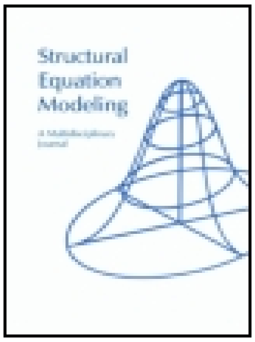

Routledge
Taylor & Francis Group

Structural Equation Modeling

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: https://www.tandfonline.com/loi/hsem20

Doing the Four-Step Right

Stanley A. Mulaik & Roger E. Millsap

STRUCTURAL EQUATION MODELING, *7*(1), 36–73

Doing the Four-Step Right

Stanley A. Mulaik
School of Psychology
Georgia Institute of Technology

Roger E. Millsap
Department of Psychology
Arizona State University

Our response to Hayduk and Glaser will principally focus on their critique of the four-step procedure. Hayduk and Glaser project things into the four-step procedure that are not part of its conception. They fail to see the implicit context in which those who use this particular four-step procedure operate, which qualifies its application. They also have misunderstandings about the rationale for the procedure and read too much of exploratory factor analysis into its use. Hayduk (1996) proposed a method for doing structural equation modeling with as few as one or two indicators per latent variable, which he feels is incompatible with the factor-analytic underpinnings of the four-step procedure and which motivates him further to seek its overthrow. He did not clarify this sufficiently on SEMNET, so we have been compelled to read Hayduk (1996) to better understand his position, and we will point out some limitations of it from our own point of view. We further argue that Hayduk's (1996) advocacy of the use of fixed parameters in sparse measurement models is not a viable general alternative to the use of multiple indicators. Hayduk also believed the usual .05 level of significance in testing the exact fit of models favors the null hypothesis. He recommended that the significance level for a chi-square test be set at .75. We show this recommendation to be incoherent with the idea of a significance test and further show it to be unnecessary because, on the contrary, in most studies the null hypothesis is likely to be rejected.

Requests for reprints should be sent to Stanley A. Mulaik, School of Psychology, Georgia Institute of Technology, Atlanta, GA 30332. E-mail: pscccsm@prism.gatech.edu

BACKGROUND FOR THE FOUR-STEP METHOD

The idea of testing nested sequences of models, which is the basis for a four-step approach to model evaluation that is the focus of Hayduk and Glaser's critique, goes back to a paper by Roy (1958), which tried to develop sequences of independent statistical tests of the constraints placed on parameters of multiparameter models in multivariate analysis. The model is regarded as specified by constraints placed on its parameters. In factor analysis and structural equation modeling, after Jöreskog's (1969) formulation of a method for doing confirmatory factor analysis, a parameter could be constrained to equal a specific value, to be equal to the value of other parameters, or to satisfy certain equalities and inequalities with other parameters. The idea of a nested sequence of models is to arrange the various constraints of a prespecified multiparameter model in a specific, usually natural, order, and then, beginning with a minimally constrained version of the model in which unconstrained parameters are estimated conditional on the constrained parameters, to follow this order in introducing successive constraints on the parameters in successive models. Each successive model is tested for its fit to the data. If a new constraint does not produce unacceptable fit, one can accept that constraint as provisionally appropriate together with any preceding accepted constraints introduced in preceding but less constrained versions of the model. One would then go on to introduce and test an additional constraint. But if a constraint or set of constraints, when introduced, produces unacceptable lack of fit, one can conclude that its retention in any succeeding, more constrained, versions of the model would continue to produce lack of fit and would be inappropriate in any version of the model. At that point the sequential testing of constraints is halted.

But another practice arose wherein constraints that produced unacceptable fit when introduced would be noted and then freed and be kept free in all subsequent versions of the model as one introduced additional sets of constraints on the remaining parameters. Each freed parameter produces a loss of one degree of freedom and represents a degradation in the objectivity of the final accepted model.

Objective validation of a hypothesis is provisionally established by showing that it is upheld in data not used in the formulation of the hypothesis (Mulaik 1990, 1995; Mulaik & James 1995). When data are used both to test and to estimate parameters, which is the usual practice, an aspect of the data is consumed in determining the parameter estimates, which completes an incompletely specified hypothesis and is no longer available for testing. The aspect of the data used to "determine" the parameter estimates is the reproduced data. (If one replaces the observed covariance matrix with the reproduced covariance matrix and uses the latter to estimate the parameters of the model, it will determine the same parameter values.) The residuals, representing the difference between the reproduced and the actual data, are not used in determining the estimates. The residuals can be used to evaluate the constraints of the model by assessing possible lack of fit. The dimensionality of the residual space is the degrees of freedom of the model. The degrees of freedom represent the number of independent conditions by which a model may be disconfirmed by lack of fit to the data (Mulaik, 1990).

Initially the idea of nested sequences of models was used in stepwise multiple regression (Kabe, 1963) and multivariate multiple regression, described in Mulaik (1972, pp. 411–417). Although Jöreskog (1971) described a nested sequence of models for testing a model of congeneric tests, the idea was more fully developed and popularized for structural equation modeling by Bentler and Bonnet (1980). To evaluate certain kinds of structural equation models, a specific sequence of nested models that was the forerunner of the four-step procedure was elaborated by James, Mulaik, & Brett (1982). Anderson and Gerbing (1988) independently described a similar approach called the "two-step approach." An example of a "three-step approach" is given in Carlson and Mulaik (1993).

As background, it is important to distinguish two kinds of nested sequences: (a) a parameter-nested sequence and (b) an equivalence-nested sequence (Bentler & Bonnet, 1980).

A parameter-nested sequence of models is one in which every model presumes the same path structure, but may begin by fixing only enough parameters to achieve identification of the model. A parameter-nested sequence of models is then constructed by constructing successive models in which fixed parameters in preceding models of the sequence are carried over to each successive model, while new parameters are fixed in each successive model. Each successive model is more constrained than the previous model but retains constraints of the previous model.

An equivalence-nested sequence of models is a series of models that is not itself parameter nested, but nevertheless each successive model has equivalent fit to the covariance matrix as a corresponding successive model in a parameter-nested sequence of models. For example, a confirmatory factor-analysis model with free covariances among its latent variables will fit equivalently to a structural equation model that specifies free paths from every latent variable to the others and can be used in place of the structural model to assess model fit. Equivalence-nested sequences of models are often used because they are easier to formulate, conceptualize, or analyze. The four-step procedure involves an equivalence-nested sequence of models.

ASSUMPTIONS AND CONTEXT FOR USE OF THE FOUR-STEP PROCEDURE

The Method of Analysis and Synthesis

Although analogous-nested sequences of models may be formulated for testing other kinds of structural equation models with latent variables, the method described here is designed for a specific kind of structural equation model that is commonly formulated. It is presumed that the researcher has previously determined a set of latent construct variables to study and wishes to test a hypothesis about how they are causally related. This reflects the analysis–synthesis strategy pervasive in science and other fields, originating with the 17th-century philosopher Rene Descartes, of first analyzing or breaking down a problem or domain conceptually and/or physically into its clear and distinct ideas, concepts, or elements, and then proceeding to synthesize or join them together again conceptually and/or physically so that they reproduce the original phenomena or complex concept to be understood (Mulaik, 1987; Schouls, 1980). So, in this analytic–synthetic framework, we presume that the researcher has already performed some form of analysis, either conceptual, statistical (e.g., exploratory factor analysis with other data), and/or physical to identify clear and distinct fundamental variables of some domain of content, and has then synthesized these into a model of how these variables are causally related. The latent variables in the model should exhaust the important fundamental variables to be considered. If there are important latent mediating variables linking the latent variables chosen, they too should be added to the model and their linkage to the other variables specified. Ultimately the aim is to test the synthesized model of relations between the latent variables.

Constructing Indicators for the Model

Before one can test a model of the causal relations between latent variables, one must also link the latent variables to manifest indicators of them. This means one must construct or select, on the basis of prior conceptual or statistical analyses, manifest variables that are "indicators" of the latent constructs. (This differs from taking a convenience set of manifest variables at hand, inspecting them to formulate plausible latent causal variables that seemingly exist among them, formulating a hypothesized model of their causal interconnections, and then testing the model with these same variables. Although this is possible and is the way some may formulate their models, it is not the basis for structural equation models for which the four-step procedure is designed. For one thing, this other approach can be more vulnerable to artifacts and extraneous effects and often has insufficient numbers of indicators to reveal these.) To link latent variables to manifest variables, the researcher should construct or select at least four manifest indicators of each latent construct, preferably with different methods of measurement, such that among these indicators the latent variable indicated by them is a common factor, that is, the most immediate (hypothetical) common cause of them. The manifest indicators of a latent construct should also be such that there is no prior evidence or reason to believe that they have other common causes among them. We would allow indicators of one latent variable that also are simultaneously indicators of other latent variables in the study, but would recommend that simple indicators of a single latent variable are preferable to complex indicators of several of the latent variables, especially if the aim is to develop scales of homogeneous, univocal items.

Rationale for Four Indicators

The reason four or more indicators are selected or constructed for each latent variable is to overdetermine the latent variable by its relations to specific manifest indicators. Anderson and Gerbing (1988) noted that requiring four indicators of a latent variable was essential to provide an internal consistency test of the unidimensionality of a set of variables using Spearman's tetrad difference equations (Glymour, Scheines, Spirtes, & Kelly, 1987; Hart & Spearman, 1913). If the four variables—x_1, x_2, x_3, and x_4—have a common factor, then the correlations between the four variables should all be nonzero and satisfy the equations $\rho_{21}\,\rho_{34} - \rho_{23}\,\rho_{14} = 0$, $\rho_{24}\,\rho_{13} - \rho_{21}\,\rho_{34} = 0$, $\rho_{21}\,\rho_{34} - \rho_{23}\,\rho_{14} = 0$, where ρ_{ij} denotes the correlation between variables *i* and *j*. One can always fit perfectly a single common factor model to three positively correlated indicators, so no test of the single-factor model is possible with them. The model is only just-identified with respect to them, and the single common factor may be an artifact. But four positively intercorrelated variables may not have a single common factor. The single-common-factor model is overidentified with respect to the four indicators. This gives "multiple points of view" on the latent variable through each indicator and allows for the establishment (via a test) of its objective existence by being able to demonstrate an invariance across the several indicators (analogous to "seeing the same thing" from several points of view) by fitting a single common factor to them.

