ı requires a zero correlation between factors F1,F2 and F2,F4 when the latter is . Clearly the reformulated model makes a stronger demand on fitting the data, t is, that a .23 correlation is zero versus a .03 correlation is zero. Each structural ·del, in this case, implies a very different CFA in which different dimensions are :n as orthogonal to each other. Given that the two imply different measurement ·dels, one might expect differences in estimates.

Simply adding the covariance between F1 and F2 in either structural model ·uld imply that all the factors covary in both the CFA and the structural model. so doing, one finds that altering theory as Fornell and Yi (1992) did (i.e., re- :sing the roles of F3 and F4) leads to basically no change in the parameter esti- ıtes. This is not to say we will get exactly the same estimates for parameters ›resenting "measurement characteristics" (e.g., reliabilities) under both struc- al models, because one of the structural models may better replicate the pattern observed covariances among the factors obtained in the CFA, plus some sspecifications remain that do not specifically replicate the covariances among :tors. Clearly, misspecifications or ill fit may manifest themselves in changes in : estimated values for both structural parameters and measurement parameters. fact, major changes from the parameter estimates of the loadings in the CFA st stage to the parameter estimates of the loadings in the structural equation ›del may be indicative of misspecifications in the structural model. Table A1 ›ws the initial standardized loadings from the CFA reported by Anderson and ırus (1984), the results reported by Fornell and Yi when altering the theory, and : results on performing the same "theory change" when appropriately incorpo- ing the correlation between F1 and F2.

STRUCTURAL EQUATION MODELING, 7(1), 111–123

Doing the Four-Step, Right–2–3, Wrong–2–3: A Brief Reply to Mulaik and Millsap; Bollen; Bentler; and Herting and Costner†

Leslie A. Hayduk
Department of Sociology
University of Alberta

Dale N. Glaser
Sharp Health Care and
San Diego State University

Our rejoinder to the preceding commentaries worked systematically through each comment, rebutting points in sequence and in context. Unfortunately, this full reply was three times longer than *SEM* was able to publish, so we prepared this compacted rejoinder by focusing on a few central points.[1]

MULAIK AND MILLSAP

On first reading, Mulaik and Millsap's (M&M, 2000) comment sounds wonderful. On first thinking, it remains unconvincing. The problems begin right from M&M's first paragraph, which tries to connect the Hayduk and Glaser (H&G, 2000) critique to Les Hayduk's 1996 book. Unfortunately, this contains serious misconceptions that contaminate several of M&M's subsequent comments. M&M mistakenly claim that Hayduk "feels" fixed measurement error variances are "incompatible with the factor analytic underpinnings of the four-step procedure," and that fixed

---

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

†Our full rejoinder is available via a link from Hayduk at http://www.ualberta.ca/~socweb/

[1]Our full response is linked to Hayduk at http://www.ualberta.ca/~socweb/

measurement parameters constitute a "general alternative to the use of multiple indicators" (M&M, p. 36). Hayduk in fact views fixed measurement error variances as compatible with (not as an alternative to) multiple indicators, and he specifically discusses how to integrate their use with multiple indicators (Hayduk, 1996, pp. 25–30).

M&M claim that "Hayduk also believed the usual .05 level of significance in testing the exact fit of models favors the null hypothesis" (M&M, p. 36). Hayduk (1996, pp. 69, 77; Hayduk & Glaser) in fact argues the *reverse* of this. It is the favoring of the null hypothesis that leads to questioning whether the usual .05 level of significance should be used in testing model fit. We find it entirely unsurprising that M&M found the recommendation to alter the criterion incoherent, given that they got the logic backwards!

