i_{p,m}(\mathbf{R}^*)$ to reflect the uncertainty in model ranking due to sampling variability. More research concerning the statistical properties of UIF, and its potential usefulness as a model selection criterion in SEM, is warranted.

Minimum description length. Another promising alternative to traditional fit indices, MDL, has been developed in recent years in the information theoretic literature. MDL is a formalization of Kolmogorov complexity (Grünwald, 2000; Rissanen, 1996), the shortest code length necessary to fully represent a data sequence in a given encoding language. Stochastic complexity, which is used in the practical application of MDL, is analogous to Kolmogorov complexity but uses a class of models rather than an encoding language as a basis for expressing FP. The MDL principle reexpresses the generalizabilities of rival models as their relative abilities to "compress"—or parsimoniously describe—observed data (Myung, Navarro, & Pitt, 2006). The MDL criterion has been successfully used in many model selection situations. For example, Myung et al. (2004) use an MDL criterion to compare five models of category learning. Myung, Balasubramanian, and Pitt (2000) and Pitt et al. (2002) showed that MDL outperformed other selection criteria in distinguishing between data generated by two psychophysical laws differing

[9]The UIF can be applied only under certain assumptions. Those assumptions are (a) the data space is bounded and (b) the entire domain of plausible data has been sampled.

only in functional form. Pitt et al. and Su, Myung, and Pitt (2005) show how MDL can be used to investigate the generalizability of cognitive categorization models and two information integration models. Several other applications of MDL are reported by Hansen and Yu (2001).

One expression for MDL is termed the *normalized maximum likelihood* (NML):

$$\text{NML} = \frac{f(x_{obs};\hat{\theta}(x_{obs}))}{\int_x f(x;\hat{\theta}(x))dx}, \tag{4}$$

where $f(\cdot)$ is the maximum likelihood function. NML frames the FP component of MDL (i.e., the integral in the denominator of Equation 4) as a normalized sum of all maximum likelihood best fits (Rissanen, 2001b; the mean can be used instead of the sum with no loss of generality). Preferable models are characterized by relatively higher values of NML. The key advantage associated with NML over traditional selection criteria is that NML implicitly considers both parametric complexity and structural complexity components of FP.

The relationship between the NML criterion and the graphical method presented in this article is worth noting. In numerical computation of NML, best-fitting ML values are (implicitly) plotted against data for all possible data patterns and numerical integration is used to find the area under the resulting surface. If a researcher wished to simply rank several models in increasing order of FP, finding the area is unnecessary; all that is required is the expected (mean) likelihood for each model. The graphical method developed in this study using CDFs does exactly that, and it can be viewed as an empirical analog to finding the expected best fit using OLS rather than ML. Using ML, the mean minimum-fit likelihood will be an estimate of the FP component of NML. It is likely that a rank ordering of models using NML and another using mean RMSR would be similar or identical in most modeling situations, although NML may be preferable in that it has a formal grounding in information theory which the OLS approach lacks.

A more intuitive expression of the MDL principle is the Fisher information approximation (FIA) expression (Rissanen, 1996):

$$\text{FIA} = -\ln f(x_{obs};\hat{\theta}(x_{obs})) + \tfrac{1}{2}q\ln(N/2\pi) + \ln\int_\Theta \sqrt{|I(\theta)|}d\theta, \tag{5}$$

where $|I(\theta)|$ is the determinant of the Fisher information matrix of the parameters in θ. FIA is an approximation to the negative logarithm of NML (preferable models are characterized by relatively smaller values of FIA), and thus FIA is expected to rank models in the same order as NML. Note that the integration in Equation 5 is taken over the parameter space rather than the data space, as in Equation 4. FIA ex-

plicitly separates badness of fit (the first term), parametric complexity (the second term), and structural complexity (the third term).

In the SEM context, it is significantly more difficult to obtain the integral used in FIA than that used in NML. However, both integrals can be quite difficult to obtain even numerically. In some modeling contexts NML can be computed directly, with no need for data simulation (e.g., Hansen & Yu, 2001; Myung, Balasubramanian, & Pitt, 2000; Su et al., 2005). Structural equation models, however, are typically highly parameterized, making direct analytic computation of NML very difficult or intractable. Until a good analytic approximation can be identified, calculation of an MDL index in the SEM context involves fitting a model to a large number of random data sets, in which case FP is operationalized as the mean obtained likelihood. Fortunately, a tractable approximation to FIA exists in the stochastic information complexity (SIC) criterion (Hansen & Yu, 2001; Markon & Krueger, 2004; Rissanen, 1989):

$$\text{SIC} = -\ln f(x_{obs};\hat{\theta}(x_{obs})) + \tfrac{1}{2}\ln |\, N \cdot I(\hat{\theta})\,| \tag{6}$$

No integration is necessary to compute SIC, although the researcher does need access to the information matrix computed internally by most SEM software applications.

