SUBSCRIBER INFORMATION

al *Equation Modeling: A Multidisciplinary Journal* is published four times a year.
tions for the 2000 volume are available on a calendar-year basis only. In the United
d Canada, per-volume rates are US $45.00 for individuals and US $315 for institu-
other countries, per-volume rates are US $75.00 for individuals and US $345 for in-
s. Send subscription orders, information requests, and address changes to the Jour-
cription Department, Lawrence Erlbaum Associates, Inc., 10 Industrial Avenue,
., NJ 07430–2262. Address changes should include the mailing label or a facsimile.
or missing issues cannot be honored beyond 4 months after mailing date. Duplicate
annot be sent to replace issues not delivered due to failure to notify publisher of
of address.

electronic version of this journal is available through OCLC (Online Computer Li-
nter) at no additional cost (excluding OCLC's access charges) to all 2000 institutional
scribers. The 2000 electronic-only subscription rate is $283.50. To set up electronic
this journal or to obtain more information, please contact OCLC at 1–800–848–5878
2); E-mail: brownchr@oclc.org or visit http://www.oclc.org/oclc/eco/main.htm

journal is abstracted or indexed in *Current Index to Journals in Education; Current*
Statistics; Educational Technology Abstracts; PsycINFO/Psychological Abstracts;
al Research Methodology Abstracts. Microform copies of this journal are available
Bell & Howell Information and Learning, P.O. Box 1346, Ann Arbor,
06–1346.

right © 2000, Lawrence Erlbaum Associates, Inc. No part of this publication may be
any form or by any means, without permission of the publisher. Special requests for
on should be addressed to Permission Department, Lawrence Erlbaum Associates,
ndustrial Avenue, Mahwah, NJ 07430–2262. Printed in the United States of Amer-
N 1070–5511.

STRUCTURAL EQUATION MODELING, 7(1), 1–35
Copyright © 2000, Lawrence Erlbaum Associates, Inc.

Jiving† the Four-Step, Waltzing Around Factor Analysis, and Other Serious Fun

Leslie A. Hayduk

Department of Sociology
University of Alberta

Dale N. Glaser

Sharp Health Care, and
San Diego State University

It has been proposed that a clear separation of measurement from structural reasons for model failure can be obtained via a procedure testing 4 nested models: (a) a factor model, (b) a confirmatory factor model, (c) the anticipated structural equation model, and (d) possibly, a more constrained model. Advocates of the 4-step procedure contend that these nested models provide a trustworthy way of determining whether one's model is failing as a result of structural (conceptual) inadequacy, or as a result of measurement misspecification. We argue that measurement and structural issues can not be unambiguously separated by the 4 steps, and that the seeming separation is incomplete at best and illusory at worst. The prime difficulty is that the 4-step procedure is incapable of determining whether the proposed model contains the proper number of factors. As long as the number of factors is in doubt, measurement and structural assessments remain dubious and entwined. The assessment of model fit raises additional difficulties because the researcher is implicitly favoring of the null hypothesis, and the logic of the root mean square error of approximation (RMSEA) as a test of "close fit" is inconsistent with the logic of the 4-step. These discussions question whether factor analysis can dependably determine the proper number of factors, and argue against the routine use of .05 as the probability target for structural equation model chi-square fit.

Requests for reprints should be sent to Leslie A. Hayduk, Department of Sociology, University of Alberta, Edmonton, Alberta, Canada T6G 2H4. E-mail: Lhayduk@ualberta.ca

†To jive—1: to perform specific animated dance steps; 2: to tease or cajole.

Structural equation modeling emerged at the confluence of a factor analytic tradition from psychology and a path analytic tradition from sociology. LISREL's basic equations were set up to reflect a measurement model connecting the concepts to their indicators, following the factor analytic tradition, and a structural model permitting directed effects among the concepts, following the path analytic tradition (Jöreskog & Sörbom, 1976, pp. 1–2). Current structural equation programs unavoidably distinguish between observed variables and latent variables (factors or concepts), but they generally do not mandate the separate or sequential investigation of the measurement and structural segments of a model. In contrast, the four-step approach to structural equation modeling highlights the measurement-structural distinction and attempts to assess measurement prior to structure.

Early proponents of separating the measurement segment of a model (which links the concepts to the observed indicators) from the structural segment of the model (which links the concepts or factors to one another) included James, Mulaik, and Brett (1982), Burt (1973, 1976), and Herting and Costner (1985). Not everyone agreed with the suggestion to estimate a measurement model prior to a structural model (Hayduk, 1987, pp. 118–123), but the disagreements were not framed in a way that permitted a focused assessment.

Anderson and Gerbing (1988) explicitly recommended estimating a measurement (factor) model prior to estimating the structural model, but this two-step approach was countered by a critique by Fornell and Yi (1992a), and comments were exchanged (Anderson and Gerbing, 1992; Fornell and Yi, 1992b). The two-step approach survived, and was mildly endorsed by Jöreskog (1993, p. 297) and Jöreskog and Sörbom (1993, pp. 113, 128). Hayduk (1996, pp. 36–78) stirred the embers of the debate by reviewing the prior exchanges and adding several new points as fuel for the fire. He concluded by recommending against routine use of the two-step procedure.

The glowing embers of the controversy ignited into flames when Les Hayduk signed onto the SEMNET discussion group on the Internet and invited discussion of his 1996 book. Unbeknown to Les, Stan Mulaik, a factor analysis authority, and long-time advocate of the multistep, was one of the stalwarts of SEMNET. It was not long before hundreds of pages of discussion had floated around the Net.[1] The debate paused for a while when Les withdrew due to other commitments and Stan was left sighing for not having convinced the opposition. George Marcoulides encouraged Les to prepare an encapsulated version of the SEMNET exchanges. This article gradually emerged parallel to the SEMNET exchanges spanning the fall of 1997 and spring of 1998.

In the following article, we focus on Stan Mulaik's way of doing the four steps.[2] We attempt to preserve both the core and flavor of the SEMNET discussions, though we move considerably beyond the SEMNET discussion when considering the root mean square error of approximation (RMSEA) (Browne & Cudeck, 1993) as a four-step model test. We begin by summarizing the four steps, and follow this with: a listing of the limitations on the use of the four-step, a discussion of the four steps' most fundamental problem, and a consideration of the test and criteria to be used in moving between the four steps.

THE FOUR STEPS

The four-step procedure attempts to disentangle the reasons for a failing structural equation model by separating measurement failure from structural/conceptual failure.[3] The researcher begins with a base model that reflects the researcher's best guess at the nature of the forces operative in the real world. The researcher constructs three other models from this base model by relaxing the constraints implicit in the base model or by adding constraints to that base model.

For any given base model, a less restricted model can be created by replacing the directed effects between the concepts with a full set of free correlations/covariances between the concepts. This replaces the structural model, which was a mixture of directed effects, asserted null effects, and correlations among the exogenous variables, with a full set of undirected covariances/correlations between all the concepts (see Figure 1). An even less restricted model is obtained by also adding loadings (lambda's) to saturate the connections between the concepts and the indicators, so that each indicator might potentially load on any or all of the concepts.[4] The final type of model begins from the base model and adds planned restrictions (anticipated constraints)—rather than reducing the model constraints.

[1] The Internet site http://www.gsu.edu/~mkteer/semnet.html provides information on how to join the SEMNET discussion group and on how to search the SEMNET log of past exchanges/postings.

The initial exchanges between Les and Stan were instigated by Dale and occurred between March 20 and May 17, 1997. A flurry of exchanges occurred after Stan's posting of 8:23 p.m. on 9/26/97. The discussion died out in May 1998.

Reference to the SEMNET discussion will be made by author, date, and occasionally by time and approximate position in long postings. Hayduk 3/20/97 5:30 p.m.-80% refers to comments about 80% of the way through Les Hayduk's long 5:30 p.m. posting on March 20. Times may vary a bit because submit, send, and receipt times are not always identical.

We take the liberty of occasionally addressing Stan and Les by their first names to preserve the spirit of the friendly academic jousting that pervaded the SEMNET exchanges.

[2] This differs somewhat from Anderson and Gerbing's (1988, 1992) way of doing the two-step, which was the focus of Hayduk (1996, chap. 2).

[3] Mulaik 3/20/97 30%, 60%; 4/3/97; 4/4/97; 4/7/97 95%; 4/14/97 1:43 a.m. 80%; 3/23/97 3:50 p.m.; 1/24/98 12:32 p.m.; 3/25/98 1:07 a.m.

[4] Stan Mulaik recommended permitting all indicators to load on all the concepts but with the exception that for each concept there should be one indicator that loads only on that concept (Mulaik 4/4/97 1:12 a.m.). Stan did not make this a mandatory feature of doing the four-step, so we treat this as "optional."

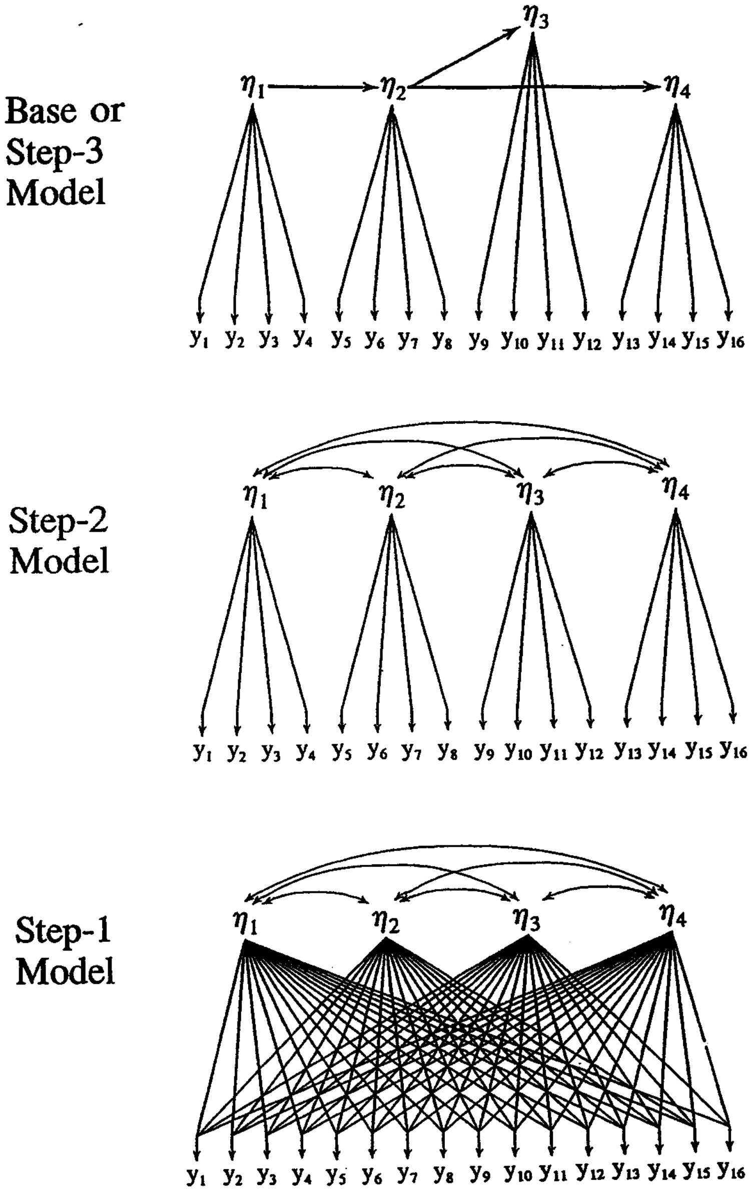
Base or Step-3 Model
η_1
η_2
η_3
η_4
y_1 y_2 y_3 y_4 y_5 y_6 y_7 y_8 y_9 y_{10} y_{11} y_{12} y_{13} y_{14} y_{15} y_{16}
Step-2 Model
η_1
η_2
η_3
η_4
y_1 y_2 y_3 y_4 y_5 y_6 y_7 y_8 y_9 y_{10} y_{11} y_{12} y_{13} y_{14} y_{15} y_{16}
Step-1 Model
η_1
η_2
η_3
η_4
y_1 y_2 y_3 y_4 y_5 y_6 y_7 y_8 y_9 y_{10} y_{11} y_{12} y_{13} y_{14} y_{15} y_{16}

FIGURE 1 Constructing the Step-1 and Step-2 models from the base model (Step 3).

