The Journal of Experimental Education

ISSN: 0022-0973 (Print) 1940-0683 (Online) Journal homepage: www.tandfonline.com/journals/vjxe20

Parsimony and Model Evaluation

Stanley A. Mulaik

To cite this article: Stanley A. Mulaik (1998) Parsimony and Model Evaluation, The Journal of Experimental Education, 66:3, 266-273, DOI: 10.1080/00220979809604411

To link to this article: https://doi.org/10.1080/00220979809604411

The Journal of Experimental Education, 1998, 66(3), 266–273

Parsimony and Model Evaluation

STANLEY A. MULAIK
Georgia Institute of Technology

ABSTRACT. Marsh and Hau (1996) argued that certain models should not be penalized for having low parsimony because an appropriate model for the data may require estimating more parameters. Mulaik argues that Marsh and Hau misunderstand the concept of parsimony, particularly its role in testing a hypothesis about an incompletely specified model to establish its objective validity. More parsimonious models represent more complete hypotheses having more ways of being tested and possibly being disconfirmed. Mulaik also shows that even within the context of the models used in Marsh and Hau's examples, there are much more parsimonious versions of those models that could have been hypothesized and tested, with good fit.

MARSH AND HAU (1996) generated and analyzed some Monte Carlo data to show that it may be inappropriate for researchers to penalize certain models for low parsimony when evaluating the fit of the models. However, Marsh and Hao's argument ultimately rests not on the Monte Carlo results themselves, but on the argument that in some cases on a priori grounds researchers must be allowed to work with less parsimonious models and not be penalized as a result in assessing their fit to the data.

My own reaction to their article is one of disappointment. They are to be applauded for noting a distinction between penalties for parsimony and penalties for estimation, but their treatment of the topic of parsimony demonstrates a near total lack of understanding of the rationale behind the way the concept of parsimony is used in science in general and by Mulaik and colleagues (Carlson & Mulaik, 1993; James, Mulaik, & Brett, 1982; Mulaik, 1991; Mulaik & James, 1995; Mulaik et al., 1989) in particular in connection with its application to evaluating models in structural equation modeling. Their bibliography shows that they were exposed to this rationale. Somehow, they chose not to deal with it. Yet this rationale would have to be addressed in any attempt, such as the one they have made, to argue that a model should not be penalized for low parsimony because it estimates so many parameters.

Let me therefore take this opportunity to briefly outline what parsimony is about and how that relates to parsimony-adjusted goodness-of-fit indices. In the 14th century, William of Occam originally put forth the principle "Entities are not to be multiplied except as may be necessary" (Jones, 1952)—implying that theories should be as simple as possible. Some of the misunderstanding about parsimony arises because originally this principle was regarded as a regulative principle in science standing on its own. Such a view persisted through to this century (Mulaik et al., 1989). L. L. Thurstone, the engineer turned psychologist who popularized the method of factor analysis in the United States, incorporated the concept of parsimony at several places in his methods of factor analysis. Thurstone argued that "the criterion by which a new ideal construct in science is accepted or rejected is the degree to which it facilitates the comprehension of a class of phenomena which can be thought of as examples of a single construct rather than individual events" (Thurstone, 1947, p. 52). He then elaborated on this proposition in a way that presaged the manner in which many use the concept of parsimony in confirmatory factor analysis and structural equation modeling today. He noted that in any situation in which a rational equation is proposed to describe the relationship between two variables, if there are only three observations and there are three independent parameters of the equation to estimate, then the equation has little scientific use because the number of degrees of freedom in the phenomena is the same as the number of degrees of freedom in the equation (i.e., it is a just-identified equation, and fit would be, of mathematical necessity, perfect).

But if one has 100 observations subsumed under an equation that has only three parameters to estimate, then the equation's conformity to the data can be of considerable scientific interest, because there is no mathematical necessity that it should fit the data. What is even more remarkable—to me at least—is that Thurstone anticipated the idea of the parsimony ratio, which I once believed was an original idea when I first proposed it in James et al. (1982). "The convincingness of a hypothesis," Thurstone said, "can be gauged inversely by the ratio of its number of degrees of freedom to that of the phenomena which it has demonstrably covered" (Thurstone, 1947, p. 52). Here, contrary to current usage for the term *degrees of freedom of a model,* he meant *degrees of freedom of the hypothesis,* or the number of parameters that must be estimated, whereas the number of degrees of freedom of the phenomena corresponds to the number of observations to fit the hypothesis to. So, the fewer the number of parameters that must be estimated relative to the number of data points that the equation is designed to account for, the more convincing is the hypothesis when the curve of that equation passes through or near all of those points.

