**Title:** *IS PARSIMONY ALWAYS DESIRABLE? IDENTIFYING THE CORRECT MODEL FOR A LONGITUDINAL PANEL DATA SET* , By: Sivo, Stephen, Wilson, Victor L., Journal of Experimental Education, 0022-0973, March 1, 1998, Vol. 66, Issue 3

# IS PARSIMONY ALWAYS DESIRABLE? IDENTIFYING THE CORRECT MODEL FOR A LONGITUDINAL PANEL DATA SET

ABSTRACT. Marsh and Hau (1996) based the assertion that parsimony is not always desirable when assessing model fit on a particular counterexample drawn from Marsh's previous research. This counterexample is neither general nor valid enough to support such a thesis. More specifically, the counterexample signals an oversight of extant, stochastic models justifying correlated uniquenesses, namely, moving-average and autoregressive moving-average models. Such models provide theoretically plausible motives for a priori specification of error correlations. In fact, when uniquenesses are correlated, stochastic models other than the conventional simplex and quasi-simplex models must be tested before positive identification of the process is possible (Sivo, 1997). In short, exchanging the mechanistic penalties for model complexity for the mechanistic specification of untenable measurement-error covariances offers no solution. Parsimony has not been dismissed based on the argument Marsh and Hau presented concerning longitudinal data.

MARSH AND HAU (1996) POSED THE QUESTION "Is parsimony always desirable when assessing the fit of a model to some data set?" in order to argue subsequently that indeed, under some conditions, model parsimony is either too stiff a requirement for or irrelevant to the assessment of goodness of fit. To overturn the assumption that parsimony is always a critical standard in the evaluation of model fit, Marsh and Hau used a particular counterexample. If that counterexample is sufficiently general and valid, model parsimony may be judged, at least in some circumstances, as an inappropriate standard in the evaluation of model fit. Marsh and Hau were similarly careful to qualify their assertion concerning parsimony by acknowledging that choosing the most parsimonious model is "usually good advice." They were concerned, instead, about the universal and "mechanistic" application of the parsimony standard in all research situations, particularly as it is evaluated by indices designed to be sensitive to model complexity.

Marsh and Hau's (1996) counterexample was not among the many examples devised to question the standard of parsimony. Instead, it arose in practice as a stumbling block to interpreting a particular analysis, traceable to some of Marsh's previously published findings (e.g., Marsh, 1993; Marsh & Grayson, 1994a, 1994b).

Review of these articles revealed that before examining certain longitudinal data sets, Marsh hypothesized that measurement errors would correlate high enough to warrant including error covariances in the quasi-simplex model. In fact, Marsh (1993) and Marsh and Grayson (1994a, 1994b) correctly concluded that not specifying the correlated measurement errors in such cases would result in systematically biased estimates of the stability coefficients. Thus, specification of

measurement error correlations in the quasi-simplex model was thought to be more than simply warranted--it was thought to be necessary. However, consideration of such model specification ushered in two issues, one practical and the other theoretical.

The practical issue concerned how to go about specifying all uniqueness covariances. Such specification would underidentify the model. To achieve an identified model, Marsh and Hau had to posit a second-order equivalent to the single-indicator model and allow all uniquenesses to covary. The second-order equivalent, or so-called multiple-indicator model, treats each item as an indicator instead of using the total test score as a single indicator. Although not the subject of analysis in the present article, this solution presents a new problem. For the solution to be applied, every item must serve as an indicator in the quasi-simplex model, and so model complexity increases arithmetically as the number of items comprising a test increases. Furthermore, a five-item test already prone to low or "negative" reliability estimates because of its brevity would be turned into five different indicators for each occasion. If a five-item test is prone to low or irregular reliability estimates, the single-item reliabilities will be much lower, per Spearman-Brown.

The theoretical issue with which Marsh needed to contend concerned the development of a rationale for specifying, a priori, measurement-error correlations in the quasi-simplex model. Marsh's (1993) approach, in this case, was to make the assumption that measurement-error correlations are always high enough in longitudinal data sets to justify their specification in quasi-simplex models. This conclusion motivated and required Marsh and Hau (1996) to assert that parsimony is not always relevant or necessary. They argued that a priori specification was warranted because Joreskog (1979), among other researchers, found correlated uniquenesses in his longitudinal data sets; thus, it follows, the principle of parsimony is not always desirable after all.