M&M's Figure 1 defends the use of at least four indicators per concept. There are, however, clear counterexamples to their demonstration. According to M&M, "the correlation between x_4 and each of the indicators of η is zero" (M&M, 2000, p. 42). The correlations between x_1, x_2, x_3, and the indicators of η are nonzero. Together these imply that models similar to M&M's Figure 1b, but measuring c with the pairs of indicators x_1 and x_4, or x_2 and x_4, or x_3 and x_4, or any triplicate like x_1, x_2, and x_4, would fail and provide clear diagnostic warnings that the measurement of c was problematic. Furthermore, a model using x_4 as a single indicator, with Hayduk's (1987, 1996) fixed theta procedure, could result in a correct determination of c's action in M&M's Figure 1a model. So models with one, two, or three indicators are capable of detecting either the problematic confounding of method m's effects among the several indicators, or c's true action. The use of even four indicators is "not infallible" (M&M, p. 42), and neither are one, two, or three indicators, but this should not detract from the overall point, that while four indicators are the minimum for testing a single common factor *in isolation*, once one introduces more concepts with their indicators, the minimum number of indicators required for testing is reduced substantially.

The most central failing of the four-step concerns the Step-1 model's use of "the same number of factors as the number of latent variables" in the Step-3 model (M&M, p. 43). The key observation is that while the two numbers may be the same, the factors and the latents are not, so that Step 1 and Step 3 are including the same number of different things.

The freeing of coefficients in moving from the Step-3 model to the Step-1 factor model loosens the model's control on what the latents originally (in the Step-3 model) stood for and against. Loosening the theoretical constraints permits the factors to become variables that do not correspond to the latents that the theory specified as belonging in Step 3. The Step-1 factors are permitted effects that theory claims the latents do not deserve, and correlations that theory claims should be impossible. These are key defining features of latent variables, and to alter these by

freeing the loadings and factor correlations alters the meaning of the latent. The latents ought to be constrained by (confounded with) theory because this is what forces the latents to be the correct latents to include in the model.

This is central to why Jöreskog's "proof" provides no support for the four-step. The proof begins with "Given the common factor model" (M&M, p. 44). To appeal to this proof is to make the unfounded assumption that the entire loss of theoretical and methodological control on the identity of the latent concepts does not permit the latents to change into anything other than the variables postulated by the theory grounding Step 3. The procedural manufacturing of connections between latents, and between items and latents, to construct the Step-1 model, frees the latents of Step 1 to become variables that are inconsistent with the theoretical and methodological dictates of the original model. Without the subtle, yet erroneous, conflation of concepts with factors, Jöreskog's proof is disconnected from the Step-3 model, and so are M&M's Appendix proof and their simulation of common factor models having a "known number of factors" (M&M, p. 46).

M&M view the factor model as testing "a specified number of common factors k without confounding that issue with 'measurement' issues about the specific relations of indicators to latent variables or structural relations between latent variables" (M&M, p. 43). What M&M view as confounding, we view as necessary and useful methodology and theorizing that provides the latents their identity. The latent–latent and latent–indicator relationships, and specified lack of relationships, are theoretical and methodological assertions that force or constrain the latent variables to be precisely the proper latents to include in the model. From the perspective of a researcher interested in the Step-3 model, it is irrelevant to know if some model containing the same number of another kind of entity might manage to fit the data.

A second perspective on this issue arises when M&M inject the term "provisional" into their discussion of the hypothesis tested by Step 1. The question concerns whether Step 1 can determine if the base model contains the proper number of latents (M&M, pp. 54–56). The context for this discussion is an imagined series of Step-1 factor models using 1, 2, 3, 4, 5, 6, 7, 8, ... factors for a given set of indicators. For a small number of factors (say 1 or 2 factors), the chi-square test indicates that the model fails to account for the indicator covariances. A larger number of factors (say 3, 4, 5, 6, or 7) will fit the data. Eventually (say at 8), the growing number of factors results in the model being underidentified because there are more coefficients to estimate than data points, so no test results are available.[2] Let us further imagine that there are five latent variables in the researcher's Step-3 model, so the Step-1 factor model with its five factors fits and passes the Step-1 test.

[2]It is possible to have models whose true number of underlying latent variables are among this underidentified set of Step-1 factor models, and these are what M&M are trying to avoid by demanding at least four indicators per factor.

So what does passing the Step-1 test inform the researcher about? The Step-1 fit does not inform the researcher that he or she has the proper number of factors, because any Step-3 model with 3, 4, 5, 6, or 7 factors would have passed the Step-1 fit test, even though only one number might claim to be the proper number. According to M&M, "Step 1 is not designed 'to locate the correct number of factors' but to test the provisional correctness of a hypothesis regarding the number of factors" (p. 55). So, in our example, 3 through 7 factors are all to be accepted as provisionally correct numbers of factors. But in what sense have we even provisionally *tested* a hypothesis about the proper number of factors if the boundary between the test's failing or fitting has no specifiable correspondence to the boundary between a proper and improper number of factors?