While overidentification may be achieved with two and three indicators and the paths from them via their parent latent variable to indicators of other latent variables, what is tested in this case is whether a causal connection exists between a (possibly irrelevant unmeasured) variable common to the three indicators and the indicators of other latents. To illustrate, suppose we have data generated according to the model in Figure 1a, where a common factor *c* is confounded with a method factor *m* in three of the four indicators of *c*. Suppose further that we want to test the hypothesis that *c* is a cause of η with this data. Inspection of the model in Figure 1a will show that *c* is not a cause of η. Nevertheless, suppose we had only the three indicators x_1, x_2, and x_3 of *c* and modeled this hypothesized relationship between *c* and η as in Figure 1b in a test of this hypothesis. Right off, we cannot perform a test that, among themselves, x_1, x_2, and x_3 have a single common factor. A tetrad-differences test of a common cause would require four variables. We can test only whether x_1, x_2, and x_3, when joined with various indicators of η, have a common factor. Indeed they do. But we would be very wrong if we thought this common factor is *c*, because it is

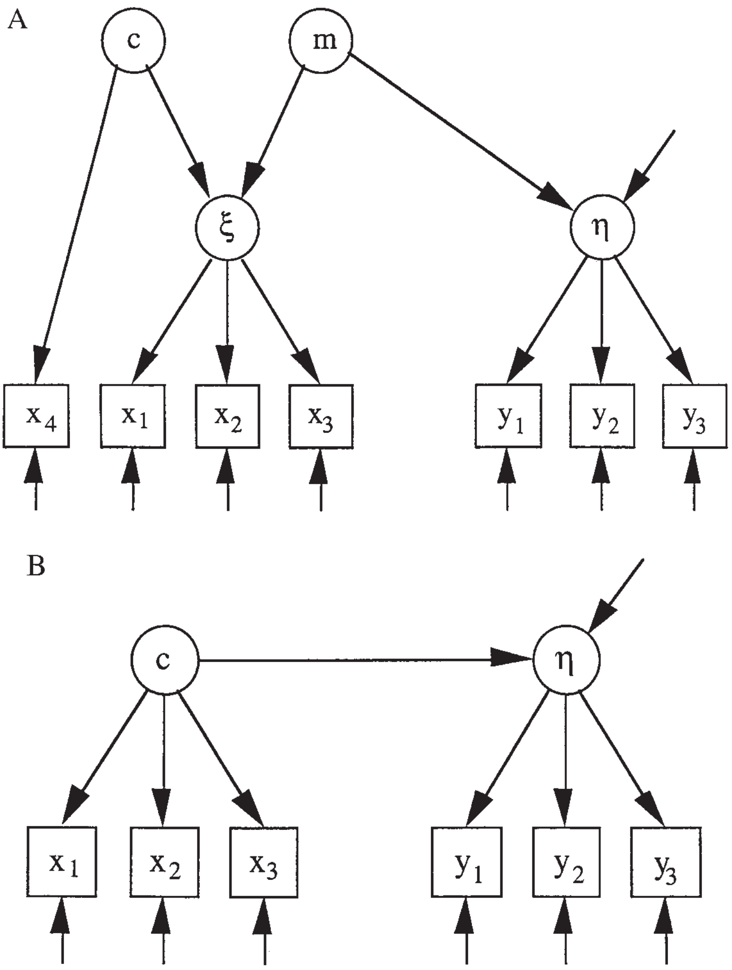

FIGURE 1 (A) A common factor *c* is confounded with a method factor *m* in three of the four indicators of *c*. (B) A model of the three indicators that would speciously suggest that *c* is a cause of η.

m. Nevertheless, because the model in Figure 1b would fit the six manifest variables perfectly, it would create the illusion that *c* is a cause of η, when in fact, *m* is the common cause of x_1, x_2, x_3, and the indicators of η. By including a fourth indicator x_4 of *c* in the model in Figure 1b, a test of the existence of a common factor among the four indicators—x_1, x_2, x_3, and x_4—is possible that is not available with just the three indicators. In this case a tetrad-difference test among these four variables would still yield a vanishing difference and support a common factor among the four, which is *c*. However, this test is distinct from tests in the structural model of whether the common cause *c* of x_1, x_2, $x_{3,}$ and x_4 is a common cause of η and its indicators, which clearly it cannot be in Figure 1a. The correlation between x_4 and each of the indicators of η is zero. They do not have a common cause. The test is not infallible, however. If x_4 is infected with the same methodological factor, one could be misled again. Nevertheless, carefully selecting four indicators of *c* that are measured by different methods reduces the possibilities of confounding common-method artifacts with the constructs measured.

Including four or more indicators is also a way to increase the degrees of freedom without arbitrarily fixing structural parameters to achieve that purpose. This provides more conditions by which to test the concepts used to construct or select the indicators as having a common factor. When formulating hypotheses about a specific constraint placed on a structural parameter, it also allows one to test this constraint separately using a chi-square difference test by comparing the chi square of a model with the constraint to the chi square of a less constrained comparison model that frees the constraint without losing identification of the model. (A problem with "sparse" models with many latent variables and few manifest indicators is that some zero relations specified between latents may correspond to identification conditions, without which the model would be underidentified. Ordinarily one would expect that a hypothesized variable not be defined by its having a zero effect on another variable. This assertion should be empirically testable as an empirical property of the variable and not "true by definition." But it is impossible to free such a zero parameter to see what its value is, free of overidentifying constraints, for in these cases, the model becomes underidentified. Such a situation was previously noted by Mulaik and Quartetti [1997] in connection with models with a general factor and group factors, all of which are orthogonal to the general factor. It is not possible to free the correlations between the general factor and the group factors without creating an underidentified model. The orthogonality property is not a testable property.) Finally, as we will see, having four or more indicators of each latent construct lends itself to the use of confirmatory factor analysis in evaluating certain aspects of the structural equation model.

Specifying as Many Parameters as Possible

Although it is not yet often done, ideally the structural equation model to be tested should prespecify values for all parameters. This makes explicit values for parameters in hypotheses to be tested in the fourth step of the procedure. Values to be specified may come from previous studies or they may be simply "ballpark" guesses of the researcher who asserts values in an effort to be as explicit as possible about his or her theory. Given the variances (or metric) of the manifest variables, structural coefficients will be more important to specifying a theory than variances and covariances of latent variables, so these should be specified if possible. Of course, if the researcher feels uncomfortable about specifying all parameter values, specifying as many as possible would be an acceptable compromise.

THE NESTED SEQUENCE

For a given structural equation model that we wish to test, we may test various aspects of this model with the following nested sequence:

1. Step 1: The Unrestricted Model

This model tests the hypothesis that a common-factor model fits the covariance matrix among the observed variables for a specified number of common factors *k* without confounding that issue with "measurement" issues about the specific relations of indicators to latent variables or structural relations between latent variables. The unrestricted model is effectively an exploratory common factor analysis model for the same number of factors as the number of latent variables of the structural equation model. It should be estimated with the same method of estimation as used to test the structural equation model, for example, with maximum-likelihood estimation. As a practical matter an unrestricted model may be estimated using an exploratory factor-analysis program like SPSS FACTOR, ML extraction. It will be important to obtain the chi-square goodness-of-fit test from this analysis to be used in constructing goodness-of-fit indexes. It is essential to understand that the test of this model is a test of constraints implicitly in the fully specified structural equation model. Unacceptable fit obtained for the unrestricted model would also indicate unacceptable fit for the structural equation model. Mulaik indicated the following reasons for proposing Step 1:

> We were led to introduce Step 1 after encountering many situations in which the measurement model (of what is NOT related to what between the latents and the manifest indicators) was unacceptable in fit, and people began freeing up more and more loadings in the factor loading matrix, until they were back to effectively the unrestricted model, and they were still rejecting that, suggesting they had begun with a wrong idea for the number of factors, that they should have included more in some way.
>
> We realized then that the unrestricted model is [equivalence] nested within the measurement model, which in turn is [equivalence] nested within the structural model, etc. And so, it could be tested first and if accepted, then you knew that if you had to back track later on in Step-2 with modification indices leading you to free up a few of the zero loadings, you would still not have the worry that you had the number of factors wrong, [i.e., the lack of fit was due to specifying too few latent variables] that the problems only exist at this point in the Lambdax matrix for the specified number of factors. E.g. an indicator may load on more than one of [the] latents, which is O.K.,

> if you do not have to radically change your concept of the latents to accommodate this. (S. A. Mulaik, SEMNET April 4, 1997 01:12:49)

Specifying the unrestricted model. An unrestricted model for k latent variables is simply an exploratory common factor analysis model $\Sigma = \Lambda\Phi\Lambda' + \Psi$ with k factors fixed in number. In his definitive paper on the maximum likelihood solution for the exploratory common factor analysis model, Jöreskog (1967) proved the following: Given the common factor model, $\mathbf{Y} = \Lambda\mathbf{X} + \mathbf{V}$, where $\mathbf{Y}$ is a $p \times 1$ random vector of observed random variables, Λ is a $p \times k$ ($k < p$) matrix of factor pattern coefficients, $\mathbf{X}$ is a $k \times 1$ random vector of common factor random variables distributed MVN($\mathbf{0}$,$\mathbf{I}$), and $\mathbf{V}$ a $p \times 1$ random vector of unique factor variables distributed MVN($\mathbf{0}$,Ψ^2), with cov($\mathbf{X}$,$\mathbf{V}$) = $\mathbf{0}$, and $\mathbf{Y}$ distributed MVN($\mathbf{0}$,Σ), where $\Sigma = \Lambda\Lambda' + \Psi^2$; then, for a specified number of common factors k for which the common factor model is identified, a sufficient condition that the maximum likelihood fit function

$$F_{\mathrm{k}}(\Lambda,\Psi^2) = \log|\Sigma| + \mathrm{tr}(\mathbf{S}\,\Sigma^{-1}) - \log|\mathbf{S}| - p$$

is minimized conditionally on a given Ψ^2, with value $f_k(\Psi^2)$ is that

$$\Lambda = \Lambda\,\mathbf{E}_k[\mathbf{D}_k - \mathbf{I}]^{1/2}$$

and $\mathbf{E}_{\mathrm{k}}$ consists of the first k eigenvector columns of $\mathbf{E}$ and $\mathbf{D}_{\mathrm{k}}$ is a diagonal matrix of the first k largest eigenvalues (all greater than unity) of the eigenvalue matrix $\mathbf{D}$ of the matrix $\Psi^{-1}\mathbf{S}\,\Psi^{-1}$. Furthermore, $F_{\mathrm{k}}(\Lambda,\Psi^2)$ is minimized unconditionally when Ψ^2 is the diagonal matrix such that $f_k(\Psi^2)$ is a minimum over all diagonal Ψ^2, and Λ is as above, conditional on the Ψ^2 that minimizes $f_k(\Psi^2)$. The solutions for Λ and Ψ^2 that minimize $F_{\mathrm{k}}(\Lambda,\Psi^2)$ are the maximum likelihood solutions of an exploratory factor analysis for k factors.

Jöreskog (1969/1979a) originally coined the term “unrestricted model” to refer to a model that “does not restrict the common factor space, i.e., one that leaves $\Lambda\Phi\Lambda'$ unrestricted. All such solutions can be obtained by a rotation of an arbitrary unrestricted orthogonal maximum likelihood solution” (1979a, p. 23). All unrestricted models should have the same communalities and the same uniquenesses and have equivalent degrees of fit to the observed variance–covariance matrix. Originally, after reviewing proposed sufficient conditions suggested by Reiersøl (1950), Howe (1955), and Anderson and Rubin (1956), Jöreskog (1969) believed that in confirmatory factor analysis it would be sufficient to specify an unrestricted model by introducing k^2 fixed parameters appropriately distributed across all factors within the matrices Λ and Φ while leaving all remaining parameters free. He perhaps thought this was so because, in identifying the eigenvalue–eigenvector solution of an exploratory factor analysis model, he constrained $\Lambda'\Psi^{-1}\Lambda$ to be diagonal (fixing $k(k-1)/2$ off-diagonals to zero) and required Φ to be an identity matrix (fixing k diagonals to unity and $k(k-1)/2$ off-diagonals to zero), for a total of k^2 fixed parameters in all. This led to eigenvalue–eigenvector solutions based on the matrix $\mathbf{\Psi}^{-1/2}\mathbf{S}\,\mathbf{\Psi}^{-1/2} - \mathbf{I}$ where **S** is the sample variance/covariance matrix for the observed variables.