In the present article, we did not intend to defend the utility or worth of the principle of parsimony in all cases or the parsimony correction built into several fit indices, but instead to criticize Marsh and Hau's (1996) "counterexample," particularly as it is used to defend a priori specification of measurement-error correlations regardless of their theoretical pertinence to a given model under study. This critique, by implication, challenges Marsh and Hau's argument concerning parsimony as a standard for evaluating models.

## Time-Series Methodology: A Tenable Alternative

Marsh and Hau argued for the a priori consideration of correlated measurement errors in their second-order quasi-simplex model, independent of whatever substantive issue is under investigation in a particular study. In this case, the correlated uniquenesses are unsubstantiated except that several researchers have reported often finding correlated measurement errors in their longitudinal data (e.g., Joreskog, 1979, 1981; Joreskog & Sorbom 1977, 1989; Marsh, 1993; Rogosa, 1979). A priori consideration of correlated measurement errors without regard to the theory proper to the variable relationships under study in a particular research situation is troublesome, given that Glass, Willson, and Gottman (1975) and Sivo

(1997) demonstrated that many longitudinal data sets do not evidence correlated uniquenesses. To better understand how correlated measurement errors may be substantively tenable in particular research situations, a review of the taxonomy offered by time series theorists may be helpful. Similarities between the longitudinal panel data and time-series data types have been duly noted in prior research, and the theory of discrete time-series accounts for how errors can correlate over time. Both support the need for investigating error covariance for each situation.

Box and Jenkins (1976) proposed a method for forecasting time series values for data characterized as dependent or autocorrelated. Autocorrelation in a series is generated by probabilistic models using stochastic processes; by examining the features of the autocorrelation at hand, one may identify the stochastic process in question (Box & Jenkins). Box and Jenkins explained that stochastic processes are statistical phenomena developing over time congruent with probabilistic laws.

According to Box and Jenkins (1976), time series data often may be modeled for two distinct stochastic processes: autoregressive (AR) and moving average (MA). AR models are constructed to allow the current realization of a time series to be expressed as a function of previous realizations of the same time series:

$$X_t = Phi_1X_{t-1} + Phi_2X_{t-2} + ... + Phi_pX_{t-p} + Epsilon_t,$$

where X denotes an observed score taken on some occasion (t) deviated from the original level $X_0$ of the series; Epsilon denotes an independent, identically distributed error associated with a given occasion (t); and Phi denotes a correlation among temporally ordered scores at some lag (e.g., t - 1 or a lag of 1, t-2 or a lag of 2). MA models are constructed to allow the current value of a time series to be expressed as a function of autocorrelated errors:

$$X_t = Epsilon_t - Theta_1Epsilon_{t-1} - Theta_2Epsilon_{t-2} - ... - Theta_qEpsilon_{t-q},$$

where X denotes an observed score taken on some occasion (t) deviated from the original level $X_0$ of the series; Epsilon denotes independent, identically distributed error associated with a given occasion (t); and Phi denotes a correlation among errors at some lag (e.g., t - 1 or a lag of 1, t - 2 or a lag of 2). When both processes are concurrently responsible for the dependency in a series, the autoregressive moving average (ARMA) model is defined as

$$X_t = Phi_1X_{t-1} + Phi_2X_{t-2} + ... + Phi_pX_{t-p} - Theta_1Epsilon_{t-1} - Theta_2Epsilon_{t-2} - ... - Theta_qEpsilon_{t-q} + Epsilon_t.$$

Commenting on Marsh's models in a article concerning how to use structural equation modeling (SEM) to fit time series models to longitudinal data, Willson (1995) noted that the basic autoregressive model is but a restricted form of Marsh's (1993) depiction of the simplex model. In fact, many researchers have mistakenly referred to the simplex model as a first-order AR, suggesting the association with time series models (e.g., Joreskog & Sorbom, 1989; Rudinger, Andres, & Rietz, 1991). The difference between the two models is that the simplex model allows all true scores and errors to be independently estimated, whereas the autoregressive

model constrains the true scores and errors in the series in specific ways.

Willson (1995) further indicated that Marsh (1993) did not test a model that assumes an MA process, wherein the errors rather than the observed scores propogate, although Willson suggested that the errors may instead possess an MA structure. If autocorrelation were anticipated to be present in both the true scores and the errors, an ARMA model may be fit to the data. Sivo (1997) found that although measurement errors are not correlated in all longitudinal data sets, when in fact such measurement error correlations are found they often exist only at a particular lag nearest to the diagonal of the error covariance matrix, suggesting that an MA or ARMA process exists in the data.