This model is used only if the base model and the less constrained models function acceptably.

The four-step procedure recommends a sequential testing of the models with the least restricted model being tested first. The first model tested (the Step-1 model) has as many loadings as possible and is saturated with correlations (covariances) among the concepts (factors). This model resembles a factor model employing an oblique rotation. The second model (the Step-2 model) includes only the anticipated loadings, namely the loadings included in the base model. The loadings that had been added merely to create the first-step model are now fixed at zero, but the saturated covariances among the concepts are retained. This makes the second model a confirmatory factor model in that specific items are located under specific factors/concepts, while the factors remain interconnected by all their possible correlations.

The third model replaces the factor correlations/covariances with the anticipated directed effects and specified absences of effects, among the concepts, though some correlations among the exogenous concepts may remain. This Step-3 model is the base model that is the researcher's prime focus.

So saturating the concepts' interrelations with correlations or covariances takes us from the base (Step-3) model to the Step-2 model, and adding loadings spanning the item range gets us to the Step-1 model.

If all these models have gone well, a fourth step might place additional constraints on the base model to examine whether some of the parameters in the third-step base model are unnecessary, equal, or significantly different from specific values. Step 4 may address questions of substantial research interest, but the variability in the types of constraints entered to get from the Step-3 to the Step-4 model (fixed zero, fixed nonzero, equality or proportionality constraints, in any number and combination) renders Step-4 models too diverse for useful generalizations, so our comments focus on the first three of the four steps.[5]

These steps are intended to provide a nested series of models, where the nesting is supposed to pinpoint problems with the base model. If the base model is the proper model, the first three steps should all proceed smoothly and with acceptable model fit. Models fail only when they impose unwarranted constraints, and because the models used at the first and second steps are less constraining than the third step (base model), these models should all fit well if the base (third-step) model is the correct model.

If the base model is incorrect, the nature of the failings in the base model are supposed to be located as purely structural if the Step-1 and -2 models fit while the Step-3 model fails. The failings are supposed to be failures of measurement if the

[5] Steps 2, 3, and 4 are the James, Mulaik, and Brett (1982) three-step approach, and Step 1 is a straightforward extension to an exploratory factor analysis model. On what the steps are, see Mulaik 3/20/97 25%, 3/23/97 3:50 p.m., and Hayduk 3/27/98.

Step-1 model fails, or if both the Step-1 and -2 models fail. Failure of the first-step model is supposed to indicate the model contains too few concepts, and hence the measurement of the concepts is called into question. Fit at Step-1 followed by failure at Step-2 is supposed to indicate that one has the proper number of concepts, but that some adjustment to the identity of the concepts, via the addition of loadings, is required.

If at any step the model fails to fit the data, the researcher should stop and not go on to the later steps until the problems with the current model are resolved.[6] The later models are more constrained than the current, and now failing, model so all the later models would also fail.

So the researcher loosens the structural constraints by replacing the anticipated directed and null effects among the concepts in the base model with a full set of free concept covariances, and thereby comes to know whether the structural model is problematic. Similarly, one loosens the constraints on the measurement model, by replacing the theoretically specified loadings of the indicators on the concepts with a full set of free loadings, thereby locating whether the measurement segment of the model is problematic.

Think of the above as a medicine man's sales pitch. Before we buy the potion we should ask whether we are suffering from an illness the potion is supposed to cure, and whether there is evidence of any curative power for whatever illness the potion does claim to cure. So, for what types of models might the four-step procedure not be appropriate, and among those where it is claimed appropriate, can it do what it claims it is able to do?

RANGE OF APPLICABILITY

First, the four-step procedure is appropriate only for researchers having a base model. If one is merely exploring, the procedure can't be used. Similarly, if one has a base model that is saturated with effects among the concepts, the procedure will not be able to separate measurement from structural concerns because the Step-2 and Step-3 models will provide identical fit statistics. The researcher must start with a real and nontrivial structural model.

Second, to do things Stan's way, there must be at least four indicators for each of the concepts in the model.[7] This means that the four-step procedure cannot be used if the model contains variables like sex, age, income, or education, or any concept with a single indicator. This is a severe restriction on the applicability of the four-step procedure because most models contain at least one concept having fewer than four indicators.

[6] Mulaik 3/20/97 15%, 4/22/97 9:20 p.m. and 3/23/97 3:50 p.m.

[7] Mulaik 3/20/97 9:35 p.m. 40%, 3/24/97 3:05 p.m. 15%, 4/14/97 1:43 a.m. 80%, 5/17/97 5:42, 2/22/98 9:41 p.m. See also James, Mulaik, and Brett (1982, p. 164).

Why does Stan make "at least four indicators per concept" a mandatory part of the four-step process? He feels the meaning of a concept is unclear until that concept has at least four indicators.[8] This also helps to ensure that the Step-1 model is identified. The coefficients in a model containing a single factor with three indicators are identified, while a single factor with four indicators provides both identified measurement coefficients and some testing power. Hayduk (1996, pp. 25–30) discusses a procedure that results in specified meanings with single indicators, and identified loading estimates and measurement tests with two indicators, so we believe that Stan is being overly restrictive in demanding four indicators per concept.[9] But rather than trying to persuade Stan of this in the current context, we will obligingly accept Stan's minimum of four indicators per concept, and thereby exclude many models from ever being subjected to the four-step procedure.

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 practicable and if the model is to be estimated using a reasonable sample size. Using many indicators unintentionally, but importantly, limits the structural sparseness that the model can display, and structural sparseness is a feature cherished by both Stan and Les.[10] If there are only two concepts in a model, a single structural effect

[8] Mulaik 3/20/97 35%, 4/16/97 3:35 a.m. 20%, 4/16/97 5:55 p.m., 5/17/97 5:42 p.m., and 1/24/98 12:32 p.m.

[9] See Hayduk's 4/15/97 5:36 p.m. 90% discussion of how a single variable like sex can correspond to different concepts (sex as chromosomal complement, as hormonal levels, as label applied by others, as self-identification, and so on) and how one can differentiate between these concepts with different error specifications. Hayduk (1996, pp. 25–30) argues that by asserting a clear meaning for the first indicator (via assigning a fixed 1.0 loading and a fixed measurement error variance), while the second and subsequent indicators are provided free loadings and error variances, one achieves some testing ability (assuming the model has more than a single concept).

The coefficients for second and later indicators are estimable under this procedure. So in order to defend the "requirement" of four indicators per concept, Stan is in a position of having to argue that uniquely identified measurement estimates are somehow possible despite ambiguous or insufficiently determined conceptual meaning. Stan must also explain why or how conceptual meaning can remain ambiguous despite the contributions of the second and third indicators to the model chi-square and its degrees of freedom. We see no justification for such claims, and hence feel that two indicators are sufficient to initiate the testing of the conceptual and measurement structures if the first indicator has been modeled with the fixed loading and error variance procedure recommended by Hayduk (1996). The first indicator asserts a specific meaning, and all subsequent indicators test the asserted meaning.

Stan and Les agree that use of scales summing several items clustering under one factor is unadvisable (Hayduk 3/20/97 5:47 p.m. 95%, Mulaik 3/20/97 9:35 p.m. 95%), so scales are not a point of contention between us.

[10] Sparseness comes from making potentially "falsifiable" assertions of null effects between concepts, zero-error variances and covariances, proportionality constraints, and fixed nonzero coefficients. Les and Stan agreed that, from the perspective of checking out the theoretically postulated model, such assertions are at least as important to a theory as are the theory's predicted effects (Mulaik 3/31/97; Hayduk, 3/31/97, 4/6/97 60%); see Hayduk (1996, pp. 7–19).

linking those two concepts saturates the structural part of the model.[11] It takes at least two effects to interconnect three concepts, so a researcher can have only one degree of structural freedom with three theoretically connected concepts.[12] It takes three effects to interconnect four concepts, so models with four connected concepts have at most three degrees of structural freedom. Models with five or six concepts can have reasonable structural sparseness, if the connections between the concepts are minimal, but most models do not have minimal interconnections, so in practice one is usually confronting models with seven or more concepts before there is a substantial degree of structural sparseness.

With 10 indicators per concept, models with 7 or 8 concepts are big, and they require substantial sample sizes to support them. This leaves four-step advocates desirous of structural sparseness with a difficult choice. Either they can strive for relatively few indicators per concept, that is, stay close to four indicators per concept, or doom themselves to collecting huge samples and building huge models in order to gain even a moderate degree of structural–theoretical sparseness. There is little sense in stressing the four-step's ability to separate measurement from structural concerns if the structural part of the model is nearing saturation.

THE AILING CURATIVE POWERS OF THE FOUR-STEP

Locating the Proper Number of Factors/Concepts

Can the first of the four steps determine whether or not the base model contains the proper number of concepts? Both the measurement and structural parts of a model are fundamentally wrong if the base model contains the wrong concepts, so there is little sense in attempting to separate measurement from structural problems (the main goal of the four-step) until one has located the proper number of concepts.

It is the first of the four steps that must determine the number of concepts/factors, because all the later steps become untrustworthy if they are contaminated by inclusion of the wrong concepts. This leads to an obvious question: Does the fit or failure of the first-step model report on whether the model contains the proper number of factors/concepts? Stan repeatedly asserted that Step 1 is capable of doing this, though his resolve fluctuated during the spring of 1998.[13] For a variety of reasons, we continue to doubt whether Step 1 can be trusted to locate the proper number of concepts.

The first way of questioning this arises from noting that several factor models can all fit to a single data set. Imagine a model for which multiple indicators are responsive to four underlying concepts in the real world. Imagine also that we try to fit Step-1 (ordinary factor) models with 1, 2, 3, 4, 5, or 6 factors to the observed covariance data. The models with one, two, or three factors should fail because they are unable to reproduce the input covariance matrix. The four-factor model should adequately reproduce the covariance matrix, and hence should fit. The factor models with five or six factors should also reproduce the covariance matrix and hence should also fit.[14]

This creates a problem for the four-step. If the base model, namely the researcher's best guess at the real world, included four or five or six factors, the corresponding Step-1 models would fit even though only one of these could contain the proper number of factors. So the researcher would be unjustifiably proceeding to the later steps in two of the three instances. If the base model postulated five concepts, and if a model with five concepts fits at Step 1, how can we move ahead assured that we have the proper number of factors, given that Step-1 models with four and six concepts can also fit? Because there is only one possible correct num-

[11] With only two concepts, the researcher is forced to choose between theoretically disconnected concepts (no causal connection between them), or a single connection that permits the concepts to appear as "connected parts of some theory" but that also saturates the structural part of the model and renders it nontestable. Saturated structural models are not testable because they provide the same fit as does a full set of covariances among the concepts. This makes the fit of the Step-2 and Step-3 models identical, and hence uninformative.