But in an addendum to a reprint of this article (Jöreskog, 1979a, pp. 40–43) he corrected an implication of the earlier assertion, that among the k^2 constraints imposed on Λ and Φ it would be sufficient to fix the elements of Λ to nonzero elements to specify a unique solution. Although this may achieve identification or a unique solution in some cases, it will not always do so. Jöreskog noted that Howe (1955) had specified a sufficient condition (for rotation) that fixed only *zero* elements in Λ, and he restated and proved the original sufficiency conditions given by Howe (1955): (a) Fix each of the k diagonal variances of Φ to unity, and free the off-diagonal elements of Φ; (b) in each column of Λ, have $k-1$ fixed zeroes; and (c) each submatrix Λ_s, $s = 1, \ldots, k$, of Λ composed of the rows of Λ that have fixed zeros in the *s*th column must have rank $k-1$. That will place $k(k-1)$ fixed zeros in Λ and k fixed 1s in the diagonal of Φ. All remaining parameters of Λ are freed parameters. Alternatively, one could leave all elements of Φ free, while inserting in each column of Λ $k-1$ fixed zeroes and one fixed nonzero value in each column, the fixed 1s to be inserted respectively in different rows of Λ. Again all remaining elements of Λ are left free. Condition (c) must also hold.

More recently, Jöreskog has suggested the following as a way of implementing the above conditions when specifying an unrestricted model: Using prior information, previous analyses, or theory, determine for each hypothesized factor a variable that you believe will have the largest or near largest loading on that factor and low or near zero loadings on other factors, to serve as the indicator of the factor. The loadings of the indicators of different factors should be in different rows of Λ respectively. Jöreskog then says,

> In LISREL, specify a model where these rows have fixed zero elements everywhere except for the loading which was largest. Specify this to be free and all other elements of LX free. You should now have exactly k – 1 zeros in each column. Specify the factor variances to be 1 (by PH=ST). The resulting solution should be equivalent to (or a rotation of) the original exploratory solution. This can be verified by comparing the unique variances, which should be the same in both solutions. If they are not, it may mean that some of the unique variances are not identified. See my paper with Bollen in SMR 1985 [Bollen & Jöreskog, 1985]. Even if the two solutions are equivalent, the chi-square of LISREL will be slightly higher than the one in EFAP [exploratory factor analysis program] because of [a] different multiplicative factor in front of the ML fit function. However, the minimum value of the fit function should be the same. This can be verified by requesting technical output. (personal communication, February 14, 1995)

In the Appendix we have listed a proof that, under specified conditions, given an identified exploratory factor analysis solution for *k* factors, one can construct a unique transformation of the factors that would produce zeroes in the places specified by Jöreksog's method described earlier for specifying the factor pattern matrix of an unrestricted model.

We have generated a number of "population" covariance matrices for common factor models with a known number of factors, known factor loadings, known correlations among factors, and known unique variances. These covariance matrices were then analyzed by Jöreskog's maximum likelihood estimation algorithm in the SPSS FACTOR program with the number of factors specified. The chi-square statistics from these analyses were compared with chi squares from the fitting (using EQS) to the same covariance matrices of unrestricted structural equation models specified according to Jöreskog's recent recommendation with the same number of factors specified. The chi squares were very similar across the two methods, although the discrepancies between them, though still small (less than a full unit in magnitude), were slightly more than could be accounted for by differences in premultiplying the likelihood function by *N* versus *N* – 1. At the present we think these discrepancies are due to differences in programs and convergence criteria. We also determined that choosing different indicators for the underlying factors to specify the unrestricted model would still yield solutions with equivalent chi squares as long as the indicators chosen were univocal indicators of their underlying factors. If the indicators chosen were strongly dependent on several or all of the factors, sometimes the solution would yield improper solutions such as having correlations between the factors equal unity. We also found that obtaining convergence with an unrestricted structural equation model depended on giving the program good starting values. We also found with models that specified fewer factors than the number that generated the covariance matrix that sometimes very large samples were essential to detect, that the number of factors had been underspecified, say, by a single factor, because the chi squares, though considerably different from zero, were not significant.

A further presumption is that the unrestricted model is at least empirically identified. It will not be identified if the number of free parameters exceeds the number of distinct elements of the covariance matrix. This will be the case if the number of factors *k* hypothesized does not satisfy Ledermann's inequality

$$k \leq \frac{(2p+1)-\sqrt{8p+1}}{2}$$

where *p* is the number of observed variables (Harman 1960, pp. 69–73; Ledermann, 1937; Mulaik, 1972, p. 138; Thurstone, 1947). If the common factor model is just-identified or underidentified, it cannot be tested against data and no conclusions can be drawn from such a model.

However, models other than the one considered by Jöreskog may still be identified, especially those imposing appropriately k^2 fixed values obtained initially as parameter estimates from an exploratory factor analysis of the manifest indicators having the same number of factors. The fit of the unrestricted model sets an upper bound on the good fit of subsequent models of the nested sequence to be tested. So at this step this fit should be very good, otherwise one should stop and reevaluate one's hypothesis before going on further to test the model. Less-than-optimal fit suggests that the possibility exists for additional common factors or for doublet factors or other dependencies among the manifest indicators not accounted for by k common factors, the latter possibly showing up in large Lagrange Multiplier test values for tests of zero covariances among disturbances. There may be something wrong with one's analytic procedures. For reasons yet to be determined, a common factor model for that number of factors just may not be appropriate to the data and that means something is wrong with one's structural equation model as well.

Before abandoning the model at Step 1 due to lack of fit, some researchers will attempt to salvage the model by freeing zero loadings, or by permitting nonzero covariances among the "unique" factors. Nonzero covariances of this sort are occasionally supported by theory, as when a block of measured variables shares a common method of measurement, and the method factor has not been included (Marsh & Bailey, 1991). This theory would support the specification and testing of an alternative model. On the other hand, multiple relaxations of parameter constraints at Step 1 that are guided primarily by fit indexes (e.g., modification indexes) are not recommended. A number of studies have shown that repeated respecifications of this sort will fail to lead to the "true" model in many cases, and that the chosen modifications can vary widely across samples (Kaplan, 1989, 1990; MacCallum, 1986; MacCallum, Roznowski, & Necrowitz, 1992). Our view is that if modifications to the Step-1 model are introduced in this manner, the resulting modified model should be regarded as an exploratory creation that will need confirmation in future samples.

The Problem of Models With More Latent Variables

But suppose there is a different kind of model that also fits the same data but with many more latent common factors than the number in one's structural equation model. Hayduk has raised such a possibility as a challenge to the four-step procedure. The "correct"model may be a model with more latent variables, but the first step of the four-step procedure finds the number of latents hypothesized in one's structural model as acceptable. Is this not a fatal flaw of the procedure? We think not. It is easy to conceive of imagined scenarios in which an acceptably fitting simpler model may fail to accurately represent a subtly more complex situation. On the basis of mathematics alone one can conceive of countless models more complex than any given model that just might fit the data as well as or better than one's chosen model arrived at through an analysis and then conceptual synthesis of a particular domain. The problem is that these alternative models must be shown to represent something natural in the domain, and for many such models, this is not possible. But beyond that, if there are no prior reasons for giving specific values to the many additional parameters of a model with more latents, they will have to be introduced as free parameters, and one likely will not have as many degrees of freedom as one's chosen model. Many such models will be underidentified. But if such models are identified and we are to choose between models that fit within one's error tolerances, models with more degrees of freedom will be preferred. They are subjected to more tests of fit. Where models with more latents best enter into consideration is in the analysis phase prior to one's synthesis of these latents into a model, and they must have more degrees of freedom than competing models with fewer latents, to be taken seriously.

The four-step procedure is based on a presumption that the scientist's analysis has considered the evidence in the literature or situation for a more complex model and found it lacking. By selecting or constructing the indicators of the latent variables after a conceptual and/or statistical or other analysis has been performed with other data to isolate the latent variables for study, one is able to focus on generating sets of indicators that will be generally unifactorial indicators of the respective latent variables, and at most dependent on only the latent variables included in the study. At this indicator selection stage, one should today be conscious of possible ordered dependencies among indicators, like the variants of the simplex model that Hayduk put forth in the SEMNET debate, which would introduce other sources of dependency among the items not due to the latent variables yielded by the preliminary analysis. These extraneous sources of dependency should be eliminated or controlled for. This is simply an experimental design problem. The goal is to be able with good reason to assume that if one conditioned the manifest indicators on the chosen latent variables of the study, the indicators would then provide evidence of conditional independence. This then justifies testing the unrestricted common factor model, for the test is of whether, after conditioning on *k* latent variables, the residuals exhibit independence by being at least uncorrelated (which, if one can assume multivariate normality for the residuals, is an indication of independence).

Is this test infallible? No. No tests in science are infallible. The question is whether it reasonably reaches its conclusions within the framework of its assumptions. We think it does. It rejects the stated hypothesis of the unrestricted model if nonnegligible lack of fit is detected. That alerts the researcher to carefully reconsider the model and the latent variables hypothesized. Negligible lack of fit is set to be quite small so that if one accepts the hypothesis by observing only negligible lack of fit, only models with slightly better but negligible lack of fit might still be considered, and many of these may be ruled out because they have fewer degrees of freedom. Again, remember that the constraints of the unrestricted model correspond to an equivalent subset of constraints of the structural equation model, and all one is testing is to see if these constraints introduce lack of fit. Nevertheless, the test will detect in many cases failure of the assumptions of the factor model in data generated by models like the simplex model put forth by Hayduk, because the simplex model and its variants will have more common factors than those of one's hypothesized common factor model, or will behave inappropriately in Step 2. But the simplex models should also be ruled out in the analytic phase as one selects or constructs one's indicators, for these are suggested if one is aware of a natural ordering among the indicators, for example, either in time or in space, and causal mechanisms between adjacent and successive indicators in the ordering. Mathematically, one does not always have to specify an ordering among the variables to establish a simplex among them (Browne, 1992; Schönemann, 1970). Data generated by the simplex model and related "molar models" were regarded by Jones (1959, 1960) as inappropriate for exploratory factor analysis. Hayduk (1996) apparently rediscovered this in a study of personal space indicators, which is described in Hayduk and Glaser (2000).

2. Step 2: The "Measurement Model"

This is a confirmatory factor analysis model that tests hypotheses about certain relations of the manifest indicators to the latent variables. Specifically, further zero loadings are imposed to indicate which indicators are *not* dependent on a given latent variable. This is done for each latent variable presumed connected to manifest indicators. Usually, but not necessarily, the resulting factor structure will permit only a single nonzero loading for each indicator. If one gets acceptable fit of the measurement model, one is able to go on to test the structural equation model of Step 3.

Step 2 does not concern all aspects of measurement. Obviously, simply indicating which latent variables are not related to which manifest indicators does not specify all that one might in terms of "measurement," so the convention of calling this a "measurement model" can be misleading. Calling this model the "measurement model" probably draws on Jöreskog's (1979b) use of that term to refer to the relationships specified in the LISREL model between manifest indicators and latent variables as given by a factor analytic model. To some extent, Hayduk's preoccupation with whether the "measurement model" really can separate "measurement from structural concerns" in this debate reflects a view, perhaps present in the literature surrounding the Anderson and Gerbing (1988) paper, that Hayduk (1996) comments extensively on. He projects this view onto Mulaik, who actually regards the term "measurement" used here with a grain of salt, seeing that the term is used in the scaling literature to refer to "a procedure for identifying values of quantitative variables through their numerical relations to other values" (Michell 1990, p. 63). In this literature, establishing that one has a quantitative variable is a complex process that rarely is undertaken in the behavioral and social sciences when numbers are obtained in empirical settings. Most structural equation modelers simply assume, rather blissfully and without qualms, that they have quantitative variables and thus are "measuring" something with them. But the problem of "measurement" in this sense remains largely ignored and few numerical variables used in structural modeling likely meet the rigorous standards of scaling theorists for quantitative variables for which measurement applies. However, we will let sleeping dogs lie and not raise this problem here, other than to note that in being aware of it, we hardly believe that the "measurement model" deals with testing all aspects of measurement pertaining to a model.