A model with as few as three factors can fit, but the minimum number of fitting factors can not be defended as routinely corresponding to the proper number.[3] We agree that since factor models with one and two factors would fail, the fitting five-factor model has survived a challenge some other models might fail. But the issue is not the fact that our model passes Step 1. The issue concerns what passing the fail–fit boundary for the Step-1 test permits us to conclude.

Earlier in their comment, M&M tell us that after fit at Step 1, even with failure at Step 2, "you would still not have the worry that you had the number of factors wrong, [i.e., that lack of fit was due to specifying too few latent variables]" (M&M, p. 43). So, for our example of fit with 5 factors, we are supposed to conclude that 3, 4, 6, or 7 could not be the proper number of factors! We simply do not think that fit at 5 factors justifies such a conclusion. Notice that had we begun with a Step-3 model containing 4 latents, we would also have passed the Step-1 test, so M&M's statement would instruct us that 5 could just as well be included among the wrong number of factors. According to M&M we should not "worry" that we had the number wrong if we passed Step 1 with 3 or 7 factors either![4]

M&M's "provisional testing" idea camouflages a chasm between what the Step-1 test does and what M&M's four-step ideas corner them into claiming it does. M&M's way of asking why this chasm exists was to request that we "identify some inherent mathematical or statistical feature of the common factor model that renders it inadequate as a tool for identifying the number of factors" (M&M, p. 58). Unfortunately for factor analysis, the feature that allows too few factors to coordinate many items is a fundamental feature that is inherent to the factor model. It is precisely the same feature that loosens the model's control on the identity of the latent variables. The Achilles heel of factor analysis is that each latent spans all, or nearly all, the items in the Step-1 model. Fit with the Step-1 factor model is prone

[3]See the SEMNET discussions leading up to Stan's postings of 4/16/97 and 4/17/97.

[4]Anyone inclined to argue that there is no such thing as a proper–correct number of latents will be in deep trouble trying to defend M&M's Step-1 test because he or she will have to argue that Step 1 tests the "provisional correctness of a hypothesis" about something he or she claims does not exist.

to being wrongfully achieved because the four-step procedure demands the insertion of anticipatedly unnecessary item-spanning loadings. The four-step's procedural insertion of spanning loadings, and free factor correlations, are the culprits, and it is these same culprits that make the latents and factors different things.[5]

We doubt that our identifying "the feature" will make much difference to the four-step debate. The feature is in one sense just an obvious fundamental and unavoidable feature of the factor model. About all our identification of "the feature" will do is rob M&M of the gratuitous assertion that Les's prior refusal[6] to assert a problematic feature was because the "claim is false" (M&M, p. 58). The real news is that essential features of the factor model make it impossible for the procedurally constructed Step-1 model to routinely inform us about the proper number of latent variables.

A third perspective on the same big issue arises in the context of M&M's discussion of their Figures 2, 3, and 4, where they explain why a factor model with 3 or more factors will fit data matrices created by a real world composed of 10 latents in a simplex chain. M&M fail to appreciate how their own discussion undermines their use of Jöreskog's proof.[7] They demonstrate why, in a particular instance, the Step-1 factor model fit–fail boundary is radically removed from the true number of underlying latents.[8] That is, M&M's simplex-factor discussion exemplifies why it is, that conflating factors and latent concepts, despite its necessity for their interpretation of Step-1 fit, is unjustified.

M&M seemingly do not yet appreciate that the simplex-factor failure is not an isolated instance. It is merely one of many demonstrations[9] that the fit–fail

[5]M&M are incorrect in thinking that our "quarrel is not with the common factor model but with tests for lack of fit in general" (M&M, p. 58). Our quarrel *is* with the factor model when that model is constructed or dictated by procedural admonitions and not theoretical–substantive considerations. It is the procedural insertion of theoretically unnecessary, and even theoretically precluded, coefficients, that permits improperly achieved close covariance fit. This is nothing any test of fit could recover from.