A saturated structural model might make it seem like the only kind of error is measurement misspecification, but this is not the case. If the true structural model is composed of a sparse set of effects among more than the modeled number of concepts, there is both measurement and structural misspecification despite the saturated structural model. This is explored more fully in this article in the discussion of the proper number of concepts.

[12] Three concept covariances minus two estimated effects leaves one degree for freedom, from the fixed null effect, with which to test the structural model. We ignore possible equality constraints as these do not alter the gist of the point being made here.

[13] "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 indictors" (Mulaik 3/23/97 3:50 p.m.). And, "what you can test in an exploratory factor analysis is the number of factors" (Mulaik 3/27/97 11:28 p.m.) and "my first step test of the number of factors" (Mulaik 4/11/97 10:35 p.m.). "How would factor analysis lead you to have the wrong number of concepts at Step-1? You are using factor analysis to test your hypothesis about the number of factors or latent variables at this step" (Mulaik 4/4/97 1:12 p.m.).

In considering fit at Step 1, fail at Step 2, should one go back and try a model with more factors instead of the number that did work with an exploratory factor analysis that lets each factor span all the items? Stan thought, no, "... 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" (Mulaik 4/4/97 1:12 a.m.).

See also 4/14/97 1:43 a.m., p. 7, and the discussions between 4/18/98 and 2/27/98, on the hypothesis for Step 1.

[14] Stan's postings of 4/7/97 5:16 p.m. and 2/22/98 9:41 p.m. permit that any number of factors above the minimum should fit, up to the number that lead to an underidentified model. He also permits that the researcher may have hypothesized some number other than the minimum. Using progressively more factors would ultimately lead to an underidentified Step-1 model (Mulaik 2/22/98 9:41 p.m.) with the highest number of identified factors, depending on the number of observed variables and other things. But our concern is not with the location of the identification limit. We are questioning the logic of the testing process for whatever models are below that limit, wherever it might be (Hayduk 2/20/98 2:09 p.m.).

ber, any Step-1 test permitting acceptable fit with more than one number of factors is problematic for the four-step.[15]

One way to solve this problem is to demonstrate that the minimum fitting number of factors is the proper number of concepts. Stan began by exploring, then rejecting, then wavering, on the idea that the minimum number is likely to be the proper number.[16] Starting with one factor and then adding one factor at a time until the first good fit is achieved, or starting with many factors and then subtracting one factor at a time until just prior to failure, locates the minimum number of factors. Unfortunately, there is nothing that assures us that the minimum number is the proper number of factors.[17]

In fact, we suspect the minimum number of factors is likely to be the wrong number. This is because the coefficients freed in moving from the base model (the researcher's best guess at the real-world "model") to the Step-1 model, can work as fitness–Band-Aids that hide the bleeding (ill fit) created by cutting out a factor or two. A real-world "model" with *k* sparsely connected concepts can be fit by a factor model with fewer than *k* factors if that factor model contains a bunch of coefficients not required by the real world—recall the many coefficients added to get from the base model to the Step-1 model. The estimates of the unnecessary coefficients are free to be adjusted to counteract whatever ill fit was created by amputating a factor or two. A factor model with too few factors can provide an adequate fit if it is stitched together with a full set of free concept covariances (Step 2), and bandaged with a set of free factor loadings (Step 1) that keep us from seeing the bleeding model.[18]

There is no available proof that guarantees that the minimum number of fitting factors locates the true number of concepts *if the true model is of any of the thousands of styles of nonfactor "models" that might comprise the real world, hopefully, but not necessarily, represented by the base model.*[19] Stan provided some factor analysis mathematics as proof that Step 1 locates the proper number of factors, so we feel obliged to address why this is unconvincing. The basic deficiency is not in the mathematics[20] but in a key assumption that goes into the mathematics, namely that the true or proper model underlying the analysis is a factor model. That is, the math assumes that the only thing that needs to be *tested* is the number of factors to include, not whether the factor model is the proper model to use.[21] This point resurfaced when Stan[22] quoted Jöreskog on the chi-square test for the Step-1 factor model:

> In the preceding sections we have treated the problem as if the number of common factors *k* was given in advance. ... If it is possible to specify a certain value k_0 we could test the hypothesis $k=k_0$ against the alternatives $k>k_0$. ... *If the model is proper for the relevant data* and if the hypothesis is true, then the p-k_0 smallest latent roots of Σ^* are equal. (Jöreskog, 1962, p. 347, italics added)

[15] See Hayduk 5/14/97 12:34 p.m., and the exchanges of 5/16/97.

[16] Contrast Mulaik 3/24/97 3:05 p.m. and 3/26/97 4:37 p.m. with Mulaik 4/16/97 3:35 a.m. 70% and 90%, 4/16/97 5:21 p.m., 4/17/97 3:40 p.m., 4/17/97 11:55 p.m., and 4/28/98 12:47 a.m.

[17] Parsimony has been used as a justification for use of few factors, but models with many more, but sparsely connected, factors can be as sparse as ordinary loading–obese factor models, so some other defense is required (Hayduk 4/5/97, 11:19 p.m. 40%, 2/24/98 2:59 p.m.).

[18] This is demonstrated in the series of SEMNET postings that culminated in the Simplex+8.2 model as presented in Hayduk 2/26/98 7:22 p.m., which we summarize below. See also Hayduk 5/16/97 1:37 p.m.

[19] There was considerable SEMNET discussion of the connection between the minimum number of factors and parsimony. Stan contended that the fewer the factors, the more parsimonious the model. Les countered by observing that this is a model-biased view. Models with more but sparsely interconnected concepts can be as parsimonious as, or even more parsimonious than, models with fewer factors and many loadings. Les does not recognize the factor model as "the standard model" and hence is free to acknowledge that models with more but sparsely interconnected concepts can be wonderfully parsimonious. See the exchanges between Les and Stan of 3/24/98 and 3/25/98 on the bloody mess that can be created by wielding Occam's power-razor (chainsaw).

[20] The math appeared in Mulaik, 5/16/97 2:35 a.m., but see also Mulaik, 12/19/97 4:04 p.m., 12/24/97 2:21 p.m., 1/5/98 9:54 p.m., 1/12/98 9:49 p.m., and 4/23/98 6:34 p.m. The proof corresponds closely to that provided by Jöreskog (1967, pp. 445–449).

Stan began with a covariance matrix modeled with *m* orthogonal factors having loading matrix F, so that

$$C = FF' + U^2$$

where U^2 is a diagonal error variance matrix. (This parallels equation 4–30 of Hayduk, 1987, with uncorrelated factors and uncorrelated errors.) Dividing by the appropriate error standard deviations (U) twice provides

$$U^{-1}CU^{-1} = U^{-1}FF'U^{-1} + I.$$

Using the eigenvector matrix of the left side, (A) provides a diagonal matrix on the left, and hence the $A'U^{-1}FF'U^{-1}A$ matrix on the right of

$$A'U^{-1}CU^{-1}A = A'U^{-1}FF'^{-1}A + I$$

is also diagonal. (Eigenvectors may be scaled so that the sum of the squared vector values equals 1.0 by convention. Thus $A'IA = A'A = I$. Compare this with Jöreskog [1967], equation 15.) Because the matrix containing FF' has at most as many nonzero eigenvalues as there are factors defining the columns of F, examining the eigenvalues of the matrix on the left to see which are greater than 1.0 should locate the number of factors, because the remaining eigenvalues must all equal 1.0 if the above equation is to hold.

After permitting sampling fluctuations, and linking the test of model fit to a test of equal remaining eigenvalues, Stan concluded that if *m* is the selected number of factors "... and if m^* is the 'true' number of common factors, then the probability that $m > m^*$ is equal to or less than the significance level of the tests. So, this suggests that if you use a small *p* value for your significance level, you are less likely to extract too many factors. This says nothing about whether you will take too few. But I think you will with small *p*. Nevertheless, with large sample sizes, you will be able to detect small departures from equality for the residual eigenvalues. And this will lead you to taking many factors, although many of them may be of minor importance because they contribute so little to the common variance" (Mulaik 5/16/97).

Stan's concluding comments acknowledge that, despite the proof, even he feels the researcher may end up taking "two few" factors. This is a direct statement that acceptable factor analytic fit can be achieved with the wrong number of factors. Specifically, the researcher is more likely to err on the side of having too few, rather than too many, factors. (*Continued*)

This makes it clear that the hypothesis for the chi-square test is the number of factors k_0 versus its alternatives, and that this is not a test of the veracity of the basic factor model in which k_0 is found. That is, the Step-1 factor model test *assumes* the basic form of the model is correct; it does not *test* whether this factor form is correct. The researcher who believes in the base model that is transformed into the Step-1 factor model hopes that the Step-1 model contains the appropriate number of factors and hopes the factor model is appropriate, but they need to *test* both these hopes. It is precisely because the veracity of the basic factor model is not tested by the chi-square test that Stan's proof is unable to defend the four-step.[23]

This point is sufficiently fundamental that we will illustrate that factor models can indeed *fit* with the *wrong* number of factors/concepts by considering two examples that were discussed on SEMNET: one real example and one hypothetical example.

The real example arises from a data set in which the failure of a factor model shocked one of us (Les Hayduk). Les had imbedded 10 measures of people's personal space preferences (minimum comfortable interaction distance) within an experiment. Numerous other spatial measurements had monitored the treatment effects, and these 10 measurements were placed throughout the experiment to serve as pure replicate measures under null-treatments.[24] The data were collected in the earnest belief that a single underlying factor, namely each person's basic spatial preference, was being repeatedly measured. This preference was conceptualized as the baseline spatial preference against which Les could judge whether or not the experimental treatments were having any effects. The experimental setting resulted in a small sample size but permitted control of many potential sources of error.[25]

The correlations among the 10 indicators were strong (averaging .84) and reassuring. Unfortunately, the one-factor model failed to fit the data. The failure of the one-factor model shocked Les, because to admit even two factors seemed inconceivable. It seemed there had to be one factor. If there was no "thing" being tapped by the replicate measurements, the idea that the experimental treatments effected "that thing" was in jeopardy.

The eigenvalues for the correlation matrix are 8.76, .60, .21, .13, .10, .07, .05, .04, .02, .01, and attempting to fit ordinary factor models[26] results in failure with a single factor, $\chi^2 = 69$, $df = 35$, $p = .0005$; failure with two factors, $\chi^2 = 40$, $df = 26$, $p = .04$; fit[27] with three factors, $\chi^2 = 23$, $df = 18$, $p = .16$; and nonconvergence or underidentification with four or more factors. So from the factor analysis perspective, it seems that two factors were not quite enough, three were sufficient, and more were impossible.

Unfortunately, the proper number of factors in this instance is probably 10, not 3. We know this because there is a pattern in the correlations. The indicator correlations are higher near the diagonal and decline progressively for correlations further from the diagonal.[28]

The significance of the pattern initially escaped Les, and the failure of the factor model led him to reexamine the experimental design. This reconsideration led to an alternative model with 10 concepts imbedded in a straight-line causal chain. In this model the participants' spatial preferences at any time are modeled as causally responding only to their immediately prior spatial preferences. This simplex model is identified,[29] *demands* declining correlations for more widely separated observations,[30] and in retrospect makes excellent sense as a dynamic spatial preference (Hayduk, 1994). The spatial preferences for the participants were changing from moment to moment, even though those preferences remained highly correlated because at any one moment they were strongly dependent on the preferences a moment earlier.[31]

[20]*(continued)*. We can appreciate that small sample sizes might permit real but small differences in the tailing eigenvalues to go undetected, but there are more fundamental ways of questioning this proof, as indicated in the body of the text.