For the purposes of the present discussion, however, measurement concerns relations between latent variables and manifest indicators of them. This is consistent with most treatments of measurement with latent variables. For example, Meredith (1993) argued that measurement invariance involves showing that, when $\mathbf{X}$ is a vector of manifest random variables and $\mathbf{W}$ a vector of latent variables measured by the variables in $\mathbf{X}$, then, given other variables $\mathbf{V}$ by which populations of participants may be selected, measurement invariance holds if $F(\mathbf{x}|\mathbf{w},\mathbf{v}) = F(\boldsymbol{x}|\boldsymbol{w})$ for all $\mathbf{v}$ in the sample space $(\mathbf{x},\mathbf{w},\mathbf{v})$, where F( ) denotes a cumulative distribution function. Meredith (1993) also provided a definition of weak measurement invariance, that $E(\mathbf{X}|\mathbf{w},\mathbf{v}) = E(\mathbf{X}|\mathbf{w})$ and $\Sigma(\mathbf{X}|\mathbf{w},\mathbf{v}) = \Sigma(\mathbf{X}|\mathbf{w})$ for all $\mathbf{v}$ in the sample space, where E( ) is the expected value operator, and Σ( ) is the variance–covariance of the variables in its argument. Under multivariate normality, measurement invariance and weak measurement invariance are equivalent. Suppose a measurement model for variables in $\mathbf{X}$ is given by a common factor model $\mathbf{X} = \alpha + \Lambda\,\mathbf{Z} + \mathbf{U}$ for all $(\mathbf{x}, \mathbf{z}, \mathbf{u})$ in the sample space, where $\mathbf{z}$ denotes common factor and $\mathbf{u}$ unique factor random vectors. Meredith showed that weak measurement invariance will almost certainly hold in this case if $E(\mathbf{X}|\mathbf{v}) = \alpha + \Lambda E(\mathbf{Z}|\mathbf{v})$ (implying $E(\mathbf{U}|\mathbf{v}) = \mathbf{0}$ for all $\mathbf{v}$) and $\Sigma(\mathbf{X}|\mathbf{v}) = \Lambda\Phi(\mathbf{Z}|\mathbf{v})\Lambda' + \Psi$ for all values of $\mathbf{v}$ in the sample space. This requires invariance of the factor loadings Λ and the unique factor variances Ψ (diagonal) but not of the common factor variances and covariances Φ. This argues that specifying the elements of Λ and Ψ are essential to specifying a particular common factor measurement model. But Step 2 only specifies zero elements in Λ and does not fix variances of Ψ, so it hardly constitutes an essential specification of a measurement model and concerns a more limited set of hypotheses. Many of these specifications will be imposed later in Step 4 if an essential measurement model is specified and tested.

Dealing with lack of fit of the measurement model. Sometimes this model does not fit well initially, because certain fixed zero loadings of indicators on latents are misspecified. This might indicate a serious problem for an indicator if theory specifies that a certain latent variable should have no direct relationship with an indicator of another latent variable. The indicator in question must be examined closely to see if there is not a possible basis for the indicator's also being dependent on another latent variable in the study besides the one it is intended to be an indicator for. In many cases, researchers who base their models on prior factor analyses ignore the fact, in their zeal to inject pure univocal simple structure into their models, that while an indicator may load chiefly on one latent variable in a simple structure solution, other common factors may also have smaller nonzero loadings on that indicator. Forcing these smaller but nonzero loadings to be zero can contribute to unnecessary lack of fit. So, the issue that must be addressed is whether one can free some of the offending fixed zero loadings in the anticipation that they may represent nonzero values without doing violence to one's conception of the latent variables and their interrelationships. This is a scientific judgment call. The judgment is helped by prior evidence or a prior rationale that would suggest that a latent variable presumed not to be related to a certain indicator should indeed be related to it. To locate offending zero loadings, one can test each zero loading using a modification index or a Lagrange Multiplier test. Then by proceeding to free the zero loading with the largest modification index or chi square, one reestimates the model, repeating this procedure one parameter at a time until one gets good fit. But this procedure also loses degrees of freedom, and the resulting model should be regarded as exploratory, as discussed earlier. A zero loading that produces lack of fit may also indicate something wrong with one's theory of the latent variable.

3. Step 3: The Structural Equation Model

This is the structural equation model itself, with the same zero relations between the manifest indicators and the latent variables established by the measurement model, and the same freed loadings. However, fixed zero constraints are introduced concerning the relations between certain pairs of latent variables, so that some latent variables are not dependent on other latent variables. One goes forward after this step only if one still gets acceptable fit (e.g., CFI > .95, root mean squared error of approximation [RMSEA] < .05). If only two latent variables are involved, this step, of course, would be passed over.

4. Step 4: Tests of Prespecified Hypotheses About Parameters Freed From the Outset

There are several ways to proceed: (a) One can perform a simultaneous test in which one fixes each of the previously freed parameters (only those freed from the outset) to prespecified values dictated by theory or estimates obtained from previous studies with different data. A drawback of this approach is that if one gets substantial lack of fit, it may not be obvious which fixed parameters are to blame. (b)

One can proceed to impose fixed values on freed parameters in a nested sequence of models, until one gets an unacceptable lack of fit, which identifies a misspecified parameter. One might leave that parameter then free (having provisionally rejected one's hypothesis about its value), and fix other parameters, until one has tested all of one's hypotheses about the individual, nonzero parameters. (c) One can perform a sequence of confidence-interval tests of hypotheses about individual, previously freed parameters using confidence intervals based on the standard errors of the estimated parameters. Because multiple tests are performed and these are not independent tests, a Bonferroni procedure involving more conservative significance levels might be imposed. Also note that using confidence intervals in this way reverts to statistical significance testing, whereas other approaches have replaced that with measures of model approximation.

THE HAYDUK–GLASER CRITIQUE

We have already stated the limits of applicability of the four-step procedure to the evaluation of a specific kind of structural equation model. But Hayduk and Glaser (2000) attribute some other limitations to it that we do not feel are appropriate.

Models With Saturated Relations Among Concepts Cannot Separate Measurement and Structural Concerns

"If one has a base model that is saturated with effects among the [latent] concepts, the [four-step] procedure will not be able to separate measurement from structural concerns because the Step-2 and Step-3 models will provide identical fit statistics" (Hayduk & Glaser, 2000, p. 6). While this is true, it is not so much a limitation of the four-step procedure as it is a limitation of the structural equation model specified to be evaluated. A model with saturated relations among the latents, meaning all of the structural coefficients between them are free to be estimated, specifies no testable hypothesis about their relations and so nothing about them is testable. Only if the researcher in this case has then specified certain structural parameters between the latents to have specific (nonzero) values can any further testing be performed, and this would occur in Step 4 and not in Step 3 (which in this case would be skipped).

Four-Step Procedure Not Usable With Models Having One or Two Indicators of Latent Variables

"The four-step procedure cannot be used if the model contains variables like sex, age, income, or education, or any concept [latent variable] with a single indicator" (Hayduk & Glaser, 2000, p. 6). Again not true. James, Mulaik, and Brett (1982) illustrated a "two-step" procedure with an example in which one of the latent variables was indicated by a single variable. The idea is to begin with a minimally specified version of one's model, or a version in which no further freeing of parameters can be done without losing identification or the essential framework of one's model. If certain fixed parameters cannot be freed at this first stage without losing identification for the model, then one must begin with these parameters fixed. But other parameters can be freed without losing identification, so an unrestricted version of the model with respect to those latent variables having four or more indicators can still be joined with latent variables having single indicators, as long as the single indicators have structural parameters and error variances specified. It may not be possible to use an exploratory factor analysis algorithm to perform the analysis to obtain the chi squares, but one can always use a regular structural equations modeling program.

If one has models with only one, two, or three indicators per latent variable at the most, one can do structural equation modeling, and one can seek to establish the minimally specified model from which to begin evaluating the complete model by a nested sequence of models analogous to the four-step procedure. But there are problems with models of this kind, and we will return to this point later.

Models With Many Indicators per Latent Variable Are Impractical

Hayduk and Glaser (2000) said that "A third limitation comes from observing that if a model has many indicators per concept, that model must contain fewer concepts if the model is to remain practical and if the model is to be estimated using a reasonable sample size" (p. 7). Hayduk and Glaser apparently believe that one cannot do a four-step analysis with many latents and more than four indicators per latent variable without having "huge samples," which would place an inordinate burden on researchers. Structural equation modeling already requires rather large samples to enable one to use statistical inference with chi squares and other statistics because these statistics have their theoretical properties only in large samples. Because having four or more indicators provides for more degrees of freedom and more possibilities of detecting the presence of artifacts and misspecifications of the model, one will have to pay the price of whatever sample sizes are necessary to make statistical inferences while taking advantage of the benefits of many indicators. But we note that Hayduk and Glaser have not backed up their claim with any hard evidence that models with many latent variables and more than four indicators will require prohibitively large samples. We note that H. Marsh provided evidence contradicting their claim:

> The use of four indicators per factors is a good recommended minimum (in relation to typical current practice of using even fewer), although my current simulation studies are leading me to suggest that “more is better.” In simulation studies I systematically varied the number of indicators per factor (2 to 12) and sample size (50 to 1000) with a simple three-factor model. For all levels of *N*, more indicators per factor did better in terms of convergence to a proper solution, accuracy and precision of parameter estimates, and factor reliability estimates. The advantages of having more indicators was most evident for small *N* where a large number of indicators seemed to compensate for small-*N*. Whereas this was based on one simulation study, the results seem to be holding in new research. (SEMNET, May 18, 1997, 09:30:22)

These simulation results were subsequently reported in Marsh, Hau, Balla, and Grayson (1998).

Some experts argue that the sample size should exceed some integer multiple of the square of the number of estimated parameters to ensure that the Hessian matrix will not be empirically singular due to sampling error, creating conditions of empirical underidentification and causing lack of convergence to solutions. Certainly this will be a greater problem for the first step because so many parameters must be estimated in that step. Although we have no proof of this, we think the problem can be somewhat diminished if one can use an exploratory maximum likelihood factor analysis algorithm to estimate the unrestricted model’s chi square, because in that algorithm the order of the Hessian matrix only equals the square of the number of unique variances that have to be estimated, instead of the square of the number of estimated parameters, so convergence is more likely possible in smaller samples.

But if the sample size is inadequate to deal with the large number of parameters of an unrestricted model of Step 1, it is certainly reasonable to skip that step and start with Step 2, “the measurement model,” where the number of parameters to estimate is much less. So, Hayduk and Glaser have not totally vitiated the multistep method to model evaluation with this argument.