[6]Contrast this with Hayduk SEMNET 3/20/98 12:55 p.m.: "I think/hope we will eventually get to 'improperly freed coefficients' as the culprits." The numerous theoretically vacuous, yet procedurally freed, loadings and latent correlations are the culprits.

[7]The underlying model in this instance is "known" not to be a factor model (as the proof assumes), but is some other model. In fact, with 10 indicators, a 10-factor model is severely underidentified, so M&M's simplex discussion also inadvertently demonstrates that there are instances when none of the factor models that fit would have the proper number of latents.

[8]It is possible to locate 10 latents in such a model, if one has a consistent dedication to theory (for an example see H&G). The freeing of coefficients in moving to M&M's Step 1 is a direct attack on the requisite theoretical constraints.

[9]For one instance, ask M&M about their simulation output where "models that specified fewer factors than the number that generated the covariance matrix that sometimes very large samples were essential to detect that the number of factors had been underspecified" (M&M, p. 46). This says that even with reasonable *N*, and even if the true model is *another* factor model, M&M find that the Step-1 factor model can't be trusted to locate the proper number of factors. For another example, see the Simplex+8.2 discussed in H&G (2000). See Hayduk SEMNET 2/20/99 11:51 a.m. for yet another example. Then there is the General1 model surrounding Hayduk SEMNET 11/19/98 6:10 p.m.

boundary of the Step-1 factor model has little to do with the number of latents required in the true Step-3 model. We suggest you read the corresponding several pages of seemingly technical stuff as being like a distracter on a multiple-choice exam. Most of the statements are correct, but the overall point is reversed to trap the unwary.

M&M's section on "The Problem of Models With More Latent Variables" (pp. 47–49) repeats a slight-of-expression that contains the same essential inconsistency. The issue appears this time as the difference between modeling the world out there and fitting the data. The four-step begins with a researcher trying to understand the inadequacies of the Step-3 model as a representation of the world "out there." In contrast, M&M see Step 1 as a data-fitting task.[10] Inattention to the difference has ugly consequences. If all one focuses on is the data at hand, there is no true number of latent variables "out there" to act as a benchmark against which to judge the number of factors used to fit the data. By losing sight of (perhaps hiding) the real world out there, M&M overlook (perhaps hide) the possibility that great Step-1 fit should be read as extremely bad and misleading news if that fit is achieved with fewer factors than there are latents in the real world.

M&M's real world has "peek-a-boo" existence: Now you see it in Step 3, now you don't in Step 1. Unfortunately, they can have it only one way, and that way means acknowledging the Step-3 world out there, because there is no four-step procedure without the Step-3 model that initiates the exercise. The worldless, mere data-fitting idea must be dropped if the four-step is to retain its logical integrity. When M&M find that 3 correlated factors are capable of fitting data generated by 10 latents, they pass this off as Step 1's failure to "accurately represent a subtly more complex situation" (M&M, p. 47). Yes, the fit difference is subtle, but the difference between a real world with 10 latents or 3 factors is not. The clanging of a subtle fit difference against a substantial factor-latent difference tolls the passing of the four-step.

To close our discussion of M&M, we enumerate a few of the points deleted due to space limitations:

1. M&M's granularly salted measurement (pp. 49–50) conflicts with their four-step claims by equivocating on what Step 1 tests.
2. M&M's own criteria (p. 54) force Step 1 (but not Step 3) to require prohibitively large samples.
3. Starting from a saturated factor model (M&M, p. 59) and working backward through slowly degrading fit as more coefficients are fixed provides another way to see why factor models fit with the wrong number of latents.
4. What reasoning did M&M use to connect favoring null to the need to change the .05 criterion?

[10]See also Mulaik SEMNET 4/30/98 12:52 p.m.

5. M&M's appeals to naturalness is an appeal to factor analytic tradition, as is illustrated by a similarly nested, yet naturally overlooked, Step-2* model that reinstates the structural/theory constraints while preserving the spanning loadings (briefly mentioned in H&G, pp. 33–34, but expanded in our full reply).
6. *The* four-step procedure is not useable with models having one or two indicators of latent variables, unless M&M's Step 1 can determine the proper number of latents with only one or two indicators per latent.