Both the FIA and SIC expressions permit insight into how functional form can influence FP. As the redundancy (correlation) among parameters increases, $|I(\theta)|$ decreases. Consequently, models with relatively independent parameters will tend to be less parsimonious, whereas models possessing parameters with overlapping roles will tend to possess less FP, all things being equal. The FIA expression additionally illustrates an interesting relationship between the MDL criterion and Bayesian model selection. Asymptotically, FIA can be expressed as a rescaled Bayesian information criterion (BIC) plus a term whose importance diminishes with increased sample size, showing that BIC and MDL produce essentially the same ranking of models at extremely large sample sizes (Myung et al., 2006).

Facilitating the use of NML in SEM software would permit routine comparison of any number of models, nested or nonnested (Rissanen, 2001a), or evaluation and comparison of nonlinear models (Myung & Pitt, 2004). In the meantime, researchers can follow the general procedures outlined in this article—fitting models to large numbers of random correlation matrices and examining CDFs of fit indices. Alternatively (or in addition), NML can be computed using the mean likelihood obtained after fitting models to large quantities of random data, or approximated using the SIC expression in Equation 6.

Defining the Data Space

Defining the data space for computation of UIF and NML is not straightforward (Meehl, 1990). Nevertheless, some general guidelines can be offered. First, the

data space must be limited to those data patterns that are possible. In the SEM context, the possible data space might consist of all **R**s because no theory predicts data resulting in a nonpositive definite dispersion matrix. Second, the data space can be further refined or limited to those considered *plausible* (Roberts & Pashler, 2000), given knowledge of the population to which the researcher intends to make inferences or to generalize results. If two variables are experimentally manipulated so that their correlation will always be very close to zero, for example, such a constraint can be built into the data generation procedure. In the present context, plausible data consisted of positive-manifold **R**s for the sake of illustration, but this constraint could be further refined or relaxed as more knowledge is gained about the kinds of data likely to occur in a given milieu.

A second question concerns how one should sample from the data space once bounds have been established. I agree with other researchers (Dunn, 2000; Myung et al., 2006; Myung & Pitt, 1998; Rissanen, 1989, 2001b) that the fairest way to level the playing field for model selection is to sample from the data space uniformly, with no recourse to experience. This is equivalent to applying what is sometimes call the *principle of indifference*. Another possibility is to favor some regions of the data space over others, without completely ruling out any one portion. In light of these conflicting views, it is not immediately obvious what should constitute the appropriate data space in the SEM context, but judgments about relative FP may depend heavily on this choice. For example, Botha et al. (1988) used three methods of generating random correlations and noted that UIF values were sensitive to how the data were generated. Some auxiliary theories about how the world works will necessarily be involved in these decisions, and these auxiliary theories can be just as fallible as the theories under study. Future work on this problem is clearly warranted.

Other Factors Contributing to FP

Any factor influencing the antecedent probability that a model will fit well should contribute to FP. Besides the number of free parameters and the specific functional form associated with a model, there are at least four other factors contributing to FP. These factors, which include restriction of parameter range, the probability distribution specified in the likelihood function, sample size, and some features of the research design, are described in more detail by Pitt et al. (2002).[10] Here I focused

[10]Sample size moderates the impacts of other factors in determining FP. As sample size approaches infinity, the number of free parameters in large part determines a model's FP. However, at small sample sizes, other factors (e.g., functional form) also have an impact. The role of N can also be understood in terms of geometric complexity (see Pitt, Myung, & Zhang, 2002; Rissanen, 1996). The influence of parameter range on FP can be seen explicitly in the FIC expression of MDL, where this range determines the limits of integration in $\int_{\Theta} \sqrt{|I(\theta)|}\, d\theta$.

primarily on illustrating the roles of *q* and functional form in determining FP; the impact of these other factors deserves more attention.

Some may wonder what specific aspects of functional form may contribute to increased or decreased relative FP. There are a few clues that can be used to informally gauge a model's relative FP. The number of free parameters is often the best clue to relative FP. All else being equal, having more free parameters yields a model with greater FP.