Step-1 Tests Cannot Determine the Correct Number of Factors

Hayduk in SEMNET and Hayduk and Glaser (2000) are preoccupied with arguing that Step 1 is unable to determine the correct number of factors or concepts. The argument arose when Hayduk seized on the following statement made by Mulaik in SEMNET:

> a step preceding the measurement model evaluation, would be to test Jöreskog’s “unrestricted factor analysis model,” which is equivalent in fit to an exploratory factor analysis solution for a specified number of factors. This allows one to test whether one

> has the correct number of latent variables without confounding that test with the specification of relations between specific latent variables and specific manifest indicators. (Mulaik, SEMNET 23 March 1997 15:50:19 EST)

Mulaik and Hayduk had two different things in mind regarding the phrase "correct number of factors," used here. Mulaik was thinking in the context of the tests of the four-step's nested sequence of models in which models are accepted or rejected provisionally in terms of having acceptable or unacceptable lack of fit. You accept a model as provisionally "correct" if you obtain "acceptable lack of fit" (e.g., nonsignificant chi square, a CFI > .95). Accepting the unrestricted model meant you could accept as provisionally correct those constraints in the full structural equation model that are equivalent to the constraints of an exploratory common factor model with factors equal in number to the number of latent variables in the structural equation model. But Hayduk apparently interpreted "correct number of factors" in a stricter absolute sense, and then, despite Mulaik's denial that this is what he meant, argued that Step 1 does not determine the correct number of factors and can lead one to accept the wrong number. Hayduk furthermore seemed to project onto the Step-1 test the procedure in exploratory factor analysis of seeking the minimum number of common factors that fit the data acceptably: Once you had rejected the hypothesis for *k* factors, you would then automatically test the unrestricted model for *k* + 1 factors, and if that failed to fit, for *k* + 2 factors, and so on, until you had found the number *k* where you just fit the data acceptably. Or if you tested for *k* factors and obtained acceptable fit, and then tested for *k* – 1 factors and found that acceptable, you would proceed to *k* – 2 factors, and so on until you got unacceptable lack of fit, and just prior to this you would have the minimum number of needed factors. In other words, this would be a step in which you sought to determine by a trial-and-error technique the "correct number of factors" as the minimum number of factors that just fit the data. This was not Mulaik's view of what was done at Step 1.

Step 1 is not designed "to locate the correct number of factors" but to test the provisional correctness of a hypothesis regarding the number of factors. Mulaik said,

> I am hesitant to recommend a series of tests in which I successively drop one factor until I get a substantial lack of fit, i.e., in a search of the "minimum number of factors." That is not theory driven, and it would leave me in a quandary if I got a substantially smaller number of factors that also fit acceptably. Which of my theoretical factors should I discard? No, for the time being, unless others have suggestions otherwise, I would then proceed [having obtained acceptable fit] with the theoretical number k as being acceptable, and move on to Step-2 to test additional constraints on how certain indictors are unrelated to certain latent variables. (Mulaik, SEMNET 7 April 1997 17:16:15)

And he also added

> Step-1, in fact, all of the Steps are tests designed to assess if introducing a theoretical constraint leads to unacceptable fit. It is not primarily an exploratory technique we are describing here, although there is some room for that, if [in freeing parameters] you pay the penalty in loss of degrees of freedom and have not committed yourself at an early STEP in a way that conflicts with constraints you originally intended to introduce at a later STEP.

But Hayduk seemed not to be concerned with whether the model fit or not for a specified number of factors. He felt that the unrestricted common factor model, regardless, could not determine “the correct number of factors.” Because the chi-square test for the unrestricted model was equivalent to the exploratory factor analysis chi-square test for a specified number of common factors, it was inherently incapable of detecting the “correct number of factors” in certain instances. It would likely accept a model that had far too few latent concepts. What motivated him was his encounter with data that was seemingly generated by a simplex model. And here he made a subtle conflation of “number of common factors” with “number of latent concepts.” (One of the latent concepts–variables of the simplex model influences only one manifest indicator, and so it could not be a common factor.) Hayduk and Glaser (2000) described this encounter of Hayduk with a simplex. We will not repeat it in detail, but note that Hayduk began with a hypothesis that his 10 indicators would have a single common factor. Despite the fact that the first eigenvalue of his covariance matrix overwhelmed the remaining, he still did not get acceptable fit to the data. He did get acceptable fit with the chi-square test (using an exploratory factor analysis maximum likelihood program) for three common factors. But because he failed to get acceptable fit for the single common factor hypothesis, he decided (we would say wisely) to rethink the single factor hypothesis. This led him to consider a simplex model for the 10 variables, which has 10 latent concept variables, which fit even better than the three common factor model. The fact that the three common factor model fit optimally among all exploratory common factor models suggested to Hayduk that something is very wrong with testing for the number of latent variables using this chi-square test. You could get acceptable fit even when you have too few latent variables.

We have a number of answers to Hayduk and Glaser’s concerns with the first step of the four-step procedure.

To begin with, we note that Hayduk’s initially hypothesized single common factor model failed to fit his 10 indicators. In this instance a first-step test for the number of factors did its job: It detected that his model of a single common factor fit unacceptably. At that point there was no further point to continue with the subsequent steps of a four-step procedure. And Hayduk did what one should do when the first-step test rejects the model because it gets unacceptable fit for the hypothesized number of factors: He proceeded to reanalyze the situation and to consider alternative models, leading him to a simplex model for his indicators. (Unless he intended to get some other data, he should have paid no further attention to the covariance matrix among his indicators while formulating alternative models. If he then turned around and tested the simplex model against the same covariance matrix he used to formulate his model—adjusting the model to get best fit—no one would consider that convincing proof of the objective validity of the new model. What he should have done, and seems to have done, is examine how the variables were measured and how they stood with respect to one another in the experiment.) The simplex model presumes a natural ordering for the variables in time or space, which the common factor model does not. So, the simplex model represents an ordering in the measurements that the common factor model does not. The superiority of the simplex model to the common factor model in this case does not come from an assessment of fit—both the simplex model and the common factor model fit the covariance matrix acceptably—but from considering something outside the covariance matrix. The four-step procedure is only concerned with what the covariance matrix supports or does not support in the model tested in terms of fit.

But Hayduk remained disturbed by the fact that a three-factor exploratory factor analysis model fit the covariances among his 10 indicators. Hayduk and Glaser (2000) argued that "If the failure of the one-factor model had merely led to trying a two-factor and then three-factor model, Les would have ended up claiming that the data contained 3 factors, which is well below the 10 latent concepts that are probably there" (p. 14). But slipping into a full-blown exploratory factor analysis mode after failing to get an acceptable Step-1 unrestricted model to fit is not a part of the four-step model testing procedure, for it implies a complete abandonment of one's initial structural equation model on which the four-step procedure is based and use of the data at hand to formulate a new best fitting model (which cannot then be tested against that same data). So, in the context of his personal space study, this issue is irrelevant to the use of the four-step procedure.

We can imagine that there just might have been a situation in which Hayduk approached his 10 indicators with a three-factor hypothesis: There is a wide distance factor, a middle distance factor, and a close-up factor. And he would have gotten acceptable fit. But Hayduk argues that that would have been wrong. There are 10 latent concepts in the simplex model that best fit the data. So, he argues that the common factor model must be biased to settle for too few factors. He wonders whether there may be countless data sets best represented by models with more latent constructs, that will be fit acceptably by the common factor model with fewer common factors.

We will agree that examples can be constructed in which the common factor model with *k* factors will fit adequately in data generated by more than *k* factors. Hayduk and Glaser are asserting a stronger point however: The common factor model has an inherent tendency to fit with too few factors. In other words, factor analysts should expect that the model will fit with *k* factors when the true model involves more than *k* factors. The problem is that Hayduk and Glaser have not given enough evidence to support this stronger claim. To do so, we would like them to identify some inherent mathematical or statistical feature of the common factor model that renders it inadequate as a tool for identifying the number of factors. In the SEMNET exchanges, Hayduk consistently refused to present any such evidence. We think there is a good reason for his refusal to do so: His stronger claim is false. In fact, the experience of most researchers with confirmatory factor analysis is that it will often reject models at Step 1 or 2 that have received support from less demanding exploratory component or factor analysis. This experience has become so common that some researchers have argued (speciously, we believe) that confirmatory factor analysis is flawed because it rejects such models (McCrae, Zonderman, Costa, Bond, & Paunonen, 1996).

Suppose however that the data are not generated by a common factor model, but rather arise through some alternative latent variable model. Hayduk and Glaser argue that in this event, the Step 1 or 2 test will not operate properly because these tests are based on the assumption that the common factor model itself is adequate. But again we point out, this assumption is, by design, already implicit in the hypothesized structural equation model to be tested for lack of fit by the four-step procedure. The Step 1 and Step 2 tests for fit simply test separately the constraints of the common factor model apart from the other constraints in the structural equation model. So, the common factor model is an aspect of the full structural equation model as we have formulated it, and this aspect must be tested. Their quarrel is not with the common factor model but with tests for lack of fit in general. Yet we cannot imagine any tests against data that do not rely on lack of fit as an indication that something is wrong with a model or hypothesis. But logic informs us that getting acceptable fit is no absolute indication that a model is true: No test for lack of fit will reject a false model that nevertheless fits the data to which it is applied. The model can at most be accepted as only provisionally true. The fact that mathematics allows us to conceive of numerous distinct models that fit the same data equally or nearly as well prevents us from regarding acceptable fit as a final determination of the truth of a model. The only thing we can do to choose between equally well-fitting models is seek to show that most of these models do not represent anything objective, either with the data at hand or with additional data, and therefore these models can be discarded.

Again, it is probably true that one can in some cases generate data from a model that is not a common factor model that will be fit adequately by a common factor model. We certainly agree that investigators should carefully consider at the outset whether the variables under study can best be modeled with a common factor representation.

But the problem is only of concern if you believe, as do Hayduk and Glaser, that the factor model will routinely fit when the data are not generated by a factor model. Hayduk and Glaser have presented no evidence to support this claim. Isolated examples do not carry the day. Furthermore, Hayduk's argument in SEMNET (February 20, 1998) in connection with the simplex model, that the exploratory common factor model will become saturated (by Ledermann's inequality) and fit the data generated by a simplex perfectly (with a zero chi square) with fewer than the full number of latent factors of the simplex, is not an indication that a common factor model is always inclined to fit with fewer than the proper number of factors. The example cited of a saturated model with a zero chi square is not an example of a test of fit. The chi square has 0 *df* and therefore is uninformative as to the validity of the common factor model in that case.

But the reason we think their quarrel is with tests of fit and not with the common factor model of the four-step procedure is because their argument would seem to apply equally to tests of the Step-3 base model. A structural model may fit the data and yet be a false model. In what way do we resolve the issue by skipping the four-step and simply testing the base structural model when it too can fit yet be a false model? If a structural equation model based on a common factor model that fits is a false model, then it is possible that the structural equation model itself will also fit when tested, and yet be a false model. If we cannot use good fit of a model to its data as a basis for provisionally accepting the validity of a well-conceived model, what is the purpose then of testing the model's fit to data? What criteria then can we use for accepting or rejecting the model? They do not give any. It seems that their argument is a two-edged sword that undercuts their own position. The solution to the problem of the good-fitting false model is not to abandon testing and provisionally accepting models with good fit, but to abandon good-fitting models in favor of models with equal or better fit that have more degrees of freedom.