This amounts to the same thing as saying, as I have in my writings with respect to the parsimony ratio, that the more degrees of freedom an acceptably fitting model has (in terms of conditions by which it could have failed to fit the data)

relative to the number of data points to be accounted for by the model, the more convincing is the model as a representation of that data. One begins with as many degrees of freedom by which to fail to fit the data as there are data points and loses a degree of freedom for each parameter estimated. In Mulaik (1990), I showed that the number of potential dimensions of the residual space—the space in which the data are free to differ from the reproduced model—equals the number of independent data points minus the number of parameters estimated, that is, the degrees of freedom of the model. In any case, Thurstone should be studied by the present generation of structural equation modelers for the insights he had into the modeling craft. Even so, Thurstone still regarded parsimony as a regulative principle standing on its own. He offered no reasons why models with fewer estimated parameters relative to the data points subsumed under them are more convincing.

The Austrian philosopher of science Sir Karl Popper, whose greatest impact was on English-speaking scientists, argued that parsimony was not a regulative principle standing on its own, but depended on a more fundamental principle: "The epistemological questions which arise in connection with the concept of simplicity," he said, "can all be answered if we equate this concept with *degree of falsifiability*" (Popper, 1934/1961, p. 140). He seemed to sense that a hypothesis that requires fewer estimated parameters can be subjected to more possibly disconfirming tests than a hypothesis that estimates many parameters. In Mulaik (1990), I showed why this is so in terms of the number of dimensions in which a hypothesized model is free to differ from the data points when some of the parameters of the model are estimated from that same data. In Mulaik and James (1995), we identified parsimony with disconfirmability of a model. Even so, this still does not explain fully why the model with more degrees of freedom should be more convincing when it fits the data.

In Mulaik (1995), I offered an explanation: We test a hypothesis against data independent of the data used in the formulation of the hypothesis because that is the way we judge the objective validity of the hypothesis when it fits the data. Ordinarily, given a set of data, anyone can construct any number of distinct models by trial and error and the multiplication of parameters that reproduce the same data. Such models necessarily fit that data. Thus, fit to that data is no longer a sign of the validity of the model constructed by trial and error or some iterative process of increasing approximation to the data, because numerous models can be made in the same way to fit the same data, not all of which can be true. Furthermore, such models lack a mark of objectivity, that is, their validity is independent of the particular perspectives, viewpoints, and frameworks unique to the individuals formulating the models. But testing a model against data not used in its formulation does provide provisional evidence for or against the objective validity of the model.

In Mulaik (1995), I argued that the basis for the judgment of objective validity for a model that fits data not used in the formulation of the model is derived

from a schema of perception by which we perceive a world of objects independent of ourselves. The psychologist J. J. Gibson (1966, 1982) argued that perception must be understood in terms of the way it is used by an organism as it moves around in and interacts with its environment. Objects visually are invariants through time of features of the optic array of the organism that otherwise varies in systematic ways with the motions and changes in position of the organism as it moves through its environment and thus are independent of the acts of the organism. Object perception for Gibson involved a simultaneous differentiation of what was caused by the object and what was caused by the subject of perception using information in the sensory array. He called the perception of objects *exteroception* and the perception of the effects on the sensory array due to acts of the organism *proprioception.* For example, the organism gains information that it is moving away from or moving toward an object by the way points on the apparent surfaces of the object appear to contract inwardly or to "loom" outwardly on lines radiating from the focal point of vision focused on that object. On the other hand, the object is an invariant pattern that persists through such dilations and contractions. I call this schema for the simultaneous perception of objects and the effects of acts of the organism and their separation the *subject–object schema of perception.* We use this schema, for example, when we see an object from one point of view, develop an interpretation of what we see, and then evaluate that interpretation by moving to a different point to see if that interpretation is consistent with what is seen from the new point of view. This is illustrated nicely by what is known as the Ames chair (Zimbardo, 1985) and the Ames room (Weiten, 1992) illusions. Data from a different perspective and independent of the data used to formulate an interpretation are used to "test" the interpretation and may show it to be "false."

Schemas like the subject–object schema, the path schema, the container schema, the chain schema, the part–whole schema, the up–down schema, and the radial schema are derived from embodied perception, which integrates sensory information received by us either simultaneously or fractions of a second apart. Those schemas also make up much of the framework of our conceptions of the world as metaphors that integrate recalled information received at widely spaced points in time and space (Johnson, 1987; Lakoff, 1987; Lakoff & Johnson, 1980; Mulaik, 1995). Science uses the subject–object schema as a metaphor for establishing objective conceptions of the world (Mulaik, 1995)—hence, the value that science places on demonstrating the invariance of results that are independent of the observer, the observer's means and instruments of observation, and the observer's conceptual framework or point of view and hence, also, the basis for testing hypotheses with data not used in their formulation. Furthermore, the more such tests of a hypothesis from different perspectives and with different measuring instruments are performed and the tests passed, the more confidence scientists have in the provisional objective validity of the hypothesis. Hence, the test of a