## Modeling Correlated Measurement Errors in Longitudinal Data

Clearly, the practice of testing multivariate time series models accounting for correlated errors against longitudinal data is rare when compared with the practice of testing simplex or quasi-simplex models. Indeed, the predominant discussion of fitting simplex or quasi-simplex models to longitudinal data in the literature motivated Marsh (1993) to suggest the one-factor model. Perhaps the reason time series models specifying correlated errors have not been used much is that "econometricians have concentrated on the linear simultaneous equation model, in which there are stochastic disturbances in equations ['shocks'] but not on measurement errors in variables ['errors']. Meanwhile, other social scientists--e.g., psychometricians and statistical sociologists--have focused their attention on errors-in-variables models in the form of true score theory and factor analysis" (Geraci, 1977; p. 163). Presumably, "other social scientists," particularly psychometricians, have not focused much attention on stochastic disturbances. Also, researchers in time series analysis have typically recommended a large number of data points, often cited as at least 50, to use ARMA models. Unfortunately, the distinction between requiring such numbers for forecasting and complex model identification and requiring only a few time points for simple AR, MA, or ARMA models has not been forcefully pointed out.

Just as it is legitimate to use SEM to observe whether stochastic simplex models (or their more restricted autoregressive form) fit longitudinal panel data, it is equally plausible to observe whether stochastic MA or ARMA models fit as well, especially when measurement errors covary. Moreover, this approach requires that the researcher theoretically decide, a priori, what relationships are substantively tenable. Models must be specified at some hypothesized lag according to what is theoretically plausible, or lag correlation structure can be examined. Because time series analysis has been generally used with single -subject designs, longitudinal panel data have both some advantages and some limitations with time series methods.

## Longitudinal Panel Data and Time Series Models

With single-subject designs, the autocorrelation structure is estimated from the observed number of time points. Because this is a straightforward power function, methodologists suggest a minimum number of time points, such as 50, to support

reasonable standard errors. For panel data, the individual subjects may be thought to be replications. Cross-time covariances will then be more stable than covariances for single subjects, and far fewer time points will be needed to estimate lagged relationships. Of course, this requires assuming identical time series processes for all individuals or reasonably similar processes that can be aggregated meaningfully into a single representation. It is an empirical question whether an identified process reasonably represents all, most, some, or none of the subjects comprising the panels. It is also an empirical question whether it is important to require similarity of process for the group and the individual. If the same process is observed across groups and across time, the estimation of parameters is appropriately made, regardless of the processes found in individuals. The substantive issue of why the processes differ will be a different investigation.

With longitudinal panel data, it seems reasonable that a minimum of four time points are needed to specify either an AR Lag 1 or an MA Lag 1 process. Although more complex processes may be at play, potentially affecting estimation and statistical tests, the limitation is in the generalizability of the data, not in the estimation and testing within the sample itself. Models more complex than those we have been able to measure and estimate are always possible. For as few as five or six time points, more complex error-covariance structures such as AR Lag 2 or ARMA (2,2) are possible and are likely to exhaust most commonly found models (Glass et al., 1975).

Unlike general simplex models, the time series formulations restrict the adjacent covariances (for Lag 1) to the same values. Such parsimony is relevant to examining error-covariance structure, because such a restriction can free a significant number of degrees of freedom.

Marsh (1993) was correct when he said that researchers who analyze longitudinal data should not rely solely on the simplex or quasi-simplex model; however, stochastic models other than the conventionally used simplex and quasi-simplex models should also be tested against longitudinal data sets, particularly when the estimation of measurement-error covariances is substantively relevant. In this case, second-order models are unnecessary because measurement-error covariances are expected only at some lag, and so the time series models specified to have measurement-error covariances are sufficiently parsimonious to be estimated.

## Conclusion

The weight of Marsh and Hau's (1996) argument rests on their proposed counterexample. This counterexample was not one of many possible counterexamples, but represents a stumbling block evident in some of Marsh's previously published research (e.g., Marsh, 1993; Marsh & Grayson, 1994a, 1994b). It is a stumbling block because Marsh (1993) was compelled to specify measurement-error correlations in his quasi-simplex or one-factor models (or else risk biased stability estimates), though he had no theoretical justification for this mechanistic practice. Consequently, Marsh and Hau (1996) were apparently forced to argue both that parsimony should not always be a standard and that the a priori assumption of correlated measurement errors should be made when longitudinal

data are studied, despite their potential irrelevance to the particular theory advanced. However, Sivo (1997) found that not all the widely used exemplar longitudinal data sets have correlated errors, so such an a priori claim is presumptuous. Moreover, Sivo found that whenever simplex or quasi-simplex models are tested, both MA and ARMA models must also be tested before one decides which process is present in a given data set because, on gross observation, one stochastic model may appear to capture the error process in a data set when, in reality, another process is present.