But what is the correct number of latent variables in Hayduk's personal space preference study with 10 indicators? Hayduk and Glaser seem to believe that 10 is the correct number. But on what basis? Is it because the simplex model fits acceptably? But that may not be sufficient. Is there any reason to suppose that fewer latent variables than 10 would handle the data just as well, making the ontological status of some of the latent variables of the 10 latent variable simplex problematic? Well, let us look at the simplex model for 10 indicators that Hayduk refers to. We illustrate it with the path diagram in Figure 2. It seems indeed that this model has 10 latent variables. But Jöreskog (1979c) observes that this model for three indicators is equivalent to a common factor model with a single common factor; for four indicators it is equivalent to a two common factor model. The reason for this loss of dimensions arises out of indeterminacies at the front and tail ends of the model, which are due to there being only single indicators of each latent variable. In the full model in Figure 2, Jöreskog would say that without loss of generality we can regard the latent variables $\eta_1, \ldots, \eta_{10}$ as scaled to unit variance. And for the inner variables parameters, $\alpha_2, \ldots, \alpha_9, \beta_3, \ldots, \beta_9$, and the error variances on the manifest

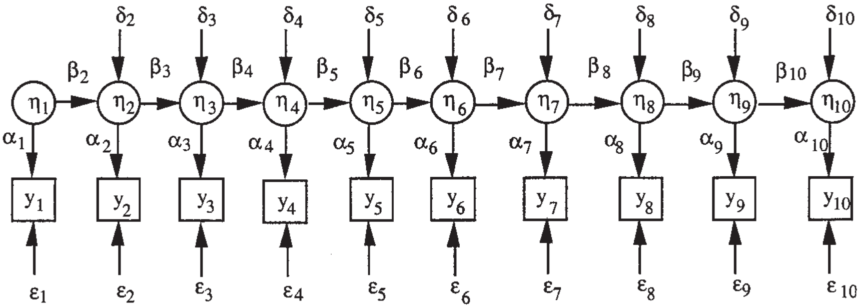

FIGURE 2 A 10-indicator quasi-Markov simplex model.

variables θ_2, …, θ_9, are identified; but for the outer indicators only the products $\alpha_1\beta_2$ and $\alpha_{10}\beta_{10}$ are identified. To eliminate this indeterminacy he suggests we make $\beta_2 = \beta_{10} = 1$. This makes it possible to consider a similar model (see Figure 3) that has only 8 and not 10 latent variables, η_2, … η_9, which fits the data just as well. These results suggest that, with single indicator models, some latent variables may lack independent grounds for postulating their existence, that they are artifacts of the model. Only when there are other variables dependent on the same latent variable is the independent existence of the latent variable reinforced. It also points out that merely having a latent variable in a model that is identified does not guarantee that it exists independently of the model. So, what is the correct number of latent variables? It is not clear why Hayduk and Glaser seem to think this is obviously 10 in this case.

Are Simplex Latent Variables Artifactual? But there is another problem with the simplex model with single indicators of the latent variables. Jöreskog (1979b) gives the degrees of freedom of this model as $p(p + 1)/2 – (3p – 3)$, where p is the number of manifest variables. In the case of a 10-variable simplex, there are thus $55 – 27 = 28$ *df*. Regard now each degree of freedom as corresponding to a fixed parameter that might be freed and have the model still be identified. Freeing any more parameters would result in the model being underidentified. Now, because the simplex model specifies that each latent variable is only a cause of the latent variable that immediately follows it, that means one must fix $(p – 1)(p – 2)/2$ structural coefficients of paths connecting latent variables not adjacent to one another to zero. In this case we have $(10 – 1)(10 – 2)/2 = 9(8)/2 = 36$ zero structural coefficients relating latent variables to one another. But we see that with the model having only 28 *df*, we could free only 28 of these 36 zero structural coefficients and have the model remain identified. That leaves 8 zero structural coefficients that cannot

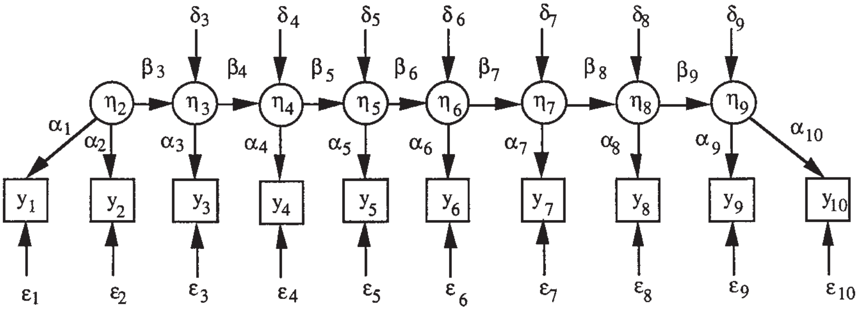

FIGURE 3 An 8-latent variable model equivalent to the 10-latent variable simplex model.

be freed without making the model an underidentified model. Because the latent variables of the simplex supposedly have substantive content, one would think that relations between them should be empirical and hypotheses about them empirically testable. But there is no way to test empirically that any of these 8 zero coefficients, respectively, are equal to zero independently of constraints that might be placed on any of the 28 remaining structural coefficients. One could not perform, for example, a chi-square difference test between a model in which 28 of the structural coefficients are free and the 8 zero coefficients fixed to zero, including the one in question, and a model in which 28 of the structural coefficients are free and 7 of the original zero coefficients are zero while the one in question is a free parameter. The first model would have 0 *df* and the second would be underidentified. Effectively, then, eight of the zero structural coefficients are part of a definition of the latent variables. Which coefficients are fixed to zero "by definition" is an arbitrary choice of the researcher. Any testable hypotheses about other structural coefficients are then relative to the eight that are fixed by definition to achieve identification of the model. The interpretation of the other 28 structural coefficients is then relative to the variables defined in part by having the eight fixed zero coefficients appropriately among them. But the fact that these variables and causal relations between them cannot stand alone empirically and be testable empirical relations raises the question of their artifactual status; they are creations of the researcher.

The same would not be the case, however, for a simplex of latent variables, each having four or more indicators, respectively. Then, all structural coefficients between latent variables will be overidentified and hypotheses about their empirical status testable. Herb Marsh, in fact, suggested on SEMNET (January 29, 1998) that researchers obtain multiple indicators of the latent variables of the simplex, and cited a paper of his where he had made this recommendation (Marsh, 1993). He also noted that in this case a simplex with multiple indicators becomes nested with a confirmatory factor analysis model, which we in turn note would make such a simplex model testable by the four-step procedure. But the simplex model is only a special case of models that Hayduk calls “sparse.” They frequently involve many latent variables relative to a small number of manifest variables; but identification and overidentification of structural parameters in these models are achieved by fixing many structural parameters to zero or other values, and we suspect in many cases these introduce unrecognized researcher artifacts into the models.

Approximation of the Common Factor Model to the Simplex. But we can also show why common factor models with many fewer common factors will fit simplex-generated data in many cases to within acceptable limits of lack of fit. Let us suppose that the variances of the disturbances on the latent endogenous variables of the 10-indicator quasi-Markov simplex model in Figure 3 are all small except for that of δ_5 and δ_8. Then we may be able to approximate the 10-latent variable simplex with a three common factor model as shown in Figure 4. For example, from Figure 3 we see that the covariance between y_3 and y_5 would be $\alpha_3\beta_3^2\beta_4\beta_5\alpha_5 + \alpha_3\theta_3\beta_4\beta_5\alpha_5$. But from Figure 4 we see, using path-tracing rules, that the covariance would be only $\alpha_3\beta_3^2\beta_4\beta_5\alpha_5$. Similarly, the variance of y_3 in Figure 3 is $\alpha_3^2\beta_3^2 + \alpha_3^2\theta_3$. But the variance of y_3 in Figure 4 is $\alpha_3^2\beta_3^2$. So, if α_3 and other disturbances’ variances are small, near zero, the common factor model will reproduce the covariances and variances very well. A similar point is made by Bast and Reitsma (1997; we thank Harry Garst for pointing this out on SEMNET [January 19, 1998]).

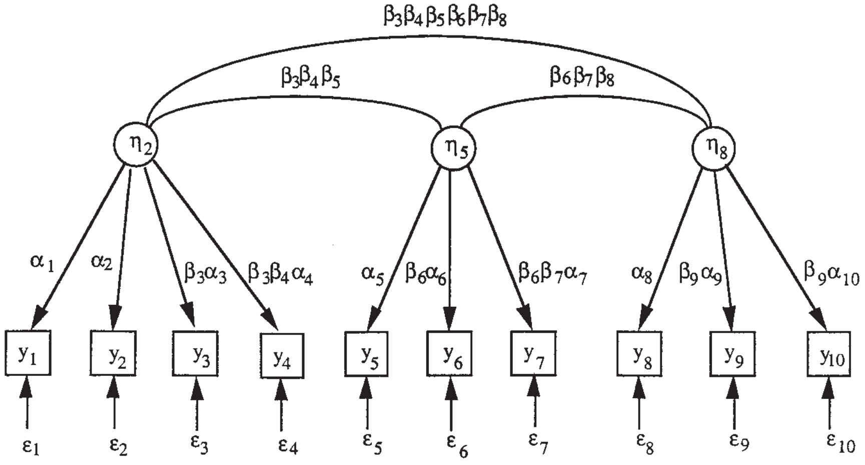

FIGURE 4 A three factor model that approximates a quasi-Markov simplex.

So, not only are there uncertainties as to the "correct number of factors" in an absolute sense due to underdetermination of parameters in a model, but also because in many cases models with different numbers of factors can fit the same data to within limits of acceptable fit. Ambiguities like this cannot be avoided in assessing the validity of models on the basis of fit to data, especially when there is at most one indicator per latent variable. To demand otherwise is to go counter to the logic of drawing inferences from model fit.

Can One Reconsider the Number of Factors at Any Step After Getting an Acceptable Fit at Step 1?

Hayduk and Glaser argue that we should be willing to question the number of factors despite a fitting Step-1 model. They argue that "four-step proponents will not be able to tell us to keep reconsidering whether we have the proper number of concepts at Step 2 and Step 3, without questioning the utility of the four-step procedure itself" (p. 17). This is nonsense. The four-step procedure is principally concerned with whether the various constraints of a fully specified structural equation model lead to lack of fit or not with the data. If after getting an acceptable Step-1 unrestricted model, one gets lack of fit in a later step, that should always be an occasion for reconsidering one's model and in turn the number of latent variables. It will be a judgment call based on external theory and knowledge beyond the data whether to completely abandon the model at this point, or to use modification index searches to find fixed parameters to free to get acceptable fit within the initially hypothesized framework.

The No-Peeking Problem

Hayduk and Glaser (2000, p. 2) argue, "Coefficient estimates are data-coordinated (Hayduk, 1987, p. 139), so looking at the estimates is tantamount to peeking at the data. Hence, four-steppers are not supposed to look at the estimates obtained at prior steps." They then go on to say, "This is problematic if substantial unexpected loadings, or absences of expected loadings, at step-1 are signs of measurement problems. … the four-stepper should not notice this and should not alter the model accordingly, lest the next model test be compromised. So the poor researcher may be damned if he or she does look (via compromising the later test), or doesn't look (by missing the problematic estimates)."

Our position is that if the researcher modifies the model on the basis of seeing the data or the estimates based on the data, then some penalty must be paid in terms of loss of degrees of freedom. There is operationally no difference between rejecting a hypothesis about a parameter and then freeing it to get a value determined by the data, and beginning with a badly fitting starting value in a parameter estimation algorithm and adjusting that value to better fit the data. In both cases a parameter determined by the data should be regarded as a free parameter. Otherwise, any improvement in fit of a modified model is an unaccounted-for researcher-based artifact. Hayduk and Glaser do not indicate that they do this. Furthermore, the focus on loadings in Step 1 is out of place because all rotations of the loadings of unrestricted models will produce equivalent fit, and all unrestricted models are rotationally equivalent. Jöreskog's specification of the unrestricted model makes no binding commitment as to what indicator is or is not related to what factor. The factors of a specific unrestricted model need not correspond to the factors of the measurement model or the structural equation model, although it should be rotationally equivalent to an unrestricted model that is nested within the structural equation model. It is simply a way of specifying an unrestricted model. The commitment about what indicator is or is not related to what factor and to what degree first enters the picture in Step 2 and is completed in Step 4. The four-step procedure is not an exploratory technique but a systematic procedure for evaluating the specified constraints of a given model. Any expected loadings such as Hayduk and Glaser describe should be explicitly specified. The test of their validity will occur in Step 4 if other specified constraints in preceding steps are consistent with the data or have been freed to allow acceptable fit (with corresponding loss in degrees of freedom).