Second, it is important to note that every structural equation model is a system of simultaneous equations for observed data. The more equations in which a given free parameter appears, the more important that parameter is likely to be in permitting the model to fit observed data. If the researcher can construct a model in which all rival models are parametrically nested and then identify what parameters in that model are constrained to produce each rival model, some insight may be gained into the relative FP of each rival model. This idea, incidentally, explains the importance of implied zeroes; exogenous covariances tend to appear in more equations than do other parameters. For example, both Models A and B in the introductory example can be considered parametrically nested in a hypothetical Model C that contains both parameters β_2 and β_3. Model A is obtained by constraining β_3 to zero. Model B is obtained by constraining β_2 to zero. In our hypothetical Model C, β_3 appears in three simultaneous equations, whereas β_2 appears in only two. Model A may therefore be more parsimonious because a parameter with relatively greater influence (β_3) was constrained to obtain it.

In general, the more ways there are to trace connections between pairs of variables, the more FP a model is likely to possess. Clearly the number of free parameters contributes to the number of ways in which pairs of variables are linked, but it is important to consider *which* parameters are free. Freeing some otherwise fixed parameters will achieve greater FP than will freeing other parameters, allowing a model to fit a more diverse array of data patterns than a model with the same number of free parameters but with a different pattern of constraints. Because the additional FP granted by freeing a parameter depends on which parameter is freed, FP is an important factor to consider even in the comparison of nested models. Ultimately, however, it is the data-fitting capacity of each model that determines its FP, so the most direct (if not the most practical) method of determining relative FP is to fit the competing models to large amounts of random data.

Can a Model Have Too Much FP?

A natural question to ask at this point is whether a given model can have too much FP. There are no clear benchmarks for making a judgment like this. Regardless of its FP, a model's purpose is to represent theoretical predictions, and theory may be vague or specific. It is unclear what it would mean for a model to have too much FP if it accurately reflects a complex theory. Fortunately, we need no such

benchmarks in the context of model selection. All competing models may legitimately possess high FP or high parsimony; what matters is each model's generalizability relative to rival models.

Sampling Variability and FP

An issue alluded to but not dwelt on is that of sampling variability. In addition to addressing the importance of a model's ability to constrain possible outcomes, Roberts and Pashler (2000) also addressed how inattention to sampling variability may hinder the process of model evaluation. When *N* is small, adjusted fit indices—and thus the ranking of models in terms of generalizability—may show considerable variability across repeated sampling. This uncertainty in ranking clearly poses a problem for model selection. When *N* is large, however, it is reasonable to expect rankings to remain relatively stable, permitting greater confidence in model selection. A possible solution to this problem is to derive confidence intervals for adjusted fit indices (e.g., UIF or MDL) to reflect uncertainty due to sampling.

Limitations

Although I believe the inferences drawn from the example illustrations to be valid and generally applicable, there were some features of the strategy that might limit their generality.

Random dispersion matrices. I generated correlation matrices in which all elements were nonnegative. This restriction is acknowledged to be unrealistic in some situations and may have consequences in situations where negative correlations are a real possibility. In addition, correlation matrices serve well enough for illustrative purposes, but in practice many situations will require random covariance matrices. Covariance matrices retain information about scale; any models not invariant to changes in scale (e.g., latent growth curve models and models simultaneously applied to multiple samples) should be fit to covariance matrices, not correlation matrices (Cudeck, 1989). Future research should be devoted to the generation of random covariance matrices so that relative FP may be accurately quantified for such models.

Focus on OLS estimation. I have emphasized OLS estimation to the exclusion of other common estimation methods. OLS was chosen because it proved to be more computationally robust than ML in preliminary simulations, is less prone to breaking down when confronted with near-singular correlation matrices (Browne, MacCallum, Kim, Andersen, & Glaser, 2002), and may be more appropriate than ML in many modeling contexts, given assumptions about the nature of error imposed by ML (Briggs & MacCallum, 2003). I acknowledge that ML esti-

mation is used far more often than OLS in practice, but there is no reason to believe that the pattern of results would be substantially altered by use of different discrepancy functions.