[21]Hayduk 5/16/97 1:37 p.m.

[22]Mulaik 4/23/98 6:34 p.m.

[23]The SEMNET discussions between 4/20/98 and 4/28/98 highlighted the difference between testing the number of factors that are sufficient to make a factor model fit, and testing whether the factor model with any number of factors is the proper model. The factor model test examines only the first of these, and leaves untested whether the factor model is the proper model.

[24]More of the context is provided in Hayduk (1996, pp. 97–100) and the relevant methodological details and data appear in Hayduk (1985). The correlation matrix was also provided in Hayduk 1/27/98 2:47 p.m.

[25]Trying to rebut the point being made here by emphasizing the smallness of N is a tangent because the proof fails even with a fictitious counter example. We could merely postulate that the same correlations appeared with an arbitrarily larger N, and our point would remain the same.

[26]In 1985 Les considered only a single factor and used some nontraditional specifications. Dr. N. M. Lalu kindly ran these factor analyses using SPSS specifically for this paper.

[27]Adopting Stan's $p < .05$ fit criterion for the moment.

[28]Stan's proof is silent with respect to patterns in the data matrix. There is nothing in this proof that lets us do anything different about determining the proper number of factors as a result of this specific data pattern, or as a result of any of the hundreds of less obvious patterns we point to shortly. Stan's proof places no constraints on the data matrix, other than that it be a correlation–covariance matrix for items gathered "in good faith" as indicators of some factor(s)/concept(s). Les had the faith; walked the multiple-indicator methodological walk; and was shocked, devastated, and even sent into temporary denial when the factor model failed.

Pointing out that there was an overlooked data feature does not excuse the factor model's failure to locate the proper number of factors. What is of concern is that Stan's proof does not ask for or permit additional evidence. It does not suggest that any kind of evidence whatever could stand against the proof. In this instance, the researcher's diligent seeking of, and belief in, a single factor, the high correlations, and the small trailing eigenvalues that would probably take a sample in the thousands to detect as unequal, are all in accord with the requirements of the proof, and everything seems to be OK—everything, that is, except the conclusion.

[29]With fixed nonzero thetas, as explained in Hayduk (1985).

[30]Les did not initially appreciate the significance of noting that this pattern is demanded by a simplex model, whereas this and many other patterns can be well approximated by a three-factor model.

[31]See Hayduk (1985, 1994, 1996, pp. 97–101) for discussion of this.

What does this tell us about a Step-1 factor model? 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 concepts that are probably there. The really bad news is that the identified model with three factors *fits* acceptably. How can a model with 3 factors fit acceptably when the real world has 10 factors/concepts? Something has gone drastically wrong, despite Stan's proof. The problem seems to be that the real world has many concepts connected by few effects, while the factor model posits relatively few concepts having many effects.

Before we consider this further, though, it is important to note that this failure is not unique to this specific data set, nor is it attached to the straight-line simplex model with its characteristic correlation pattern. *There are many real-world models that can similarly result in a factor model fitting with fewer than the proper number of concepts.* In the SEMNET discussions, Les constructed a series of hypothetical models (the SimplexPLUS models) that added effects into a basic straight-line simplex model, and thereby progressively transformed the simplex model into other models that came to resemble ordinary recursive models. The Simplex+8.2 model, for example, adds eight effects onto a simplex backbone, and its correlation matrix does not display an obvious simplex pattern.[32] There are 10 concepts in the Simplex+8.2 model, yet a factor model fits with only 3 common factors.[33]

The personal space example illustrates that there are real instances of models that fit with too few factors, and the series of hypothetical SimplexPLUS examples demonstrate that there are potentially hundreds of covariance matrices from ordinary recursive models that can be mistakenly fit by too few factors.

The correlation matrices for the SimplexPLUS series of models progressively lose the telltale correlation pattern, but these matrices continue to have large initial eigenvalues and small trailing eigenvalues that drive the chi-square test for Step 1.[34] What is not obvious is that the eigenvalues are acting in consort with a hidden accomplice, namely the unnecessarily free coefficients in the "unquestioned" factor model. Let us look at this from the perspective of saturated models.

A saturated model (factor or otherwise) guarantees that the model will fit by being able to exactly reproduce the observed covariance matrix. Saturated models fit, not because the number of factors/concepts or anything else about the model is right, but because of the number of free, and possibly needlessly free, loadings/coefficients attached to the factors are sufficient for reproducing the data. The relevant issue here is, does deleting one loading from a saturated factor model lead to a large and significant chi-square ill fit if the true real-world model has more than the saturating number of factors? The answer is clearly no, it will not. The chi-square fit will only slowly decline with each lost loading, because the remaining loadings, even if misspecified or useless in the sense of not corresponding to effects in the real world, still manage to imply covariances that match the data fairly well.

The extra coefficients added to turn a sparse structural model into a full set of concept covariances, and the extra coefficients added to convert focused loadings into loadings spanning a wider range of items, are like fitness–Band-Aids that can obscure the hemorrhaging (ill fit) created by having too few factors/concepts in the base model.[35] Too few factors, assisted by all the unnecessary coefficients added to create the Step-1 factor model, may lead to passing the Step-1 fit test, even though the real-world "model" has more factors/concepts with fewer loadings and fewer structural effects.

What the four-step is lacking, and may forever lack, is a demonstration that the coefficients freed in getting to the Step-1 factor model are not hindering the assessment of the number of factors/concepts. We see the hindering nature of these coefficients (the extra loadings and free concept covariances) when they take a perfectly good 10-concept model and render it underidentified as a 10-factor

[32] The discussion of the SimplexPLUS models began with Hayduk 1/26/98, was formalized in Hayduk 2/14/98, and took its most helpful form in the context of the Simplex+8.2 model of Hayduk 2/26/98 7:22 p.m. The Simplex+8.2 model is created to have a backbone of 10 concepts connected in a straight causal line (with all effects .8). The first concept has additional .3 effects leading to concepts 3 to 10. The first concept has variance 1.0 and each of the other concepts has a structural error of .15. Each of the 10 single indicators has a .25 measurement error variance. This specification results in a recursive model whose correlation matrix is

$$\begin{array}{llllllllll}
1.0 \\
.70 & 1.0 \\
.72 & .73 & 1.0 \\
.73 & .72 & .79 & 1.0 \\
.74 & .70 & .76 & .82 & 1.0 \\
.75 & .70 & .75 & .80 & .83 & 1.0 \\
.76 & .70 & .74 & .77 & .81 & .85 & 1.0 \\
.77 & .69 & .73 & .76 & .79 & .83 & .86 & 1.0 \\
.77 & .69 & .72 & .75 & .78 & .80 & .83 & .86 & 1.0
\end{array}$$

[33] Mulaik 3/21/98 9:09 p.m.

[34] For example, see Mulaik 2/18/98, 3/9/98. Stan reported that the eigenvalues of $U^{-1}CU^{-1}$ for the Simplex+8.2 correlation matrix are all greater than 1.0 when the true error variances are used, just as the proof requires. The trailing eigenvalues are close to 1.0 and could easily lead to five of the eigenvalues being discounted. But again, the problem with Stan's proof does not seem to be in the mathematics of the eigenvalues.

[35] "Without the structural model to constrain the factor covariances, the freely floating factor covariances are a sponge that can absorb some slack in 'the wrong number of factors' at Step-1" (Hayduk 5/9/97 1:48 p.m.). "I am arguing that, if the real world places constraints on the factor covariances, and the Step-1 model does not have those constraints in place, then we can have made a wrong decision about the number of factors because of the lack of those constraints" (Hayduk 5/12/97 12:04 p.m.). See also Hayduk 4/3/97 6:11 p.m.

Stan's general position was that it would be impossible for a five-factor model to produce an acceptable fit if the real world contained six concepts and a sparse set of coefficients (see two Mulaik postings of 5/13/97). But also see "So, at any step there is the possibility that we have taken too few factors" (Mulaik as quoted in Hayduk 5/12/97 12:04 p.m.).

model.[36] We should see the similarly hindering nature of the unnecessarily free coefficients when the factor model is identified but padded well enough to keep us from feeling the pinch of having only 3 factors instead of 10 concepts.

A proof relevant to the four-step procedure would require that one can start with any real-world model whatever (the Step-3 base model being the researcher's best guess at this), and still have the Step-1 factor model created by adding multiple anticipatedly useless coefficients (the extra loadings and free concept covariances) nonetheless provide a definitive statement of whether the base model does or does not contain the proper number of concepts.

Unfortunately the researcher has no way of knowing when a bunch of unnecessary loadings are unjustifiably providing large initial eigenvalues that can lead them astray by providing an acceptable chi-square. Any given set of indicators might be reflecting multiple interconnected concepts, despite the researcher's best efforts and intentions, and despite even the power of experimental controls. Yes, we can do our best at locating multiple indicators (Les did), and yes we may even account for most of the item covariances with a few factors (three factors explain 96% of the common variance in the personal space measures),[37] but we can still have a wildly wrong number of factors/concepts. Three factors are not even close to the 10 that provided the Simplex+8.2 data, or the 10 that most sensibly match with the personal space data. Unfortunately, the researcher usually has no way of knowing if this is happening or not, and hence even with acceptable fit at Step 1 the researcher might be proceeding to Step 2 with the wrong number of concepts.

[36] And similarly, we see it when the freed coefficients render models with 9, 8, 7, and 6 concepts underidentified as factor models with 10 indicators.

[37] The argument that larger eigenvalues correspond to more "important" factors became central after the Mulaik posting of 4/11/97 8:39 p.m. This is connected to the discussion of whether one should test exact model fit or approximate model fit (see the discussion of the RMSEA below). The connection between eigenvalue and importance was examined in the SEMNET exchanges following Hayduk 2/18/98 10:52 a.m. It became clear that the eigenvalues for simplex models do not correspond in "importance" to the concepts, but at "best" they correspond to the errors on the concepts. This has the unsettling consequence of claiming that eigenvalues find "well explained endogenous concepts" to be unimportant and unworthy of being retained in the models.

Another approach to seeing that eigenvalues need not correspond to, "importance" can be attained from considering a circumplex model, which is like a straight-line simplex except that the final concept influences the first, so the model becomes a circle of effects. A circumplex of 10 concepts with equal effects and equal errors is composed of 10 identically important concepts, no matter how one defines importance. Unfortunately, the eigenvalues for a covariance matrix created from such a model will show the characteristic pattern of large initial eigenvalues tapering off to near zero values. This demonstrates that both large and small eigenvalues may not indicate importance, because all the concepts in the model are of equal importance.

These observations contradict the factor-analytic idea that the smallness of trailing eigenvalues makes subsequent factors "of minor importance because they contribute so little to the common variance," as Stan put it, but this is a tangent in the context of the four-step, and so we do not pursue this.