More on Latent Variables With Less Than Four Indicators

In both the target article and Hayduk (1996), Hayduk promoted a strategy for factor model specification that combines the use of relatively few indicators (one to three per factor) with restrictions on both the loading and unique factor variance of one indicator. For example, Hayduk (1996) gave a hypothetical case with three indicators in which (a) one indicator is taken to be the "best" indicator (the "gold standard" of Hayduk & Glaser) of the intended common factor, (b) the loading for that indicator is fixed to one, and (c) the unique factor variance is fixed to a value chosen by the researcher. Parameters for other indicators may also be fixed, but this is not essential. Parts a and b of this strategy are not new because one loading is often fixed for identification purposes. Of greater concern is the use of (c) as an alternative to the use of additional indicators. This strategy has implications not only for the four-step debate, but also for the way SEM researchers think about measurement.

The rationale for (c) is that the researcher fixes the "meaning" of the common factor by fixing the unique variance of the gold standard indicator. Hayduk (1996) suggested that, rather than basing the value for this variance on previous empirical research, the researcher should choose a value that captures the researcher's intended meaning for the factor: “Step three quantifies your assessment of how similar or dissimilar your concept is to the best indicator. This assessment is under your control because you are free to change the meaning of your conceptual fiction (your η), and hence it is you who controls the size of the gap between your concept and any given indicator” (p. 25). Arguing along similar lines, Hayduk (1996) says that “It is my image of η_1, and not the real world (y_1), which determines the proximity between η_1 and y_1.” Fixing the unique variance is equivalent to “adjusting for a specific amount of measurement unreliability.” In this argument, the “meaning” of the factor is defined as the “proximity” between the best indicator and the researcher’s image of that factor. This proximity is chosen in (c) by fixing the unique factor variance, and the meaning of the factor is thus determined. Most importantly, Hayduk (1996) contended that the use of one to three well-chosen indicators under the above strategy is preferable to the use of more indicators in the traditional Step-1 model of Mulaik:

> My experience is that it is tough to get even two indicators of most concepts to cooperate, and rare to find three well-behaved indicators of a concept. Consequently, I discourage my students from using initial models with many sets of multiple indicators. All this does is overlay the diagnostics of indicator problems on top of the diagnostics of structural problems, which leads to confusion, not clarification. (p. 30)

This “experience” seems oddly inconsistent with the position of Hayduk and Glaser, who argue that the Step-1 factor model will typically produce an acceptable fit with too few factors.

We acknowledge that the strategy just described is occasionally useful, especially where multiple indicators are simply unavailable, or where a sufficient research base exists to justify the use of few indicators with fixed parameters. Our objections concern the use of this strategy as a routine approach to building SEM models. We believe that this strategy is ill conceived, for at least two reasons. First, the meaning of the common factor is not really established by fixing the unique factor variance for one indicator. Second, the adoption of this strategy in preference to the use of multiple indicators is likely to degrade, rather than enhance, the interpretation of the factor.

Our view is that Hayduk greatly exaggerates the theoretical significance of the unique factor variance as a determinant of the meaning of the common factor. From basic factor theory, the unique factor variance is composed of variance attributable to measurement error, and also reliable variance that is specific to the indicator. Hayduk (1996) recommended that the unique variance be fixed by considering the percentage of the total indicator variance that one believes should correspond to error. The difficulty is that this percentage is not simply a property of the indicator and the “concept,” but also of the population under study and can be affected by such factors as restriction of range. Unfortunately, the “real world” must play a role here. To take a simple example, a fallible indicator of height will have proportionally greater measurement error variance (and lower reliability) in a population that is nearly homogeneous in actual height than in a more diverse population. In this case, the variation in the reliability of the indicator does not constitute variation in "meaning." If we ignore this phenomenon and simply operate under a fixed percentage of error, we do not thereby achieve a fixed "meaning" for the resulting factor. Instead, we distort the factor structure. The alleged precision that the researcher gains by fixing the unique variance is an illusion.

But early in a research program, rather than trying to establish the meaning of the factor by fixing parameter values, we believe that a better approach is to employ a sufficient number of indicators. We see the question of "meaning" as being equivalent to the question of the construct validity of the indicators: To what extent do these indicators really measure the intended factor? Presumably, the researcher has adopted a factor model because of uncertainty on this question, either due to measurement error, or to the presence of specific variance in the indicators, or both. The hard truth is that the construct validity question is an empirical one that involves the "real world," and it cannot be answered by simply fixing parameter values in the factor model. In the context of factor analysis, the best evidence for the construct validity of the indicators is provided by showing that a substantively rich set of such indicators fit a common factor in real data. By a "rich set," we mean a set of indicators whose content is broad enough to reduce or eliminate competing interpretations for what the factor might be. The strategy promoted in Hayduk (1996) and in Hayduk and Glaser works against this approach by encouraging researchers to rely on few indicators, and by giving the illusion that any failure of interpretation that results can be recouped by "fixing parameter values."

The Criterion Level and Favoring Null

Hayduk and Glaser (2000) argue "that the probability criterion [for a chi-square test] should be set much higher (larger) than .05 to compensate for the researcher's favoring of the test's null hypothesis, rather than the alternative hypothesis" (p. 21). Assuming sample size N is less than 500, Hayduk (1996) recommended using $p >$ .75 as the criterion for accepting one's model (one's null hypothesis). We note that in terms of setting a significance level for accepting or rejecting one's hypothesis, this corresponds to $\alpha = .75$. One has to wonder, considering all the numerous discussions and debates that have taken place among statisticians over significance testing in the past 60 years, why this criterion for accepting a null hypothesis never was advocated nor had it gained acceptance among them. Perhaps it is because this criterion is rationally incompatible with the concept of a difference so large that it is reasonable to regard it as not due to chance, because, under the null hypothesis, differences this large or larger occur only with a small probability, for example, $p \leq$ .05. If $p \leq .75$, corresponding differences this large or larger do not occur with a small probability under the null hypothesis. (See Mulaik, Raju, & Harshman, 1997, for further discussion of significance testing.) Perhaps because Hayduk has been known to conduct SEM studies with small samples, he feels a need to improve power by choosing a larger α. But for most SEM modelers who work with the usual large samples so that they can take seriously the *p* values associated with their chi-square statistics, this sounds like rather poor advice. Their problem is not that they are all too readily inclined to accept the null hypothesis. Almost invariably they reject it, because with large samples the power of the chi-square test is very high just at the point where sample size is large enough to justify the use of *p* values by asymptotic distribution theory. They are detecting all kinds of minuscule discrepancies between their exactly specified models and the data, and hence are invariably rejecting their null hypotheses. To set α at .75 would only make matters worse. There is no implicit favoring of the null hypothesis. On the contrary, there is an implicit tendency to reject it.

Bentler and Bonett (1980) were among the first to explicitly note this problem and to suggest the use of indexes of approximate good fit in addition to chi square (an index of exact fit). Hayduk and Glaser seem to attribute this to James, Mulaik, and Brett (1982) and, in doing so, seemingly have failed to review the literature and to give proper credit for the origin of this practice. In any case, the introduction of indexes of approximate fit reflects a recognition that in large samples the problem of sampling error has given way in importance to the problem of what constitutes an important discrepancy between one's theory and the data. There may be small and probabilistically significant discrepancies between one's model and the data, while the model nevertheless provides a very good approximation to the data. But are these discrepancies important discrepancies? Do they reflect minor systematic errors, minor unrecognized causal influences, like the influence of friction that is often ignored in introductory physics textbooks in developing the theory of a spring? In other words, is one's model essentially correct? Or do the discrepancies reflect real, important clues as to the need to consider a major reformulation of one's theory? There is no universal answer to give to these questions. It all depends on scientific judgment and the current state of knowledge in the given application. Nevertheless, we feel it is important to recognize that obtaining a significant chi square is prima facie evidence that one's model is (provisionally) not exactly true. We do not abandon the chi-square test. We simply note that an index of approximate fit, when high, reflects an approximation. Knowing that something is an approximation implies also knowing that it is not exactly true.

Scientists are frequently content with theories that are merely good approximations when no better theories are available. Giere (1988, p. 190) indicated that the typical physics experiment produces results that are accurate to within 2%. But most physical theories, even good ones, make predictions that are only within 20% of the data and would be rejected by the use of chi-square statistics (which physicists rarely use). Hedges (1987) noted after an extensive review of both the social science and physics literatures that physical theories are no more accurate in general than are social science theories. Social scientists take seriously models that are good approximations, just as physical scientists do. Recognizing they have only good approximations also motivates scientists to seek even better fitting models.

Hayduk and Glaser argue that the use of indexes of close fit do not match the requirements of the four-step. If one has only a good approximation at the first step, does this mean one is only close to the proper number of factors? Chi square may indicate that we do not have the proper number of factors, while an index of close fit indicates we are close to the proper number. They regard it as unreasonable to move to a next step after getting an unacceptable chi square in the first step and criticize the use of indexes of close fit to allow one to do so.

Again we say that Step 1 is not designed to locate the proper number of factors (in an absolute sense); it determines only whether the number of factors hypothesized is “proper,” that is, allows for acceptable fit. Furthermore, getting a good “close fit” does not necessarily mean one is close to the “proper number of factors.” There may be many additional factors, but their contribution to the common variance is small, perhaps confined to doublet factors as well. Whether to continue with the model in the face of a significant chi square but a very good approximation at Step 1 is a scientific judgment call, as we have already indicated. In any case, because the close fit of a Step-1 unrestricted model will be better than any more constrained model in the nested sequence, our recommendation is to set the standard of close fit in Step 1 very high, for example, CFI > .97. Then, by Step 3, one can be content, say, if the CFI drops to no less than .95, which is a criterion of good close fit.

Hayduk and Glaser appear to believe that allowing researchers to use indexes of close fit will contribute to the degeneration of a scientific discipline by encouraging scientists to get by with approximations rather than highlighting the remaining failings of models. We think this is nonsense. With large samples as is typically required, the discrepancies that statistical tests detect are minuscule. So, demanding high degrees of close fit for provisional acceptance of a model in its essential respects will in most cases not lead to an overly soft science. Furthermore, the use of indexes of close fit does not preclude highlighting the failings of models as indicated by chi square and other statistics. There will always be Les Hayduks around to critique the work of individuals who are excessively lax. Indexes of close fit are not replacements for chi square; only supplements to it.

RMSEA

Hayduk and Glaser appear to wish to identify Mulaik with the RMSEA and then to critique its use, perhaps as a way of undermining further Mulaik’s position. But Mulaik has no strong commitment to the RMSEA. He is not an author of the index, although he highly respects the individuals who are. Nevertheless, he acknowledges the current popularity of the RMSEA as an index of approximate fit, and this was why he suggested it could be used with the four-step procedure. Other than these comments, we will leave discussion of Hayduk and Glaser's critique of the RMSEA index to James Steiger, who is an author of the index.

CONCLUSIONS

Hayduk and Glaser's concerns that the four-step procedure is designed to separate measurement from structure during model assessment are off target. We agree that measurement issues are encountered at all steps. This is a debate Hayduk wants to make with others who make such claims, and we leave him to debate this with them. What the four-step procedure allows one to do is to separate the respective constraints within a structural equation model, and then systematically test them in a natural order that is implied by the structure of the model itself and the distinction between manifest and latent variables. This allows one to a degree to isolate the sources of lack of fit among the constraints of one's model.