M&M feel "the four-step procedure has easily survived" our critique (p. 70). We are equally confident that the four-step has been mortally wounded and ought not be used. This leaves the decision about the current state of the four-step's health/illness where it belongs, namely on your—the initially neutral reader's—shoulders.

We turn now to the comments of the other discussants, following the same compacted presentation style. Our full response is available as indicated at the beginning of this rejoinder.

BOLLEN

Regarding whether we "question the use of the chi-square test statistic" (Bollen, p. 75), we in fact favor the use of χ^2, though we object to the acceptance–rejection boundary used by M&M. We also object to the interpretation of Step-1 factor fit as indicating a proper number of latents, whether that fit is adjudicated by chi square or another test. The reason the four-step should be singled out, when no other procedure provides a routine way to locate the proper number of latents, is that no other procedure claims to be able to do this, though M&M do. Step-1 fit followed by later failure is supposed to mean "you would still not have the worry that you had the number of factors wrong" (M&M, p. 43, and Mulaik SEMNET 4/4/97 1:12 a.m.). We agree that determining the proper number of latents is particularly difficult, as Bollen's examples illustrate. Since Bollen's observations question the assertion quoted above, they challenge the four-step, not our critique of the four-step.

With regard to Bollen's jigsaw idea, consider a jigsaw procedure that begins by respecting the latent (structural) model, in contrast to Bollen's jigsaw that begins with no latent/structural theoretical constraints. Our jigsaw starts out with the entire postulated latent (structural) model and uses a single indicator for each concept (Hayduk, 1996, pp. 25–30). We then add one indicator at a time and look to see if "the expanded model still fits" (Bollen, p. 80). There are a variety of ways of searching through the indicators (e.g., use all the indicators of one latent first, or try for two indicators of each latent before using a third indicator for any latent, etc.) and, hence, we should be able to locate the indicators whose "obscuring spurious or suppressor relations were missed by treating the" indicators "separately" (to paraphrase Bollen, p. 80). Note that both Bollen's jigsaw, and our theory-respecting jigsaw, are incompatible with the four-step, and, hence, are unavailable to four-step devotees. We agree that "Embracing set modeling strategies or steps discourages researchers from thinking about specification and creates a dependence on more mechanical procedures that are far from problem free" (p. 81).

BENTLER

Bentler observes that, for preplanned nested series of models, the initial model chi square is independent of subsequent chi-square-difference tests, but is not independent of subsequent overall model chi-square tests. This argues against M&M's use of model chi squares for Steps 2 and 3, since a model chi square is used at Step 1. M&M's four-step procedure requires model tests, so any improvement in the statistical purity of testing that could be garnered by using subsequent chi-square-difference tests would demand substantial revision of the four-step procedure.

Bentler makes two further observations about chi square: namely that the single model chi-square test that precedes a series of difference chi squares should begin by testing the least constrained model; and that "there is no reason to expect that the standard .05 level will lead to a generally correct error rate" (Bentler, p. 84) if a complex model evaluation sequence is used. So the four-step loses one kind of statistical purity (the Step-3 test has an unknown error rate) because it uses a series of tests beginning with Step 1. Both kinds of statistical purity are lost (both of Bentler's objections apply) if one starts the series of tests at the more constrained Step-3 model. Since it is the series of tests that creates these problems, the solution seems clear. Do the base model Step-3 test first (thereby preserving a clean Type I error rate), and do not engage in any procedurally dictated series of tests.[11]

We got a reciprocal chuckle[12] out of Bentler's self-opposing claims that because the LISREL "model is built literally of two parts, it certainly makes sense to me to first test whether the general FA structure ... is plausible" (Bentler, p. 85);[13]

[11]Do have substantively dictated preplanned tests in mind—preferably tests constituting full alternative models, or entirely different substantive model segments (just in case the model fails, or possibly to illustrate the potentially equal abilities of competing models). But do not follow a generic matrix-dictated series of tests.