This is a failing from which the four-step procedure is unlikely to recover. Step 2 is unlikely to force any acknowledgment that the researcher has the wrong number of concepts. If the Step-1 model fits and the Step-2 model fails, the researcher can merely reinsert some of the loadings that had been deleted in moving from the Step-1 model to the Step-2 model, and the Step-2 model will fit without readjusting the number of concepts.[38] Similarly, ill fit at Step 3 can be counteracted by adding in a few more structural effects, and the researcher again will not be prompted to seriously requestion whether they have the proper number of factors.

Yes, we *could* question whether we have the proper number of concepts at failing Step 2 or Step 3 even after observing a well-fit Step 1. In fact we *should* do this. We *should* be willing to requestion the number of factors *despite a fitting Step-1 model*. But the 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. If we are not assured of having the proper number of concepts, then the measurement issues have not been settled even by Step 3.

The consequence of this is that we are lacking precisely the benefit the four-step was supposed to provide, namely a clear statement of whether a failing Step-3 model has measurement problems or structural problems. The measurement concerns, in the guise of a dubious number of concepts, linger all the way back to the Step-3 model. The absorbent Band-Aids provided by the freeing of the coefficients during the formation of the Step-1 and Step-2 models actively obstruct any determination of whether the base model contains the proper number of concepts.[39]

Another way to approach the question of whether Step 1 determines the proper number of factors is to consider whether a fitting model can be the wrong model. A SEM truism states that a good fit is insufficient to conclude that the model is the correct model. If we apply this truism to the Step-1 model, this claims that the Step-1 model can fit with *k* factors but that this is still the wrong model. So either we can have fit with the wrong number of factors/concepts, or we have an instance

[38] Stan thought 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" (Mulaik 4/4/97 1:12 a.m.).

[39] One might consider trying to solve the problem of the Step-1 model fitting because of capitalization on unnecessary yet free coefficients by adjusting the test of the Step-1 model's fit for the number of excess coefficients in the model. Unfortunately, for the four-step advocates, this constitutes a direct dismantling of the four-step because this requires incorporation of knowledge of the base model to create the adjustment. This reintroduces aspects of the base model into decisions about measurement, which the four-step advocates are trying to avoid.

From the perspective of those not committed to the four-step, it would seem reasonable not only to use the number of excess estimated coefficients in adjusting the Step-1 model test, but also to include the actual model locations of the unnecessary coefficients. This takes us right back to the base model, so we might as well just use the entire base model, and routinely consider a possibly incorrect number of factors as one of the reasons the base model might fail.

where, contrary to the truism, acceptable fit does mean one has located the correct model. We see no reason to grant that moving toward the factor model provides a way of gaining indubitably trustworthy knowledge.

The consequence of the preceding is that the Step-1 test does not reliably determine whether the researcher's model contains the proper number of concepts. It merely tests whether the researcher's best guess at a model, the base model, propped up by a bunch of coefficient-protheses here, and held together by a bunch of coefficient-stitches there, might be fit enough to be carried out of the emergency room (Step 1) and onto the next station. If the emergency room can't reattach the missing appendages (concepts) then it is unlikely they will regenerate while the patient/model lies in the next ward (Steps 2 or 3).

The inability of the factor model to convincingly locate the true number of factors is not a problem that is unique to the four-step procedure. It is a problem that arises whenever a factor model is tested in lieu of a real-world model having more concepts. The above arguments explicitly critique factor-analytic determinations of the true number of factors by splitting the academic hair separating a "test of the number of factors in an assumed factor model" from a "test of the assumed factor model."

Now, onward and downward to another set of four-step problems.

Structural Effects, Fixed Measurement Errors, and the Proper Factors

Imagine that a researcher believes there is a gold-standard indicator of some concept, so that the concept must have a specific close connection to that indicator if the concept is to be the concept/factor the researcher wants in his or her model. The researcher's base model includes this indicator, along with other indicators of the concept, and the researcher uses the fixed 1.0 loading (lambda) and fixed corresponding measurement error variance (theta) procedure for the golden indicator as discussed by Hayduk (1996, pp. 25–30).

The first use of the fixed measurement error variance (theta) cannot be at Step 3 of the four-step because the fixed theta seems to be a measurement concern, and measurement concerns are supposed to have been determined at Steps 1 and 2. If the fixed theta is entered at Step 1, this fixed value may contribute to model failure because the gold-standard indicator does not in fact share only a single concept/factor with the other weaker indicators. There may be some single factor that might fit with all the indicators, but that factor is not the factor/concept the researcher wants, because it is not sufficiently closely connected to the gold-standard indicator.

The problem with inserting a fixed theta at Step 1, from the perspective of a four-step advocate, is that Step 1 would now be testing something other than just the number of factors. Fit would hinge on a specific factor having an identity closer to that intended or required by the researcher. But if we postpone the entry of the fixed theta (measurement error variance) until Step 2, and Step 2 fails because of this fixed value, we have another problem. Step 1 is supposed to have told us we have the proper number of factors (we passed Step 1), yet Step 2 says that one of the factors cannot be the factor the researcher wants in his or her model. That is, the factors in the Step-1 model are not necessarily the same factors that are in the Step-2 or Step-3 base model with the specific gold-standard meaning imposed by the fixed theta. To go back to Step 1 and add another factor, we would have to challenge the claim that Step 1 determines anything definitive about the required number of factors/concepts. Alternatively, the researcher might back off the gold-standard indicator, irrespective of how many strong and convincing studies have been done to establish that gold standard. Both options are uncomfortable for the four-step advocate. It is not easy to back away from the idea that fit at Step 1 determines the proper number of factors/concepts, nor is it easy to claim that application of the four-step should be restricted to areas of study for which there are no solid or trustworthy indicators. This would mean the four-step is to be reserved for use in disciplines that are lacking in established measures, so that one never encounters a really well-established indicator of any concept/factor.

This conundrum gets worse if we apply a parallel argument to the structural part of the model. Suppose the researcher's theory demands that a specific concept be linked to some other concepts with some very specific effects. If this concept is not coordinated in this way, it simply is not the proper concept. That is, the substantive area has sufficient theoretical integrity to demand specific theoretical connections between some of its conceptual entities. Now for the parallel problem. Suppose we leave entry of these theoretically motivated constraints until Step 3, and then we encounter model failure because of these constraints. Steps 1 and 2 may have been passed, but forcing the concept to be "the intended theoretical concept" makes the model fail, just as forcing a more specific meaning on the concept with a fixed measurement error variance might make a later model fail.

It seems that one reasonable option might be to consider whether the model might require insertion of another concept/factor to accommodate the theoretically demanded conceptual meaning—despite one's having successfully passed Steps 1 and 2, which are supposed to have determined the number and identity of the factors/concepts.[40] This says that the measurement issues have not been convincingly settled even by successful model fits at Steps 1 and 2, and that measurement remains an open issue all the way to Step 3.

The phenomenon grounding the fixed measurement error variance and demanded-theoretical-structure problems seems to be that while "some" factors might pass Steps 1 and 2, there is insufficient control of those factors to be sure that they are "the same factors/concepts" demanded by the theoretically and methodologically dictated constraints. Again, the promised separability of measurement

[40] See Hayduk 3/21/97 1:37 p.m. and the related exchanges.

from structural concerns is lost, and the separation of measurement from structure seems artificial and incomplete.

The "No Peeking" Problem

To maximize the independence of the test at the various steps, the researcher using the four-step should not use knowledge of the estimates obtained at any prior step in making decisions about the model being tested at a later step. This is because using the data to modify the model compromises the ability of the data to test that model. 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.[41]

This is problematic if substantial unexpected loadings, or absences of expected loadings, at Step 1 are signs of measurement problems. Small loadings might be indicating a factor/concept that is not as close to specific important indicators as it ought to be, and the factor intercorrelations might indicate some concept has moved closer to, or further from, specific other concepts than is reasonable, but 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).[42] These "damned if you do, or don't" conundrums are artifacts of the sequential testings inherent in the four-step process.

What Test and What Critical Value Should Be Used?

Two of the problematic details of the four-step procedure concern the test to use in deciding whether to proceed from one step to the next, and the criterion value to use for that test.[43] We will begin by considering the .05 level typically used with the chi-square test of model fit, and later examine the RMSEA as an alternative test.[44]

[41] Mulaik 3/27/97.

[42] See the exchanges following Hayduk 3/26/97.

[43] Stan insists that these are tests, and that the researcher stop and not go on if a test is failed at any step (Mulaik 3/20/97 15%, 3/23/97 3:50 p.m., 3/24/97 25%, 4/16/97 3:35 a.m. 50%, 4/22/97 9:20 p.m.). There was no discussion of what the four-stepper would do "if there were no significance tests."

When Les and Stan addressed the issue of whether it was sufficient to pass the overall model fit criterion, or whether the model had to fit in detail (e.g., no modification indices over five or so), there seemed to be consensus that detailed, and not just overall, fit was required at each step. Unfortunately, there are no stated criteria for detailed fit, so one never really knows if one should proceed or not (see Mulaik 3/24/97 3:05 p.m. 30%, and Hayduk 3/25/97 4:04 p.m.). We ignore this contradiction by confining our attention to the issue of overall fit at each step.

[44] Nested models usually justify the use of a difference-chi-square test, but Stan did not propose this. He uses three separate model chi-square tests, rather than one model chi-square test and two difference-chi-square tests via nesting. If a single model-test was used at Step 3, Stan would have a difference test only at Step 1, and this would make the Step-1 test different from the usual factor-model test—which a factor analysis proponent might find uncomfortable. Alternatively, if the single model-test was used at Step 1, the Step-3 test would now be a difference-chi-square test, and we would have no overall test of the base model, which would be uncomfortable for anyone who believes in testing the researcher's real model. Stan has not yet, to our knowledge, articulated why he prefers to have the Step-1 test be a usual factor model chi-square test, rather than a difference-chi-square test, which nesting would ordinarily warrant. We leave it to others to pursue the implications of this.

The Criterion Level, and Favoring Null

The four-step procedure is stuck with having to specify some boundary between when a model does, and does not, fit well enough to permit proceeding to the next step. The traditional rejection (stop and do not go on) criteria has been a probability of less than .05 for the chi-square test. We argue that the probability criterion 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.

To see why, consider the role played by random sampling fluctuations in ordinary hypothesis testing. Imagine a researcher who postulates a relationship between two variables, gathers some data, and finds that the relationship in the observed data is statistically significant ($p < .05$) because the observed correlation is too far from the null hypothesis of zero relationship for it to be easily dismissed as a sampling fluctuation. It would be rare for (there would be a probability less than .05 of) a relationship of the observed strength to arise due to mere sampling fluctuations if there were indeed no relationship in the population, so the researcher is permitted to point to the data as nonchance evidence in favor of whatever theory prompted postulation of the relationship.

The point is that, before the researcher is permitted to use the data as evidence supporting his or her favored theoretical position, he or she is required to demonstrate that the supposed evidence cannot be explained away as a mere sampling fluctuation around some other value that is reasonable and yet inconsistent with the theory—in this instance the null hypothesis of no relationship between the variables. The researcher proceeds in a way that secures the assent of a disbeliever by demonstrating that the observations are unlikely to be mere chance occurrences, namely sampling fluctuations around a value that is inconsistent with the theory.