Hayduk and Glaser fail to appreciate the context in which the four-step procedure is used, where conceptual analysis leads to identification of latent concepts to study and these in turn are synthesized into a causal model, and multiple manifest indicators of each of the latent concepts are then carefully selected or constructed. This justifies testing a confirmatory common factor analysis model in connection with the indicators after careful consideration is given in indicator construction to ruling out known alternative causal relations among the indicators. But, of course, the subsequent tests are not infallible against these possibilities. Nor can there be any such tests.

We make more explicit the advantages of performing studies with four or more indicators per latent variable: They increase degrees of freedom, provide empirical support for the independent existence of the latent variable and its properties, demonstrate one has a clear and replicable conception of what is involved in selecting or constructing indicators of the latent variable, improve estimates of parameters, and make possible the detection of extraneous variables that may be confounded with the latent variables under study and thereby might lead the researcher to draw incorrect inferences as to causal connections between the latent variables.

Hayduk and Glaser's argument that the fundamental problem of the four-step procedure is its inability to ascertain in Step 1 whether the researcher has the proper number of latent concepts is really an argument that can be used against any test of a model based on assessing whether it fits the data or not. Their argument can even be used against their own use of tests of fit with structural equation models that are evaluated without going through a multistep procedure. They effectively complain that, when the test reports acceptable fit, the model may still be false. Well, this is not news to most experienced researchers who use tests of fit and treat acceptable fit as only provisionally authorizing acceptance of a model. Step 1 merely ascertains whether for the number of latent concepts hypothesized a common factor model for the same number of concepts would fit or fail to fit the covariance matrix and thereby authorize provisional rejection or acceptance of the hypothesized number of factors. We also show that the concept of "the proper number of concepts" is problematic even in Hayduk and Glaser's example of the simplex model, for which they confidently feel they have ascertained the proper number of concepts.

Hayduk's (1996) recommendation to use a significance level of .75 in chi-square tests to avoid favoring the null hypothesis is shown to be incoherent with a conception of differences not due to chance because differences that large or larger occur under the null hypothesis with low probability. We also show that there is no implicit favoring of the null hypothesis in usual structural equation model research that uses large samples to enable researchers to perform statistical inferences. In fact, it is the other way around: The tendency is invariably to reject the null hypothesis.

We have also exposed other misconceptions of the four-step procedure held by Hayduk and Glaser. In conclusion, the four-step procedure has easily survived their critique.

REFERENCES

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411–423.

Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability: Vol. V* (pp. 111–150). Berkeley: University of California Press.

Bast, J., & Reitsma, P. (1997). Matthew effects in reading: A comparison of latent growth curve models and simplex models with structured means. *Multivariate Behavioral Research, 32,* 135–167.

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

Bollen, K. A., & Jöreskog, K. G. (1985). Uniqueness does not imply identification: A note on confirmatory factor analysis. *Sociological Methods and Research, 14,* 155–163.

Browne, M. W. (1992). Circumplex models for correlation matrices. *Psychometrika, 57,* 469–497.

Carlson, M., & Mulaik, S. A. (1993). Trait ratings from descriptions of behavior as mediated by components of meaning. *Multivariate Behavioral Research, 28,* 111–159.

Giere, R. N. (1988). *Explaining science*. Chicago: University of Chicago Press.

Glymour, C., Scheines, R., Spirtes, P., & Kelly, K. (1987). *Discovering causal structure.* New York: Academic.

Harman, H. (1960). *Modern factor analysis*. Chicago: University of Chicago Press.

Hart, B., & Spearman, C. (1913). General ability, its existence and nature. *British Journal of Psychology, 5,* 51–84.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL.* Baltimore: The Johns Hopkins University Press.

Hayduk, L. A. (1996). *LISREL issues debates and strategies.* Baltimore: The Johns Hopkins University Press.

Hayduk, L. A., & Glaser, D. N. (2000). Jiving the four-step, waltzing around factor analysis, and other serious fun. *Structural Equation Modeling, 7,* 1–35.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? *American Psychologist, 42,* 443–455.

Howe, W. G. (1955). *Some contributions to factor analysis* (Rep. No. ORNL–1919). Oak Ridge, TN: Oak Ridge National Laboratory.

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

Jones, M. B. (1959). *Simplex theory*. Monograph Series No. 3. Pensacola, FL: U.S. Naval School of Aviation Medicine.

Jones, M. B. (1960). *Molar correlational analysis*. Monograph Series No. 4. Pensacola, FL: U.S. Naval School of Aviation Medicine.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–133.

Jöreskog, K. G. (1969/1979a). A general approach to confirmatory maximum likelihood factor analysis, with addendum. In K. G. Jöreskog & D. Sörbom, *Advances in factor analysis and structural equation models* (J. Magidson, Ed., pp. 21–43). Cambridge, MA: Abt Books. Originally published without addendum as, Jöreskog, K. G. (1969). A general approach to confirmatory maximum likehood factor analysis. *Psychometrika, 34,* 183–202.

Jöreskog, K. G. (1979b). Structural equation models in the social sciences: Specification, estimation and testing. In K. G. Jöreskog & D. Sörbom, *Advances in factor analysis and structural equation models* (J. Magidson, Ed., pp. 105–127). Cambridge, MA: Abt Books.

Jöreskog, K. G. (1979c). Analyzing psychological data by structural analysis of covariance matrices. In K. G. Jöreskog & D. Sörbom, *Advances in factor analysis and structural equation models* (J. Magidson, Ed., pp. 45–100). Cambridge, MA: Abt Books.

Kabe, D. G. (1963). Stepwise multivariate linear regression. *Journal of the American Statistical Association, 58,* 770–773.

Kaplan, D. (1989). A study of the sampling variability and z-values of parameter estimates from misspecified structural equation models. *Multivariate Behavioral Research*, *24,* 41–57.

Kaplan, D. (1990). Evaluating and modifying covariance structure models: A review and recommendation. *Multivariate Behavioral Research, 25,* 137–155.

Ledermann, W. (1937). On the rank of the reduced correlation matrix in multiple-factor analysis. *Psychometrika, 2,* 85–93.

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin, 100,* 107–120.

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

Marsh, H. (1993). Stability of individual differences in multiwave panel studies: Comparison of simplex models and one-factor models. *Journal of Educational Measurement, 30,* 157–183.

Marsh, H. W., & Bailey, M. (1991). Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement*, *15,* 47–70.

Marsh, H., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much: The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181–220.

McCrae, R. R., Zonderman, A. B., Costa, P. T., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analyses versus Procrustes rotation. *Journal of Personality and Social Psychology, 70,* 552–566.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58,* 525–543.

Michell, J. (1990). *An introduction to the logic of psychological measurement.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Mulaik, S. A. (1972). *The foundations of factor analysis*. New York: McGraw-Hill.

Mulaik, S. A. (1987). A brief history of the philosophical foundations of exploratory factor analysis. *Multivariate Behavioral Research, 22,* 267–305.

Mulaik, S. A. (1990, June). *An analysis of the conditions under which the estimation of parameters inflates goodness of fit indices as measures of model validity.* Paper presented at the Annual Meeting, Psychometric Society, Princeton, NJ.

Mulaik S. A. (1995). The metaphoric origins of objectivity, subjectivity and consciousness in the direct perception of reality. *Philosophy of Science, 62,* 283–303.

Mulaik, S. A., & James, L. R. (1995). Objectivity and reasoning in science and structural equation modeling. In R. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications.* Thousand Oaks, CA: Sage.

Mulaik, S. A., & Quartetti, D. G. (1997). First-order or higher-order general factor? *Structural Equation Modeling, 4,* 193–211.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and a place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65–115). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Reiersøl, O. (1950). On the identifiability of parameters in Thurstone's multiple factor analysis. *Psychometrika, 15,* 121–149.

Roy, J. (1958). Step-down procedure in multivariate analysis. *Annals of Mathematical Statistics, 29,* 1177–1187.

Schönemann, P. H. (1970). Fitting a simplex symmetrically. *Psychometrika, 35,* 1–21.

Schouls, P. A. (1980). *The imposition of method: A study of Descartes and Locke.* Oxford: Clarendon.

Tepper, K., & Hoyle, R. (1996). Latent variable models of need for uniqueness. *Multivariate Behavioral Research, 31,* 467–494.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

APPENDIX

Let Λ_0 be the principal axes factor pattern matrix of an identified maximum likelihood exploratory factor analysis solution. (For this same solution the factor correlation matrix $\mathbf{\Phi}_0 = \mathbf{I}$ because the principal axes solution involves orthogonal factors with unit variances.) Let Λ_u be a rotated factor pattern matrix $\Lambda_u = \Lambda_0\mathbf{T}$ constrained to have zeros in the locations that are fixed zeros in the pattern matrix of an unrestricted solution, where $\mathbf{T}$ is a linear transformation matrix, constrained so that diag $\mathbf{\Phi}_u$ = diag$[\mathbf{T}^{-1}\ \mathbf{T}^{-1\prime}] = \mathbf{I}$, where Φ_u is the variance–covariance matrix of the rotated factors. The matrix $\mathbf{T}$ may be constructed indirectly from a transformation matrix $\mathbf{L}$ used to transform Λ_0 to a factor structure matrix $\mathbf{V}_u$, that is, $\mathbf{V}_u = \Lambda_0\mathbf{L}$, with zeros in the same corresponding positions as in Λ_u, where $\mathbf{L}$ must be constrained so that diag$(\mathbf{L}'\mathbf{L}) = \mathbf{I}$. By well-known common factor theory (Mulaik, 1972; Thurstone, 1947) $\mathbf{T}$ will be given by $\mathbf{T} = (\mathbf{L}'\mathbf{L})^{-1}\mathbf{D}^{-1}$, where $\mathbf{D}^2 = \text{diag}[(\mathbf{L}'\mathbf{L})^{-1}]$. Let $\mathbf{L}_s$ denote the sth column of $\mathbf{L}$, and $\mathbf{V}_{us}$ the sth column of $\mathbf{V}_u$, then $\mathbf{V}_{us} = \Lambda_0\mathbf{L}_s$ and $\mathbf{L}_s$ has the property $\mathbf{L}_s'\mathbf{L}_s = 1$. Let us now focus on the zero elements of the unrestricted solution in the column $\mathbf{V}_{us} = \Lambda_0\mathbf{L}_s$. Let Λ_{0s} denote the $(k-1) \times k$ matrix formed from the rows of

Λ_0 that when postmultiplied by the column vector $\mathbf{L}_s$ yield, respectively, the zero elements of the unrestricted solution in the column $\mathbf{V}_{us}$, that is, $\Lambda_{0s}\mathbf{L}_s = \mathbf{0}$. From this we may write $\mathbf{L}_s' \Lambda_{0s}'\Lambda_{0s}\mathbf{L}_s = 0$. The unit length vector $\mathbf{L}_s$ with this property is the eigenvector corresponding to the zero eigenvalue of the less than full rank $k \times k$ matrix $\Lambda_{0s}' \Lambda_{0s}$. This matrix, furthermore, should have no more than 1 zero eigenvalue, otherwise the solution is not unique. (This corresponds to condition c, discussed in the article, which requires that the matrix Λ_s have rank $k - 1$.) This then provides a solution for the *s*th column of **L** and in turn for each of the respective columns of **L**, and ultimately of **T**. Because the maximum likelihood solution of the unrestricted model (if identified) maximizes the likelihood of that model and any other model rotationally equivalent to it, it must have equivalent fit to the exploratory common factor analysis solution for the same number of factors.