[12]We initially read the paragraph outlining Bentler's chuckle (p. 85) as being a laugh at Les's expense. A more careful reading suggests that Bentler was indirectly complimenting Les for his independence from the confines of measurement/structure distinction that is artificially built into the LISREL model, despite the term *LISREL* appearing in the titles of Les's SEM books. Bentler seemed to enjoy Les's breaching LISREL's representation mold and thinking like an EQSer. Our chuckle comes from Bentler's, fortunately temporary, entrapment in LISREL's measurement–structure distinction despite the absence of this distinction from Bentler's own system, as we discuss below.

[13]Specifying two parts does not defend which of the two deserves to be treated first. Bentler's selected sequence is stated without any rationale, but we will not press this point because the fundamental inconsistency in Bentler's reasoning renders this moot.

while in "the Bentler–Weeks model, there is no special distinction given to the 'measurement' and 'structural' models ... [consequently] the H&G objection to starting a four-step procedure with an FA model is entirely rational, and I would agree with it" (Bentler, pp. 85–86). It seems that Bentler agrees with us when he uses his own EQS style of model representation, and only disagrees with us when he adopts the notation style used by EQS's major competitor! What wonderful irony. Naturally, we would prefer that Bentler stick with his way!

Contrary to Bentler's assertion, his discussion of the Gold Standard indicator G (pp. 87–89) suffers a great "loss of generality" due to his choice of a model having one factor and 101 indicators. This model discards all structural theory because the latent has nothing to influence, or to be influenced by. We noticed the loss of theory—apparently Bentler did not. We agree that researchers are not likely to find the required value for the measurement error variance in the "available ... real research" (Bentler, p. 88). The required value is to be found in the researcher's available and real *theory*—which seems to be a place that is unavailable to Bentler. While Bentler views the 101st free measurement error variance as a "reasonable" admission of "lack of knowledge" (p. 88), we see the freeing of this coefficient as a ploy to keep the researcher from being precise about his or her intended conceptualization, and as avoiding a strong test of that intended meaning.

Yes, the "question of factor meaning has been tackled in the literature as the question of factor indeterminacy" (Bentler, p. 88), but these discussions are being turned into a historical backwater by researchers who have come to appreciate that an ounce of precision in their assertions about their latents is worth more than all the weighty discussions of factor indeterminacy. Yes, the "meaning of a factor cannot be stated with precision until there are enough indicators" (Bentler, p. 88), but the meaning of a latent CAN be stated with precision if there is enough theory and even one indicator.[14]

Bentler really does mean it when he says, "Personally, I do not believe that testing mean and covariance structure models necessarily requires precise knowledge of what the latent variables 'really' are" (Bentler, p. 89). We simply ask, what is

[14]We read Bentler's Gold Standard discussion as directed more at Hayduk (1996) than at H&G (2000), but even here it is way off the mark. In Hayduk's procedure the measurement error variance is determined on the basis of theory, and how well that theory coordinates with the methodology used in obtaining the indicators. When Bentler says "the requisite knowledge seems unlikely to be available in real research" (p. 88), he is either saying researchers routinely do not know their theory and methodology, or he is missing the point of Hayduk (1987, pp. 119–122; 1996, pp. 25–30). With even 100 indicators all doing their best, Bentler would not permit, let alone encourage, researchers to specify their theory more precisely by asserting how even one single (possibly iron or bronze) indicator ought to behave with respect to the latent variable.

the point of Step 3 if the researcher has not done his or her utmost to control the concepts to be the proper theoretical entities?[15]

HERTING AND COSTNER

The comment by Herting and Costner (2000, hereafter H&C) strays considerably from the four-step advocated by M&M and Mulaik on SEMNET.[16] The H&C two-step makes no use of M&M's Step 1 (H&C, p. 95), and ignores M&M's Step 4. H&C's second step corresponds to the base model the researcher is interested in (M&M's Step 3), but H&C's first-step confirmatory factor analysis (CFA) does not correspond precisely to M&M's Step-2 CFA. H&C's Step 1 is somewhere between Steps 2 and 3 of the M&M four-step because H&C feel that, if exogenous latents are specified as orthogonal in the base model, those latents should appear as orthogonal in the CFA Step-1 model (H&C, Footnote 2), which contrasts with the free factor covariances in M&M's Step 2.[17]