Finding a $p < .05$ (passing the usual significance test) is a reply to anyone who argues that "it's not the forces described by the theory that resulted in the data appearing as they did (the nonzero correlation); it was a mere chance sampling fluctuation around some other value (no relationship), so these data should not stand as support for the theory." The reply being provided by the researcher is that this argument is not tenable because it appeals to sampling fluctuations that, though possible, are rare, because they occur in less than 1 out of 20 random samples. Science has been well served by demanding that theorists quell their claims until they have evidence that a reasonable counterexplanation can be annulled via a demonstration that the counterexplanation has to be stretched unreasonably by appealing to a "very small and unlikely chance occurrence" (a probability of .05 or less) to account for the evidence/data.

Suppose now that the theorist proposes that there is no relationship, *so that the researcher favors what would ordinarily have been treated as the null hypothesis counterexplanation*. If we kept the usual .05 probability level, we would be saying that the theorist should be permitted to loudly trumpet their theory until there is a small chance that the theory is *consistent* with the observed data. Observing data sufficiently far from zero that it could appear only half the time by chance sampling fluctuations (a test with $p = 0.5$) would not quell the researcher's claim of no relationship (0.5 is well above the .05 critical value). Hence the researcher postulating a null relation would (we think unjustifiably) be permitted to claim the data is consistent with his or her theory, even though had that same researcher postulated a nonzero relationship, a similar .5 probability would have been judged as grossly insufficient to permit a similar proclamation that the data is consistent with his or her theory.

An analogous difficulty plagues the chi-square test of structural equation models, and the tests of the four-steps in particular. The null hypothesis of the chi-square test is that the covariance residuals created as the difference between the observed data and model implied covariances are zero in the population. If the model is correct it should permit perfect reproduction of the observed covariance matrix, and consequently researchers hoping their models are correct are favoring the null hypothesis of zero residuals (no remaining differences).

Instead of beginning a technical discussion of the power of tests and the relative seriousness of Type I and Type II errors, let us satisfy ourselves by eavesdropping on a debate between two researchers. Researcher A is doing the four-step, and you are Reviewer B, who is considering A's arguments. The debate ends with several chi-square probabilities. You as reviewer, both here and in real life, will ultimately decide whether the probability Researcher A selects is sufficient to convince you that he or she has paid appropriate attention to sampling fluctuations for him or her to confidently proceed to the next step of the four-step. By clarifying how a switch to favoring null reverses the onus regarding sampling fluctuations, we hope to convince you of the need to switch to a larger criterion value.[45] So, are you (Reviewer B) convinced that the model at this step is sufficiently trouble-free to permit Researcher A to proceed to the next step because A's model probably is free of whatever problem would normally be detected at this step?

Researcher A: My Step-1 (or -2) model does pretty well at explaining the covariances among the observed indicators. The differences between the observed covariances and the model implied covariances have been minimized by the selection of optimal coefficient estimates, and the remaining residual covariances seem small and scattered. Granted, my model does not account for ALL the covariances exactly, but few models fit perfectly unless they are saturated with coefficients, so I conclude that my model fits the data sufficiently well for me to proceed to the next step.

Reviewer B: I agree that the residuals report on discrepancies between the data and what your model claims, but I see the discrepancies as being large enough, and important enough, that they constitute a direct statement that your model is incapable of matching up with the data it was supposed to explain. The residuals are a direct assertion that there are discrepancies between the only data you have and what your model, aided by the best estimates, implies.

Researcher A: But the residuals (discrepancies) are small and unimportant! In fact, the discrepancies are so small that they can be dismissed as mere sampling fluctuations and hence they are not signs of model failure.

Reviewer B: I'm not convinced. Large residuals are strong and direct evidence of model failure, and I think the residuals you are reporting are big enough that they can't be dismissed as mere sampling fluctuations, and consequently I must conclude that your Step-1 (or -2) model should be counted as failing because it is inconsistent with the data.

Researcher A: But look at the chi-square probability ... it's insignificant. This tells us that the differences between my model and the data are not significant! The remaining differences are likely to be mere sampling fluctuations, and nothing more! The residuals are not big enough to be demonstrably anything other than chance sampling fluctuations.

Reviewer B: It is not I who have to convince you that the residuals are not chance. It is you who must convince your readers, me included, that it is reasonable for us to dismiss residuals this large as being likely due to chance. Your probability is greater than .05, and hence is statistically insignificant, but

[45] It took several promptings (Hayduk 3/25/97; 4/7/97 8:47 a.m., 7:54 p.m.) to get Stan to commit himself on this, but he finally claimed that favoring null is not a reason to alter the probability used in the test (Mulaik quoted in Hayduk 4/8/97 9:20 a.m.). This led to a side discussion in which several people participated (see the SEMNET discussions between 4/22/97 and 4/27/97), but no general consensus was reached. There was a revival of the favoring-null debate in October 1997 (this time between Les Hayduk and Bill Shipley) that ended with an agreement that some change was indeed required. No one else on the Net seemed willing to step forward to defend the continued use of .05.

you are making a mistake by thinking that this makes the differences likely to be mere sampling fluctuations. I would be inclined to believe you if the kinds of residuals you are reporting were the kinds of residuals that occur frequently, usually, routinely, or consistently as a result of chance sampling fluctuation. But you are not reporting such residuals. Your residuals are not the kind of residuals that are typical, usual, or likely to arise from sampling fluctuations.

Researcher A: Oh yes they are. They are insignificant! I am not appealing to unusual kinds of sampling fluctuations. In fact my model's residuals are "ordinary" and are the kind of residuals that would be observed quite routinely because the probability for chi-square says that residuals this large or larger would be expected to happen:

Almost all the time ($p = .95$).
Most of the time ($p = .75$).
As often as not ($p = .5$).
Sometimes ($p = .25$).
Relatively rarely ($p = .06$, or in one out of about every 17 samples).

So, what number would you find sufficiently convincing to quiet your critical tongue? We suspect you would be persuaded by A's argument with a .95 probability and be progressively less convinced as the probability declines. But this is insufficient for us to recommend .95 as a criterion for model fit. A .95 criterion would result in the relatively frequent appearance of another mistake, namely the rejection of true models displaying even moderately likely random sampling fluctuations. So we have reason to avoid both ends of the probability scale. The .06 end makes Researcher A lose the above argument, whereas the .95 end demands atypically small sampling fluctuations, and increases the likelihood of rejecting true models.

Although .5 might be a reasonable numerical compromise, this insufficiently acknowledges that science has found it prudent to place more burden on the advocate (Researcher A) than on the skeptic (Reviewer B). Hayduk recommended that one should "aim for probabilities in the .75 region" (1996, p. 77) and for "$p > .75$ or so" (1996, p. 69), thereby acknowledging the advocate's burden while simultaneously setting a fuzzier boundary to augment consideration of the distinction between "failing to reject" and "knowing it isn't." This softer line on hypothesis testing is not easy for four-steppers to accept. They need a clear decision rule, so they "know" whether to proceed or not. We will leave the four-step proponents to argue among themselves over the precise new probability criterion, knowing that the researcher's implicit favoring of the null hypothesis will make four-steppers

consistently lose the above argument until they adopt a probability considerably larger than .05.[46] Note that this problem will remain, no matter what null-favoring test is used.

Which Test: Chi-Square or RMSEA?

The early use of the chi-square test (James, Mulaik, & Brett, 1982) and chi-square's nice statistical properties led to this being the most commonly used test with the four-step. Unfortunately, Stan found that even at the lax .05 level, the chi-square test was rejecting more models than he liked, so he went looking for a less stringent test, namely a *test of close fit* rather than a *test of exact fit.*[47] There seem to be two styles of problems attached to this endeavor.

First, the logic of testing close fit does not match with the requirements of the four-step. For example, the Step-1 test is supposed to inform us whether we have the proper number of factors. What is the poor four-stepper to do with even the best possible news from a close-fit test at Step 1? The best news would be that we have *close to* the proper number of factors, while presumably chi-square tells us we do *not* have the proper number of factors.[48] It would seem unreasonable to proceed to the next step, no matter how "close" this "close" is to the proper number. The move to "close enough" constitutes a direct assault on the ability of Step 1 to confidently

[46] In the SEMNET discussion, Stan stuck to the .05 level when confronted with the problems created by favoring null (Mulaik quoted in Hayduk 4/8/97 9:20 a.m.). It remains to be seen whether other four-steppers will similarly choose the hard place in the argument over fear of a slippery probability slope.

The suggestions that models with high chi-square probabilities are necessarily artifacts of capitalizing on chance during extended data snooping, or that these are artifacts of trivial structural sparsity, were refuted in the SEMNET discussions by pointing to Hayduk, Stratkotter, and Rovers (1997) as a clear counterexample. The Rigdon (4/18/98 2:40 p.m.)–Hayduk (4/20/98 12:56 p.m.) exchange is also instructive in this regard.

[47] Mulaik 4/11/97 8:39 p.m., 4/14/97 1:43 a.m., 4/17/97 3:40 p.m., and 5/14/97 6:00 p.m. This sounds much more sophisticated when one uses stat-speak to describe how a laxer standard leads to acceptance of models with small chi-square probabilities. "In comparison with the test of exact fit, the same test statistic is used, but the value of χ^2_c will be greater because critical values in the noncentral distribution of $\chi^2_{c,\lambda}$ are shifted to the right of corresponding values in the central distribution of χ^2_d" (MacCallum, Browne, & Sugawara, 1996, p. 135).

We have found that the probability for the chi-square test is informative and worth paying attention to (Hayduk, 1996, p. 201). Stan, and others, on the other hand, are seeking another test because the chi-square test seems inclined to request more factors than they desire. One way to read this general complaint is that the chi-square test is detecting and reporting on biases in factor-analytic practice that lead to specification of too few factors. A dedicated "pursuit" of the minimum number of factors might just do this.

[48] Had χ^2 indicated we had the proper number, then there would be no need for using a test of close fit. Note that here we are only putatively granting that Step 1 actually tests for the proper number of factors.

locate the number of required concepts, and without this assuredness the four-step is robbed of its ability to pinpoint uniquely measurement problems.

Second, the chi-square probability, now presumably less than .05, will not disappear when a new test is discovered, so Researcher A's side of the favoring-null argument will be even weaker. Researcher A will begin by pleading for permission to be excused from the standards to which others are held because he or she has the audacity to admit right up front that his or her model is not likely to be the correct model. This will probably be attributed to a weak literature, or a complex world, but prudence demands that we suspect that this may be due to a misreading of the literature, or the researcher's complex mistakes.

For researchers propounding close fit, the Step-3 model isn't possibly true, it is just close; and the coefficients added to get to Step 2 keep the model just close; and the coefficients added to get to Step 1 keep the model close, but it is not yet correct. And then the researcher will plead that we have learned something definite about measurement and structural failures. Fortunately, χ's definite statement is still there to confront anyone attempting to weave multiple almost-so stories into confident assertions.

If the base (Step-3) model has sufficient integrity to represent a specific perspective or theoretical stance, then it is newsworthy whether it fits or fails. A failing model should be published by highlighting the style of failure, and the evidence resulting in failure, and not excused as close enough to overlook the failings (Hayduk, 1996, p. 4; Hayduk & Avakame, 1990). A discipline preferring to accept close-fit models, as opposed to highlighting the remaining failings of models, is degenerating as a discipline because it is unable to encourage its practitioners to build models with sufficient integrity and clarity that either fit or failure is informative.

Throwing in a handful of free concept covariances (Step-2 model) followed by a bucket of free loadings (Step 1) directly attacks whatever precision the researcher might have imbedded in the real Step-3 model. Combining this with a plea that chi-square should be overlooked in favor of anything close seems sufficient to guarantee that whatever precision was included in the Step 3 is lost under the onslaught of the progressive loosenings.