Nonconvergence and improper solutions. For the demonstrations in this article, all solutions were retained for the construction of CDF plots, regardless of whether the minimization procedure fully converged. The frequency of nonconvergent solutions was quite small or zero for most models (.17% for Model 2B, .10% for Keyes et al.'s, 2002, Model 6, and 0% for all other models), and CDFs of RMSR looked highly similar regardless of whether nonconvergent solutions were included. When the minimization procedure failed to converge, estimation was ceased after a uniformly large number of iterations (2,000) and the value of the discrepancy function at the final iteration was used to compute RMSR. In practice, it is likely that there are many models for which the number of nonconvergent solutions becomes excessive. It is recommended that all solutions, convergent or nonconvergent, should be used because (a) the data could conceivably be obtained in practice and (b) the obtained discrepancy still has a legitimate interpretation.

As with nonconvergence, models fit to random data can be expected to occasionally result in parameter estimates lying outside logical bounds (e.g., variances less than zero or covariances implying absolute correlations greater than 1.0). Some research suggests that inadmissibility may pose a problem for interpretation of parameter estimates, but makes little difference in terms of model fit (Boomsma, 1985; Ding, Velicer, & Harlow, 1995; Gerbing & Anderson, 1987). All solutions in this study contained parameter estimates inside the permissible parameter space because RAMONA places boundary constraints on all parameters by default. Different software packages handle nonconvergence and improper solutions differently. Future research could be fruitfully devoted to investigating the consequences of dealing with nonconvergence and improper solutions in various ways for the evaluation of FP in SEM.

CONCLUSION

The primary motivation behind this article has been to introduce a more thorough understanding of FP to research psychologists using SEM. I emphasize that this work is intended to represent only a first step. It would be overly ambitious to both introduce the problems presented by FP and attempt to fully resolve all of them, within one article. I hope this exploration will provide significant steps toward the development of a new outlook on model evaluation and selection in SEM. One general strategy is to compute separate indices of FP and fit (unadjusted for FP) and to consider these two dimensions of generalizability separately. Alternatively, it may be more beneficial to combine measures of FP with measures of fit, as is

done in information theoretic criteria such as FIA. Whether FP should be evaluated separately from goodness of fit or incorporated as an adjustment factor is a question for future research to decide.

A secondary purpose of this article was to encourage the adoption of a model selection approach to conducting SEM. I advocate this approach for two reasons. First, it is highly unlikely that any model devised by mortal scientists is completely correct in all of its particulars, so the best we can expect from a model is that it serve as a useful approximation to the data-generating process (MacCallum, 2003). In practical terms, this idea translates to identifying the model that shows the best generalizability. Second, evaluation of models in isolation is prone to confirmation bias. One way to circumvent this bias is to pit competing explanations against one another. Model selection, the practice of evaluating theory-implied models relative to one another rather than to a fixed criterion, is consistent with the fundamental scientific pursuit of strong inference (Platt, 1964) and is to be strongly encouraged. This emphasis on model selection as more relevant to the aim of science than model evaluation is consistent with current thought in the philosophy of science (Lakatos, 1970; Meehl, 1990).

The good fit of a hypothesized model to observed data, although desirable, can result from the model's inherent ability to predict data patterns and may have little to do with its value as a scientific tool. Cherished models may have to be abandoned or replaced if their past successes can be ascribed more to FP than to any insight they lend into the process that actually generated the data. Adopting a model selection perspective and explicitly considering FP can help researchers avoid these problems.

ACKNOWLEDGMENTS

This article is based on Kristopher Preacher's dissertation, completed at Ohio State University.

I thank Michael W. Browne, Cheongtag Kim, Woojae Kim, Robert C. MacCallum, Jay I. Myung, Daniel J. Navarro, James H. Steiger, several anonymous reviewers, and the Carolina Structural Equation Modeling group for many fruitful comments and discussions culminating in this work.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (p. 267). Budapest, Hungary: Akademiai Kiado.

Bamber, D., & van Santen, J. P. H. (2000). How to assess a model's testability and identifiability. *Journal of Mathematical Psychology*, *44*, 20–40.

Beichl, I., & Sullivan, F. (2000). The metropolis algorithm. *Computing in Science & Engineering*, *2*(1), 65–69.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, *31*, 419–456.

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

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika*, *50*, 229–242.

Botha, J. D., Shapiro, A., & Steiger, J. H. (1988). Uniform indices-of-fit for factor analysis models. *Multivariate Behavioral Research*, *23*, 443–450.

Briggs, N. E., & MacCallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research*, *38*, 25–56.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research*, *24*, 445–455.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research*, *21*, 230–258.

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

Browne, M. W., & Mels, G. (1990). *RAMONA user's guide* [Computer software manual]. Unpublished report, Ohio State University, Columbus.