Another major difference is that H&C do not defend the fit–fail boundary at their Step 1 as warranting the kind of conclusion that M&M demand for their CFA. H&C focus on using the diagnostics of the CFA step to locate model misspecification and they consequently pay less attention to any specific conclusion as resulting from sequential model fit or failure. H&C's diagnostic slant makes them, like H&G, open to rethinking the number of factors even if the first sign of failure appears in the final model (M&M's Step 3, H&C's Step 2). H&C are convinced that neither M&M's Step 1 nor Step 2 are trustworthy guides to determining the number of latents. So, in many ways, H&C are waltzing around, not with, M&M's four-step.[18]

[15]Bentler misreads us when he says "H&G believe that the simplex, or variants of it are much more likely to occur in practice than do M&M" (p. 86). Unlike the factor model proponents, we see no reason to assume that nature is biased toward obeying the dictates of any single style of structural equation model. The Simplex, SimplexPlus, and General1 models discussed on SEMNET were purposefully different examples intended to illustrate a range of circumstances in which the factor model is prone to deceiving by indicating factor fit despite containing the wrong number of latent variables.

[16]The reason H&G (2000) did not address any other two-step approach (H&C, p. 96) was that we attempted to comply with the editor's invitation to encapsulate the SEMNET discussions between Les and Stan on the four-step.

[17]This leaves H&C in the awkward position of arguing that some constraints on latent variables should be respected in both the base and their Step-1 model (the orthogonal exogenous latents, according to H&C, Footnote 2), whereas other constraints on the latents should be included only in the base model. H&C will have to provide some justification for why exogenous–endogenous latent orthogonalities or constraints leading to near orthogonalities are not to be similarly preserved. This could be pursued in the context of H&C's Appendix discussion (H&C, pp. 108–110) where different exogenous–endogenous latent combinations become orthogonal, depending on model specification. Notice that H&C's Appendix resolves the problematic behavior of loadings by adding a covariance between exogenous latents in a CFA. By "simply adding the covariance" (H&C, p. 110), H&C contradict the procedural advice they provided in their Footnote 2.

H&C feel that "it is clear that if a CFA model cannot be satisfactorily fitted, moving to the structural model will provide no additional guidance or benefit" (H&C, p. 100). We disagree. More constrained models will provide larger and differentially patterned modification indexes. The full picture of all a model's failings may be more useful in resolving the tough problems of specification, like having an improper number of latents.[19]

We can summarize H&C simulation results by modifying H&C's conclusions a tad (note added italics).[20]

> [These] results indicate that this search procedure may be useful in achieving a better fit between model and data in certain circumstances [*namely when there is more than one specification error*], but as the number of specification errors increases [*to more than two, or if the two become intertwined*], the clues provided become less and less useful in accurately locating the specification errors. (H&C, p. 107)

Even this minimal degree of two-step advantage was achieved *after* H&C's simulation robbed the one-step of its greatest strength, namely a demanded coordination between the base model's diagnostics and the researcher's understanding of the relevant substantive theory and methodology,[21] and *after* using stopping criteria different than any of us would be permitted to use,[22] and *after* using a test model loaded

[18]H&C view the preliminary determination of the proper number of latents as not being required for diagnostic assessment. It is M&M who propose this in their claims for their Step 1, and so this way of "formulating the issue" (H&C, p. 93) was forced on H&G. The paragraphs in H&C's section "What Is the Proper Number of Factors?" (H&C, p. 93) are in direct opposition to Step 1 of M&M's four-step and are not a challenge to H&G. To see this, note how descriptive the H&C sentence beginning "This is tantamount to …" is of factor analysis and its emphasis on data fit.

[19]For a discussion of the implications of knowing some versus all a model's constraints, see the SEMNET discussion between Les and Stan spanning 11/18/98 to 11/28/98.

[20]We read the first paragraph of H&C's conclusion as repeating H&C's opposition to M&M's claims regarding their Step-1 factor analysis, but as being neutral with respect to how best to proceed.