The aforementioned difficulties arise no matter which particular test of close fit the four-stepper adopts. Additional problems may accompany the specific test of close fit one uses, such as the RMSEA suggested by Stan.[49]

The RMSEA

The RMSEA was introduced by Steiger and Lind (1980) and the underlying mathematics was developed in Steiger, Shapiro, and Browne (1985). Accessible discussions of the RMSEA have been provided by Browne and Cudeck (1993) and MacCallum, Browne, and Sugawara (1996). The core of the RMSEA is a tradeoff between ill fit and parsimony created by calculating the amount of model ill fit per degree of freedom ($\hat{F}/d$). According to Browne and Cudeck (p. 145) and MacCallum, Browne, and Sugawara (p. 134, equations 5 and 7), a point estimate of the RMSEA is provided by

$$\text{RMSEA} = \sqrt{\text{Max}\left\{\left(\frac{\hat{F}}{d}-\frac{1}{n}\right),0\right\}}$$

where $\hat{F}$ is the minimum of the fit function applying a model with d degrees of freedom to a covariance matrix based on a sample of $n + 1$ cases.[50] "Max" in this formula makes the RMSEA the square root of the term in parentheses if this term is greater than zero, or the square root of zero, which is zero, if the parenthetic term happens to be negative.

To see why this is a measure of close fit, rather than exact fit, we need to consider when the parenthetic term becomes less than zero. This is most easily seen if we rearrange the formula by multiplying the parenthetic term by 1.0 in the form of n/n. This results in

$$\text{RMSEA} = \sqrt{\text{Max}\left\{\left(\frac{\frac{n\hat{F}}{d}-1}{n}\right),0\right\}}$$

[49] Mulaik 4/9/97 5:02 p.m., 4/11/97 8:39 p.m., 4/14/97 1:43 a.m. 80%, and 4/7/97.

In discussing the RMSEA, Browne and Cudeck (1993) granted that "in the social sciences it is implausible that any model that we use is anything more than an approximation to reality. Because a null hypothesis that a model fits exactly in some population is known a priori to be false, it seems pointless even to try to test whether it is true" (1993, p. 137). The "null hypothesis of exact fit ... is invariably false in practical situations" (1993, p. 146). This may fit well with factor-analytic tradition and its progressive inclusion of more factors, but it is substantially at odds with Step 3 of the four-step and the perspective taken in Hayduk (1996), and it is contradicted any time a social science model fits well (e.g., Hayduk, Stratkotter, & Rovers, 1996).

Browne and Cudeck (1993, p. 157) warned that the RMSEA "should not be used in a mechanical decision process for selecting a model." Unfortunately, this is precisely what the four-step needs, namely a decision rule to be applied routinely, across contexts, and with considerable standardization. So Stan is pushing the RMSEA in ways the RMSEA founders discourage. This does not make Stan wrong, just lonesome.

[50] We confine our discussion to the maximum likelihood fit function though Browne and Cudeck (1993) considered the GLS and ADF functions as alternative specifications of the RMSEA.

in which we notice that $n\hat{F}$ equals the usual likelihood ratio chi-square for the fit of the estimated model to the data, so

$$\text{RMSEA} = \sqrt{\text{Max}\left\{\left(\frac{\frac{\chi^2}{d}-1}{n}\right), 0\right\}}$$

In this form it is easy to see when Max must be invoked to keep the RMSEA from becoming the square root of a negative number. χ^2, d, and n, are all positive, so the term under the radical sign becomes negative only if χ^2/d is less than 1.0. The mean of a chi-square distribution equals the degrees of freedom for that distribution, so χ^2/d will be less than 1.0, and hence the RMSEA will become zero through use of Max, whenever the model chi-square falls below the mean of the appropriate chi-square sampling distribution. This should happen about half the time, if the postulated model is the proper model.

So the RMSEA is useless for comparing pretty good models. Proper models would result in zero RMSEA values (via use of Max) about half the time, while nearly proper models would result in zero RMSEA values nearly half the time. Because all models resulting in a chi-square probability greater than .5 will provide a "perfectly close fit" of 0.0 RMSEA, the nonzero values of the RMSEA are left to distinguish among degrees of poorer model fit. So the 0.0 ideal target value for the RMSEA corresponds to a chi-square probability target of .5. Clearly, this close a close fit is insufficient to help four-steppers struggling with chi-square probabilities smaller than .05, so there is undoubtedly something else also going on in addition to the use of χ^2/d and Max to weaken the test criterion.

In fact, two other things are happening, one thing that could be very helpful for the four-step, and a second thing that, unfortunately, nullifies the potentially helpful thing. Both the helpful and unhelpful observations are connected to the suggestion to use a .05 or smaller value of the RMSEA as "close enough" to be considered an acceptably close fit.[51] We saw above that a zero target for the RMSEA has already moved us from considering an exact-fitting model to a close-fitting model, so the suggestion to accept RMSEA values up to .05 is asking us to grant that something even "close to a close fit" is still close enough. This sounds a bit devious, so let's check it out.

What would an RMSEA target of .05 mean? Surprisingly, this question has no answer because the question itself is flawed. The .05 value is not a "target." It is a sample-size-dependent adjustment to what is called "close," and a potentially huge adjustment at that. To see this, consider what requiring an RMSEA of less than .05 means for a sample size of 201, so $n = 200$. Plugging this n and the proposed .05 value into the equation for RMSEA above tells us our model is acceptably close if the chi-square to degrees of freedom ratio satisfies

$$\begin{aligned} .05 &> \sqrt{(\chi^2 / d - 1) / n} \\ (.05^2 (200) + 1) &> \chi^2 / d \\ 1.5 &> \chi^2 / d \end{aligned}$$

This informs us that seeking an RMSEA of "less than .05" with $n = 200$ is the same as seeking a χ^2/d ratio of less than 1.5. Figure 2 was created by repeating this style of calculation to locate the χ^2/d ratios required to satisfy the .05 "criterion" for other sample sizes, as well as for other potential RMSEA "criterion" (namely, .01, .03, .08, and .10).

This figure makes it clear that the .05 value is a sample-size adjustment because it determines how much we are willing to alter the required tradeoff between parsimony and fit (the χ^2/d ratio) in recognition of increasing sample size, beyond the 1.0 ratio already demanded by the Max specification in the definition of the

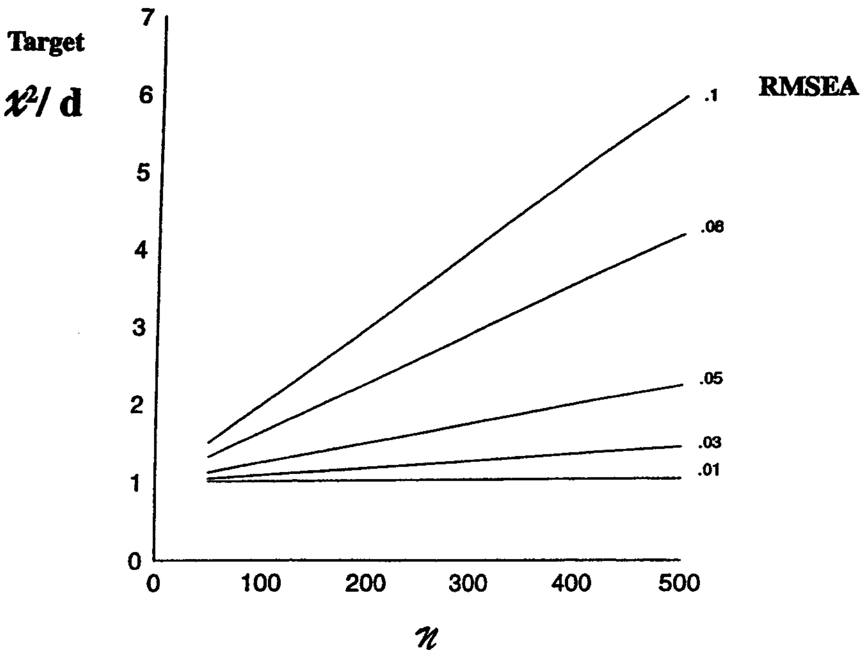

FIGURE 2 The RMSEA close-fit criterion as a sample size (n) adjustment to the χ^2/d target.

[51] Mulaik 4/9/97 5:02 p.m., 4/11/97 8:39 p.m.

RMSEA. Specifying .001 as the RMSEA "target" would require that samples of all sizes adopt a χ^2/d ratio near 1.0 as their close-fit target, while a value of .1 would judge χ^2/d ratios of about 6.0 as acceptable for samples of 500. A value of 6.0 seems positively disgusting once one notices that even values like 3.0 push one right off the edge of χ^2/d tables (e.g., Wonnacott & Wonnacott, 1970, p. 420). For an $n = 500$, an RMSEA "target" of .05 corresponds to a χ^2/d "target" of 2.25, which for 20 *df* corresponds to a chi-square probability of about .001. So what is meant by "close" with an RMSEA "target" of .05 is "far out"—in the tails that is. And then things get worse. We will be asked to accept models whose estimated RMSEA values are even within sampling fluctuations of this "target" that has now been doubly removed from a target of a proper model.

But should we, or anyone, permit any sample size adjustment at all to the assessment of what is an acceptably close fit? Our answer is no. The increasing precision that comes from using larger sample sizes should provide shorter interval estimates of the population RMSEA, so that we know more assuredly whether or not our model is within the required closeness—putatively a χ^2/d ratio of 1.0. If the model is wrong by more than close fit permits, this should be more easily detected with a large sample, so we would expect, indeed hope, that larger samples would help us detect and reject models whose true fit was just beyond the close-fit limit. Only if the RMSEA estimator is biased in that increasing sample size makes it tend to unjustifiably reject models truly under the close-fit limit, should we permit a sample-size adjustment to the close-fit target. This has not been shown to be a problem for the RMSEA, so we must object to any sample-size adjustment to the close-fit criterion. While the formula for the χ^2/d ratio implicitly contains n, because $\chi^2 = n\hat{F}$, this does not mean that an adjustment for n is required. For true models, the sampling distribution for χ^2 is not dependent on n (see Bollen, 1990). The .05 RMSEA criterion is not a statistically dictated bias correction. It is an elastic tape measure that four-steppers, and others, can stretch to let them accept models they would like to accept.[52]

The RMSEA has a known sampling distribution, and this would seem to permit the four-steppers to avoid the "favouring null" problem discussed above. By specifying a not-close-enough value of the RMSEA as a null hypothesis, the researcher could pit this against the alternative hypothesis that their model actually fits significantly better than (closer than) this not-close-enough fit. The researcher would therefore be favoring the alternative hypothesis, and hence would avoid the favoring-null problem.[53]

This would indeed be great news, but unfortunately something has gone wrong. An RMSEA value like .05 is not a criterion value; it is a specification of a sample-size adjustment. We saw in Figure 2 that an RMSEA of .05 specifies the degree to which one is willing to adjust one's criterion on the basis of sample size. So testing to see if one's RMSEA is significantly smaller than .05 is testing whether a significantly smaller sample-size adjustment (a less stretched elastic measuring tape) would have let one accept this model. It does not test whether the model RMSEA has passed some consistent close-fit criterion.[54]

So where does this leave four-steppers who propose using the RMSEA as a test? It leaves them arguing for close fit of zero, or within sampling fluctuations of zero, which is merely a way of pleading for permission to use a smaller chi-square probability than other researchers have traditionally been held to. The RMSEA can't help them solve the favoring-null problem. In fact, use of the RMSEA aggravates the favouring-null problem because it implicitly counts larger sampling fluctuations as being to the researcher's credit.