hypothesis against data that has more ways of being possibly disconfirmed is valued more highly than a test of a hypothesis that has very few ways of being possibly disconfirmed. Because estimating parameters involves using aspects of the data to complete an incompletely specified hypothesis by finding values that make the model fit those aspects of the data optimally, conditional on the prespecified parameters, those aspects of the data are lost to hypothesis testing, and goodness of fit is relevant only to the prespecified, hypothesized parameters. Consequently, models that estimate relatively many parameters are less desirable than models that estimate only a few.

Applying this general conception of parsimony and objectivity to structural equation models, we see that it favors tests of models about which one has prespecified values for parameters. Unfortunately, incompletely specified models like those described by Marsh and Hau (1996) estimate many parameters and are not examples of parsimonious models—even in the case of the so-called parsimonious model of their example, which has only 25 *df* of a possible total of 45, a degree of freedom having been lost for each estimated parameter. It is essential to understand that estimating a parameter is not the same as specifying a hypothesis about that parameter: Estimating a parameter indicates ignorance about its value and a willingness to let the data dictate a value that would allow the hypothesized model to optimally fit the data, conditional on whatever constraints have been imposed on other parameters in specifying a hypothesis about the model. This procedure permits one to evaluate the hypothesis about specified parameters in terms of the fit of the model to the data, without the estimated parameters introducing sources of lack of fit. Whatever lack of fit arises is due to the prespecified values of the parameters. Thus, even though in some cases the models considered by Marsh and Hau (1996) fit the data very well in both the "parsimonious" and "nonparsimonious" cases—which indicates that the hypothesized, prespecified values are objectively correct or approximately correct (within the context of the basic framework of the models)—the models as a whole would still be penalized from the point of view of parsimony and objectivity in that they test at most 25/45 of the conditions that could have been tested in the parsimonious case and only 16/45 in the nonparsimonious case. In these ratios, the numerator (25 or 16) indicates the degrees of freedom of the model, corresponding to the number of independent conditions that have been put to the test in the model, whereas the denominator (45) indicates the number of independent elements of the observed covariance matrix (for nine variables) against which the model can be tested. These ratios are what I call *parsimony ratios* and are indicative of the quality of the model from the point of view of its disconfirmability and the objectivity of its application if acceptable in fit.

It is important to realize that testing a model as a hypothesis concerns only the prespecified, overidentified conditions in the model and not the whole model in

terms of both specified and estimated parameters. I get the impression that Marsh and Hau (1996) regard goodness of fit as relevant to the whole model and as indicating the appropriateness of certain freed parameters they regard as corresponding to certain (nonzero?) values. I recognize that some structural equation modeling program manuals foster this misconception by recommending to neophytes that unknown but presumed nonzero values be represented by free parameters, but the computer programs themselves have no hypothesized constraints on the freed parameter values in seeking by iteration values that optimally allow the model to fit the data conditional on the prespecified parameters. Again, I stress that freeing a parameter is not the same as specifying a value or even a range of values for the parameter. The estimated value could turn out to be anything. It is not a part of one's hypothesis. It is not what is tested in the model. So why should it receive any consideration when one is evaluating the model from the point of view of what has been prespecified and hypothesized about it? The estimated value may be useful for future hypotheses to test with other data, but it is not relevant to the hypothesis being tested with the data at hand.

Marsh and Hau's (1996) concern with the parsimony ratio was that when it is multiplied by a goodness-of-fit-index value, a "parsimony adjusted goodness-of-fit index" is obtained (James et al., 1982) that shrinks the goodness-of-fit value considerably in the case of models with many estimated parameters relative to the number of independent parameters in the covariance matrix of the observed variables. They related this index to a decision rule of how one might choose between models: Accept the model with the best parsimony-adjusted goodness-of-fit index. Even though I can see merit in the way the parsimony-adjusted goodness-of-fit index rewards scientists who risk more by prespecifying more and estimating fewer parameters and being approximately right, this is not my rule for accepting models. No consideration is given to whether the models compared already have acceptable fit. I do not regard the fit of the parsimonious model when applied to the data generated with covariances among disturbances of .1 or .2, respectively, acceptable, because I usually require a relative noncentrality index (RNI) of .95 or better to regard the specified parts of a model as acceptable. The RNI of .90 is, I concede, borderline to being "in the .90s" in terms of the rule stated in Carlson and Mulaik (1993) for making these comparisons, but the RNI of .72 is clearly unacceptable. So I would not ordinarily proceed to choose between models in terms of parsimony when they do not have acceptable fit.