[21]H&C had no way to simulate a substantive appeal to theoretical or methodological knowledge, so in their simulation the one-step is, with a great loss of generality, reduced to the freeing of the largest modification index.

[22]H&C (pp. 104–105) describe models whose chi-square probabilities are .005 and .006 as being sufficiently good as to require no further emendation. Claiming unacceptable fit in these instances would have meant adding unnecessary coefficients and consequently degraded the performance of H&C's two-step. Real researchers will not be able to emulate H&C's feat because using even the traditional fit criteria would result in adding more coefficients at both Step 1 and Step 2, and this would make H&C's two-step fail more frequently.

H&C sidestepped whether favoring null should result in use of a larger fit probability criterion, but they are clear that sequential testing forces them to abandon "the formal use of tests" (p. 105). They fail to mention that abandoning formal testing attacks their ability to provide "followable" procedural recwith measurement degrees of freedom (testable assertiveness) but short on structural degrees of freedom,[23] and with other limitations.[24]

SOME COLLECTIVE COMMENTS

Here are a few observations on how the discussants collectively responded to the key themes of H&G, and what their responses indicate.

First, does the factor analytic Step 1 inform us about the proper number of factors? M&M vote yes; H&G, Bollen, Bentler, and H&C vote no. This clearly leaves the onus on M&M to find some new way to convince us of this. Without some new defense, the four-step is doomed because the current justifications of the steps are untenable if consistently qualified by a cautionary note warning that "there is no convincing evidence that this model contains the proper number of latents."

It is not only four-step advocates who will be challenged by the developing consensus that a fitting factor model does not provide convincing evidence about the number of underlying latent variables. If no new and more general defense of this claim is forthcoming, we will be watching the last slow dance of factor analysis. Academics sometimes moves at the pace of retirements.

"Does the favoring of the null hypothesis recommend changing the probability level for chi-square fit test?" To M&M the argument was incoherent, which is unsurprising given that they got the logic wrong. H&G vote yes. Silence from everyone else.

"Is a root mean squared error of approximation (RMSEA) value of .05 indicative of acceptable fit?" H&G vote no. Silence from everyone else.

The silence on the RMSEA issue probably means people want to think about this a bit more, since H&G provided a new way of looking at this. The collective silence of the neutral parties regarding the need for a much elevated probability level for an acceptable chi square is quite a different matter. This claim has been around for some time (namely since Hayduk, 1996, p. 69), so people have had time to think up objections. Failure to elicit any is therefore on much firmer ground. So beware. If your SEM model has a chi square probability of even .1, this will now be

---

ommendations. This robs them of a ruleful or procedural way of specifying when to stop doing modifications at both their Step 1 and Step 2. If H&C fail to instruct us on how much to adjust the probability limit in compensation for each prior test, there will be two steps, but there will be no procedure for doing those steps.

[23]In H&C's final Step-1 model the latent structural part of the model provides only 2 of the 45 *df*. When H&C consider models containing latents with single indicators, the structural side of the diagnostic balance will improve relative to the measurement side, and the utility of H&C's measurement Step 1 will decline below its current mediocre performance.

[24]H&C focus on errors of omission (excess misspecified coefficients would hamper their Step 1); reversed arrows are mentioned but not simulated; and concepts with single indicators (age, sex, income, education) are ignored.

read as a probably failing model, so you are exposing yourself to attack by reviewers if you interpret this as saying your model fits.

"Should researchers do the four-step or not?" The vote on this is 1 to nothing. Your lone vote wins! We would prefer you jive the four-step, rather than waltz with it, but do whatever you think will help you make real research progress.

ACKNOWLEDGMENT

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

REFERENCES

Bentler, P. M. (2000). Rites, wrongs, and gold in model testing. *Structural Equation Modeling, 7,* 82–91.

Bollen, K. A. (2000). Modeling strategies: In search of the holy grail. *Structural Equation Modeling, 7,* 74–81.

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

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

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

Herting, J. R., & Costner, H. L. (2000). Another perspective on "the proper number of factors" and the appropriate number of steps. *Structural Equation Modeling, 7,* 92–110.

Mulaik, S. A., & Millsap, R. E. (2000) Doing the four-step right. *Structural Equation Modeling, 7,* 36–74.