Collyer, C. E. (1985). Comparing strong and weak models by fitting them to computer-generated data. *Perception and Psychophysics*, *38*, 476–481.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin*, *105*, 317–327.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research*, *18*, 147–167.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin*, *109*, 512–519.

Cutting, J. E., Bruno, N., Brady, N. P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General*, *121*, 364–381.

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

Dunn, J. C. (2000). Model complexity: The fit to random data reconsidered. *Psychological Research*, *63*, 174–182.

Forster, M. R., & Sober, E. (1994). How to tell when simpler, more unified, or less *ad hoc* theories will provide more accurate predictions. *British Journal for the Philosophy of Science*, *45*, 1–35.

Gerbing, D. W., & Anderson, J. C. (1987). Improper solutions in the analysis of covariance structures: Their interpretability and a comparison of alternate respecifications. *Psychometrika*, *52*, 99–111.

Gerbing, D. W., & Anderson, J. C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 40–65). Newbury Park, CA: Sage.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, *44*, 133–152.

Hansen, M. H., & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, *96*(454), 746–774.

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.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data*. Beverly Hills, CA: Sage.

Jeffreys, H. (1931). *Scientific inference*. Cambridge, England: Cambridge University Press.

Jeffreys, H. (1957). *Scientific inference* (2nd ed.). Cambridge, England: Cambridge University Press.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8 user's reference guide*. Uppsala, Sweden: Scientific Software International.

Keuzenkamp, H. A., & McAleer, M. (1997). The complexity of simplicity. *Mathematics and Computers in Simulation, 43,* 553–561.

Keyes, C. L. M., Shmotkin, D., & Ryff, C. D. (2002). Optimizing well-being: The empirical encounter of two traditions. *Journal of Personality and Social Psychology, 82,* 1007–1022.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). Cambridge, England: Cambridge University Press.

Linhart, H., & Zucchini, W. (1986). *Model selection*. New York: Wiley.

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

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin, 114,* 185–199.

Markon, K. E., & Krueger, R. F. (2004). An empirical comparison of information-theoretic selection criteria for multivariate behavior genetic models. *Behavior Genetics, 34,* 593–610.

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

Marsh, H. W., & Hau, K.-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *The Journal of Experimental Education, 64,* 364–380.

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

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1,* 108–141.

Mulaik, S. A. (2001). The curve-fitting problem: An objectivist view. *Philosophy of Science, 68,* 218–241.

Mulaik, S. A. (2004). Objectivity in science and structural equation modeling. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 422–446). Thousand Oaks, CA: Sage.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105,* 430–445.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44,* 190–204.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences, 97,* 11170–11175.

Myung, I. J., Forster, M. R., & Browne, M. W. (Eds.). (2000). Model selection [Special issue]. *Journal of Mathematical Psychology, 44*(1).

Myung, J. I., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology, 50,* 167–179.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4,* 79–95.

Myung, I. J., & Pitt, M. A. (1998). Issues in selecting mathematical models of cognition. In J. Grainger & A. M. Jacobs (eds.), *Localist connectionist approaches to human cognition* (pp. 327–355). Mahwah, NJ: Lawrence Erlbaum Associates.

Myung, I. J., & Pitt, M. A. (2004). Model comparison methods. *Methods in Enzymology, 383,* 351–366.

Myung, I. J., Pitt, M. A., & Kim, W. (2004). Model evaluation, testing and selection. In K. Lambert & R. Goldstone (Eds.), *Handbook of cognition* (pp. 422–436). Thousand Oaks, CA: Sage.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6,* 421–425.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109,* 472–491.

Platt, J. R. (1964, October 16). Strong inference. *Science, 146*(3642), 347–353.

Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.

Preacher, K. J. (2003). *The role of model complexity in the evaluation of structural equation models.* Unpublished doctoral dissertation, Ohio State University, Columbus.

Quine, W. V. (1966). *The ways of paradox and other essays*. New York: Random House.

Raykov, T., & Marcoulides, G. A. (1999). On desirability of parsimony in structural equation model selection. *Structural Equation Modeling, 6,* 292–300.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Singapore: World Scientific.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory, IT-42*(1), 40–47.

Rissanen, J. (2001a). Simplicity and statistical inference. In A. Zellner, H. A. Keuzenkamp, & M. McAleer (Eds.), *Simplicity, inference, and modelling* (pp. 156–164). Cambridge, England: Cambridge University Press.

Rissanen, J. (2001b). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory, 47,* 1712–1717.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107,* 358–367.

Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology, 57,* 1069–1081.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

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

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

Su, Y., Myung, I. J., & Pitt, M. A. (2005). Minimum description length and cognitive modeling. In P. Grünwald, I. J. Myung, & M. A. Pitt (Eds.), *Advances in minimum description length: Theory and applications* (pp. 411–433). Cambridge, MA: MIT Press.

Thurstone, L. L. (1935). *The vectors of mind*. Chicago: University of Chicago Press.

Thurstone, L. L. (1947). *Multiple-factor analysis: A development and expansion of The Vectors of Mind*. Chicago: The University of Chicago Press.

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

Turney, P. (1990). The curve fitting problem: A solution. *The British Journal for the Philosophy of Science, 41,* 509–530.

Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 42,* 369–390.

APPENDIX
The MCMC Algorithm

The MCMC algorithm begins by choosing a starting matrix which conforms to the definition of **R**. For convenience, this starting matrix was defined as a square, symmetric matrix with 1s on the diagonal and .5s elsewhere. The only other requirement for a successful starting matrix is that it be positive definite or nearly positive definite. Thus, starting matrices could be chosen to have .9999s, 0s, or empirically derived correlations in every subdiagonal element with no detrimental consequences. All of these possibilities were tested, all with successful results. At each iteration, the algorithm takes the current matrix and perturbs its off-diagonal elements, retaining symmetry. The resulting matrix is checked for conformity to the definition of **R**. If it does not conform, the matrix is discarded and another is generated; the previous matrix (which always conforms) is used again as a starting point. If the new matrix conforms to retention criteria, it is retained and is chosen as the starting point for the succeeding iteration, and so on. In the future, if matrices other than correlation matrices are desired, additional retention criteria related to variable scale will also be necessary.

Formally, the Metropolis-Hastings MCMC algorithm was used (Beichl & Sullivan, 2000; Gilks et al., 1996). Let the $\frac{1}{2}p(p-1) \times 1$ vector X_{t+1} contain the subdiagonal elements of a matrix (*state*) at time $t + 1$. State X_{t+1} is derived from the current state X_t by sampling from a *proposal distribution*, which in this case is:

$$x_{i_{(t+1)}} = x_{i_{(t)}} + j \cdot \left(\frac{u^{1/z}}{\|w\|} \cdot w_i \right) \tag{A1}$$

where $z = \frac{1}{2}p(p-1)$, $u \sim U(0,1)$, $w_i \sim N(0,1)$, $i = 1 \ldots z$, $\|w\|$ = length of w, and j = jump size (multiplier less than 1.0). This proposal distribution function in Equation A1 generates j-scaled points on the uniform hypersphere, but the proposal distribution can have virtually any form. The shape of the proposal distribution affects efficiency (the proportion of matrices meeting retention criteria) but leaves the target distribution unchanged. However, the form of the function, within reasonable limits, makes little difference in terms of the ultimate result. Jump size (j) is a constant which affects the degree to which X_{t+1} differs from X_t. At the extremes, $j = 0$ would result in the same matrix being produced at every step, and a j that is too large will yield results practically equivalent to those generated by the UCM method, with the same lack of speed. The latter choice would constitute an *independence sampler*, the special case of the Metropolis-Hastings algorithm in which the candidate distribution does not depend on previous X_ts.

After a sufficient number of iterations (*burn-in*), the distribution of retained matrices approximates the target distribution of correlation matrices. However,

because there is dependency between each matrix and all matrices preceding it (more dependency results from smaller jump sizes), it is not always desirable to retain every matrix. Instead, matrices are selected and retained at regular iteration intervals. This *thinning number* should be large to eliminate any measurable sequential dependency, but not so large as to compromise efficiency. In the current application, the thinning number is the same as the burn-in, specifically 50 both for the 6 × 6 matrices used in Examples 1 and 2 and for the 9 × 9 matrices used in Example 3. If the number of matrices retained is large, virtually any burn-in and thinning number will be adequate. As long as the acceptance rate is relatively high, the thinning number need not be large.

The portion of the multidimensional space through which the MCMC algorithm searches is restricted to the region most likely to yield correlation matrices. As matrix order increases, the ratio of the number of unacceptable matrices to acceptable matrices increases exponentially, accounting for the inefficiency of UCM at orders higher than 7 or so. The MCMC algorithm is still subject to decreasing speed and efficiency with increasing order, but to a far lesser degree than the method proposed by Botha et al. (1988). Full details regarding the MCMC data generation method employed can be found in Preacher (2003).