The rule cited by Marsh and Hau (1996) also does not consider the magnitude of the parsimony-adjusted goodness-of-fit index. That is more important than relative magnitude. Both their models have low parsimony, and I would have little confidence in accepting their *whole* models in either case as objective truth because they have tested so little about them (Mulaik et al., 1989). For me, the parsimony-adjusted index should be at least in the mid-80s to begin to give me a

high degree of confidence in the model. I have been able to achieve such levels with several data sets from the behavioral sciences (Carlson & Mulaik, 1993; Mulaik, 1988). This can be accomplished regularly if researchers carry parameter values estimated in previous studies or experimental conditions into later studies or conditions. Higher parsimony-adjusted indices can also be accomplished simply by constraining parameters to be equal.

For example, if Marsh and Hau (1996) could rationally argue that all freed variances of disturbances are equal, they could constrain them to be equal and gain 8 *df*. If they could argue that covariances between disturbances of corresponding variables across the three waves are equal, they could gain 6 *df*. Even more, they might argue that all of the freed covariances between disturbances are equal and constrain them to be equal, gaining 8 instead of 6 *df*. Similarly, if the same respective indicators are used at each wave of measurement, the factor loadings on those indicators can be constrained to be equal across waves of measurement to gain 6 more *df*. Moreover, one could constrain all factor pattern coefficients to be equal to gain 8 *df*. Finally, it may be reasonable to expect the path coefficients between the latent variables between the first and second, and second and third waves to be equal, gaining another degree of freedom. Thus, it would be possible to formulate their model in a way that has only four parameter values to estimate, gaining a parsimony ratio of (41/45) = .91. (The null model is not nested in these models, and the degrees of freedom of this last model can exceed the degrees of freedom of the null model. Incidentally, this raises problems for the family of normed fit indices in cases like this.) So Marsh and Hau are overly pessimistic in doubting that such rigor can be typically applied in social science research.

As indicated in Carlson and Mulaik (1993), the main function of the parsimony adjustment is to make researchers who naively believe they are testing a whole model aware that they have not tested all aspects of their models when their goodness-of-fit indices shrink considerably with the adjustment. Such tests may await future studies, and it is not unusual to expect low parsimony-adjusted goodness-of-fit indices early in a research program, when one has little theory or experience on which to base specifications of parameter values. Now, I would not discourage Marsh and Hau (1996) from pursuing their models further, but I would want them to realize that the low parsimony-adjusted indices obtained from high goodness-of-fit index values indicate that they have learned relatively little that is objective about their models because they have put so little into the hypotheses about them.

NOTE

Address correspondence to Stanley A. Mulaik, School of Psychology, Georgia Institute of Technology, Atlanta, GA 30332. E-mail: pscccsm@prism.gatech.edu.

REFERENCES

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

Gibson, J. J. (1966). *The senses considered as perceptual systems.* London: George Allen & Unwin.

Gibson, J. J. (1982). *Reasons for realism: Selected essays of James J. Gibson.* E. Reed & R. Jones (Eds.). Hillsdale, NJ: Erlbaum.

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

Johnson, M. (1987). *The body in the mind.* Chicago: University of Chicago Press.

Jones, W. T. (1952). *A history of Western philosophy.* New York: Harcourt, Brace.

Lakoff, G. (1987). *Women, fire, and dangerous things.* Chicago: University of Chicago Press.

Lakoff, G., & Johnson, M. (1980), *Metaphors we live by.* Chicago: University of Chicago Press.

Marsh, H. W., & Hau, K.-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *The Journal of Experimental Education, 64,* 364–390.

Mulaik, S. A. (1988). Confirmatory factor analysis. In R. B. Cattell & J. R. Nesselroade (Eds.), *Handbook of multivariate experimental psychology* (pp. 259–288). New York: Plenum.

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

Mulaik, S. A. (1991, October). *Clarifying misconceptions about parsimony adjustments of goodness of fit indices.* Paper presented at the annual meeting of the Society of Multivariate Experimental Psychology, Albuquerque, NM.

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

Mulaik, S. A., & James, L. R. (1995). Objectivity and reasoning in science and structural equations modelling. In R. H. Hoyle (Ed.), *Structural equation modeling: Issues and applications* (pp. 118–137). Beverly Hills, CA: Sage.

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

Popper, K. R. (1961). *The logic of scientific discovery* (translated and revised by the author). New York: Science Editions. (Original work published 1934)

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

Weiten, W. (1992). *Psychology.* Pacific Grove, CA: Brooks/Cole.

Zimbardo, P. G. (1985). *Psychology and l ife.* Glenview, IL: Scott, Foresman.