SUMMARY AND CONCLUSION

The four-step proposes that measurement can and should be assessed prior to structure during the assessment of structural equation models. Most of our comments question the separability or isolatability of measurement determinations from structural determinations, and hence question the utility of sequential testing of the proposed models.

We began by observing that the four-step is only applicable to confirmatory studies postulating nonsaturated structural models, and to studies in which the base model has at least four indicators for each and every concept. Furthermore, it is

[52] Browne and Cudeck (1993) and MacCallum, Browne, and Sugawara (1996) seemed not to recognize the connection between the RMSEA criterion and sample size. They based their target value on experience with models that χ^2 rejected but they wanted to accept. Consequently, they "streeeeetched" the unrecognized sample size adjustment until they got the desired conclusion. When MacCallum, Browne, and Sugawara (1996, p. 134) claimed that a particular value "yielded conclusions about model fit consistent with previous analyses of the same data sets," they were indicating what they would like to have found for these data sets, were it of course not for the bad news provided by that pesky old basic model chi-square.

Stan's .05 criterion stretches the adjustment pretty far, but not as far as MacCallum, Browne, and Sugawara (1996, p. 134), who "consider values in the range of 0.08 to 0.10 to indicate mediocre fit." We are waiting for the tautly stretched elastic measuring tape to snap.

[53] One could start with a null hypothesis claiming that the model fit is poor (say a RMSEA of .05 or more) and reject this assessment only if the observed RMSEA were significantly smaller (closer, or better fitting) than this. This is asking if the observed RMSEA is significantly below the .05 line in Figure 2. See MacCallum, Browne, and Sugawara (1996, pp. 135–136).

[54] The self-contradictoriness of this test becomes obvious if one considers trying to locate samples that are significantly lower than one of the diagonal lines in Figure 2, while gradually pulling down the right end of the line until it lies horizontally at a χ^2/d ratio of 1.0, namely to a position where there is no sample-size adjustment. Now note that the RMSEA is defined as zero (via Max) for all samples with ratios less than 1.0, so we will never find any evidence of close fit (in the absence of a sample-size adjustment) other than an RMSEA of exactly 0.0 via invocation of Max.

probably practicable with only medium-sized models. Too few concepts make it difficult to attain a reasonable degree of structural sparseness, whereas very many concepts, with four or more indicators each, demand huge sample sizes.

The four-step process is plagued by one fundamental problem and numerous operating problems. The fundamental problem is Step 1's inability to ascertain whether the researcher has the proper number of concepts. The mandatory addition of coefficients during construction of the Step-1 model introduces a host of anticipatedly unnecessary coefficients that make the Step-1 model prone to fitting with fewer factors than actually populate the real world. The problem seems to be that the procedural demand to add coefficients that convert the base model into an ordinary factor model simultaneously converts the base model into a model whose fundamental form is untestable. The Step-1 chi-square test assumes, but does not test, whether the factor model is the appropriate model. If the assumption is valid, then the chi-square test could help us determine the number of factors required, but if the real-world model is not a factor model, then the chi-square test can provide acceptable fit with a wildly wrong number of concepts.

Neither we nor anyone else knows for sure how often researchers have been, or will be, misled by factor models that fit with too few concepts. Unfortunately, there are two signs that make it appear that this may happen more, rather than less, frequently. First, the factor model is entirely incapable of detecting any number of concepts/factors beyond the number that saturate the model. Given that it takes only a relatively small number of factors to saturate the factor model (recall the many added coefficients), the factor model test is entirely incapable of protecting the researcher from, or informing the researcher about, unexpectedly complex real worlds—and the real world seems to quite consistently turn out to be unexpectedly complex.

Second, the factor model shows only a slow and progressive decline in fit when the factor model is the wrong model and there are slightly fewer than the saturating number of coefficients. That is, if the factor model includes way too few factors/concepts, the factor model chi-square test does not jump from saying there is no test when the model is saturated, to saying the model is definitely wrong when the factor model has a few (one, or two, or five) degrees of freedom. Losing a loading or five does not radically disrupt the factor model's ability to reproduce the covariance data; the loss of loadings merely degrades the ability slowly and progressively. This means that even if the factor model has a wildly low number of factors, the warning signs of this problem are not sharply demarcated between no model test (saturation) and some model test (some degrees of freedom). So the ability of the factor model to detect the specification of wildly wrong numbers of factors does not improve instantaneously as one moves from being "at" to being "below" the saturating number of estimated coefficients.

Together these signs are extremely worrisome. They suggest that the factor model chi-square test may be more than occasionally reporting fit or near fit, when in fact the model contains a radical misspecification of the number of factors.

The four-step's operating problems are primarily associated with testing. We began our critique of testing by noting the inappropriateness of using .05 as the critical probability differentiating model fit from failure. The favoring of the null hypothesis of the test argues that a substantially larger target probability should be used. Unfortunately, Stan seems inclined to use a smaller, not larger, probability because even the .05 criterion rejects more models than he likes.

Using the RMSEA as a test of "close fit" would indeed lower the probability criterion, but this has a couple of problems, in addition to exacerbating the favoring-null problem. First, the proposed target value of .05 for the RMSEA is a sample-size adjustment, and it is not a target value. Second, the logic of "close fit" is at odds with the logic of the four-step, which demands that Step 1 locate the proper number of factors, not "close to" the proper number of factors, if one is to say anything definitive about measurement.

Then we have the no-peeking conundrum, where the researcher is not supposed to look at the estimates from the prior steps, lest he or she contaminate the later tests, yet where those prior estimates may contain useful diagnostic information. So the researcher either sacrifices the purity of his or her testing or becomes oblivious to potentially useful diagnostic information.

Then we encountered the problem that researchers with even a dash of dedication to theory will want to have the concepts in their models behave in ways that make them "the proper concepts" and not just any old concepts. To know, for example, that six concepts in an arrangement might be able to match up with the data is of limited interest to a researcher whose Step-3 model indicates the researcher's specific six concepts do not fit with the data. If "these" specific 6 concepts were chosen because they were of interest to a discipline, it seems the researcher would be well advised to publish and highlight the failure of the disciplinarily relevant 6 concepts, leaving open whether the resolution of the discipline's problems will ultimately appeal to another 6 concepts, or to 7, or to 10 concepts. A sparse 10-concept model can be even more parsimonious than a 6-factor model, so there seems to be no reason to automatically favor any other 6-concept models over models with more concepts.

We do not wish to leave the impression that we are against the use of nested models, which are at the heart of the four-step process. In fact, we encourage researchers to develop and employ whatever nested models provide the clearest assessment of whatever points are relevant to their particular literature. Where we depart from the four-step is that we do not anticipate that a specific nesting sequence is likely to be so routinely informative that it warrants routine application.

One could, for example, start with a base model and saturate the loadings, leaving the structural effects as in the base model, to get to another style of Step-2 nested model. This would impose or demand a structural theory prior to a measurement determination, so that one would be seeking measures of entities that behaved in appropriate theoretical ways. This approach suggests we would be better able to assess measurement if that assessment is made in the context of the relevant theoretical distinctions. But even this is not enough to prompt us to suggest routine use of this alternative style of nesting.

Once one starts thinking this way, one quickly notices that any nesting employed should be dependent on what the researcher wants to know. To continue the example just given, we might anticipate that only some, and not all, of the concept-to-item loadings might be used to create the nested model. Which specific concepts/factors might be especially likely to influence a wider span of items will be open to reassessment in each research context, so here again we will probably find that there is no single best nesting strategy.

If scientific history is our guide, we might as well stop anticipating that measurement and structural components will ever be disentangled at the cutting edge of science. Indeed, it is a laudable scientific contribution to have used a sparse structural theory to assist in disentangling a dense and complex measurement model (Hayduk, 1996, pp. 53–54, 69). The task is not so much to find a minimal number of concepts as it is to find ways of clearly distinguishing and respecting the uniqueness of each concept. The structural model becomes substantially sparse out of respect for maintaining the distinguishing, indeed defining, *actions* of the concepts. From this perspective, the four-step's move from Step 3 to Step 2 to Step 1 progressively robs the model of whatever theoretical precision the intentionally unique conceptual effects had to offer. This leaves the Step-1 model to try to determine the proper number of concepts in the absence of theoretical control. One gets back to the Step-1 factor model, where the minimum number is king only because someone killed the theory queen. At the scientific monarch's ball, we would prefer you not waltz the four-step; instead leap the fore-step.

ACKNOWLEDGMENTS

Dale N. Glaser now at Pacific Science & Engineering Group, San Diego, CA.

We gratefully acknowledge Stanley Mulaik's willingness to unabashedly dance the four-step in public during the SEMNET discussions of 1997 and 1998. His dedication was much appreciated because his comments focused on many of the debates and clarified many issues. His efforts were more than sufficient to have made him an author of this paper, had his views not been at odds with much of the content. For your efforts, Stan, our many thanks. We also thank N. M. Lalu for running some of the factor analyses reported herein.

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, J. C., & Gerbing, D. W. (1992). Assumptions and comparative strengths of the two-step approach: Comment on Fornell and Yi. *Sociological Methods and Research, 20,* 321–333.

Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin, 107,* 256–259.

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

Burt, R. S. (1973). Confirmatory factor-analytic structures and the theory construction process (plus corrigenda). *Sociological Methods and Research, 2,* 131–190.

Burt, R. S. (1976). Interpretational confounding of unobserved variables in structural equation models. *Sociological Methods and Research, 5,* 3–51.

Fornell, C., & Yi, Y. (1992a). Assumptions of the two-step approach to latent variable modeling. *Sociological Methods and Research, 20,* 291–320.

Fornell, C., & Yi, Y. (1992b). Assumptions of the two-step approach: Reply to Anderson and Gerbing. *Sociological Methods and Research, 20,* 334–339.

Hayduk, L. A. (1985). Personal space: The conceptual and measurement implications of structural equation models. *Canadian Journal of Behavioural Science, 17,* 140–149.

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

Hayduk, L. A. (1994). Personal space: Understanding the simplex model. *Journal of Nonverbal Behavior, 18,* 245–260.

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

Hayduk, L. A., & Avakame, E. F. (1990/1991). Modeling the deterrent effect of sanctions on family violence: Some variations on the deterrence theme. *Criminometrica, 6/7,* 19–37.

Hayduk, L. A., Stratkotter, R. F., & Rovers, M. W. (1997). Sexual orientation and the willingness of Catholic seminary students to conform to church teachings. *Journal for the Scientific Study of Religion, 36,* 455–467.

Herting, J. R., & Costner, H. L. (1985). Respecification in multiple indicator models. In H. M. Blalock (Ed.), *Causal models in the social sciences* (2nd ed., pp. 321–393). New York: Aldine.

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

Jöreskog, K. G. (1962). On the statistical treatment of residuals in factor analysis. *Psychometrika, 27,* 335–354.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32,* 443–482.

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage.

Jöreskog, K. G., & Sörbom, D. (1976). *LISREL-III: Estimation of linear structural equation systems by maximum likelihood methods.* Chicago: National Educational Resources, Inc.

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

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

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

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square tests. *Psychometrika, 50,* 253–264.

Wonnacott, R. J., & Wonnacott, T. H. (1970). *Econometrics.* New York: Wiley.