Even if all error covariances were found to be high (suggesting no stochastic process), a more orthodox and tenable approach to the data would be to determine first whether some unmodeled variable, such as practice effects, is present in the data. In this case, the correct approach would be to identify and consequently model the variable explicitly, rather than simply freeing all measurement-error covariances.

In summation, exchanging the mechanistic decision rules that penalize model complexity with the mechanistic decision rule calling for the unwarranted specification of all measurement-error covariances does not resolve the issue of how model evaluation should be effectively conducted. Certainly, parsimony has not been dismissed based on the argument Marsh and Hau (1996) presented for longitudinal data.

## REFERENCES


Box, G. E. P., & Jenkins, G. M. (1976). Time series analysis: Forecasting and control (Rev. ed.). Oakland, CA: Holden-Day.

Geraci, V. J. (1977). Identification of simultaneous equation models with measurement error. In D. J. Aigner & A. S. Goldberger (Eds.), Latent variables in socio-economic variables (pp. 163-186). New York: North-Holland.

Glass, G. V, Willson, V. L., & Gottman, J. M. (1975). Design and analysis of time-series experiments. Boulder, CO: Colorado Associated University Press.

Joreskog, K. G. (1979). Statistical estimation of structural models in longitudinal-developmental investigations. In J. R. Nesselroade & P. B. Baltes (Eds.), Longitudinal research in the study of behavior and development (pp. 303-352). New York: Academic Press.

Joreskog, K. G. (1981). Statistical models for longitudinal studies. In F. Schulsinger, S. A. Mednick, & J. Knop (Eds.), Longitudinal research: Methods and uses in behavioral science (pp. 118-124). Hingham, MA: Nijhoff.

Joreskog, K. G., & Sorbom, D. (1977). Statistical models and methods for analysis of longitudinal data. In D. J. Aigner & A. S. Goldberger (Eds.), Latent variables in socio-economic variables (pp. 187-204). New York: North-Holland.

Joreskog, K. G., & Sorbom, D. (1989). LISREL 7: A guide to the program and

applications [Software manual]. Chicago: SPSS.

Marsh, H. W. (1993). Stability of individual differences in multiwave panel studies: Comparison of simplex models and one-factor models. Journal of Educational Measurement, 30(2), 157-183.

Marsh, H. W., & Grayson D. (1994a). Longitudinal confirmatory factor analysis: Common, time-specific, item-specific, and residual-error components of variance. Structural Equation Modeling, 1(2), 116-145.

Marsh, H. W., & Grayson D. (1994b). Longitudinal stability of latent means and individual differences: A unified approach. Structural Equation Modeling, 1(4), 317-359.

Marsh, H. W., & Hau, K.-T. (1996). Assessing goodness of fit: Is parsimony always desirable? The Journal of Experimental Education, 64(4), 364-390.

Rogosa, D. (1979). Causal models in longitudinal research: Rationale, formulation, and interpretation. In J. R. Nesselroade & P. B. Baltes (Eds.), Longitudinal research in the study of behavior and development (pp. 263-302). New York: Academic Press.

Rudinger, G., Andres, J., & Rietz, C. (1991). Structural equation models for studying intellectual development. In D. Magnussen, L. R. Bergman, G. Rudinger, & B. Torestad (Eds.), Problems and methods in longitudinal research: Stability and change (pp. 274-307). New York: Cambridge University Press.

Sivo, S. A. (1997). Modelling causal error structures in longitudinal data. Doctoral dissertation, Texas A&M University, College Station, TX.

Willson, V. (1995,July). A comparison of time series and structural equation models in longitudinal multivariate data. Paper presented at the annual European meeting of the Psychometric Society, Leiden University, the Netherlands.

~~~~~~~~

By STEPHEN A. SIVO, James Madison University and VICTOR L. WILLSON, Texas A&M University

---


**Source:** Journal of Experimental Education, Spring98, Vol. 66 Issue 3, p249, 7p.
**Item Number:** 545425
~~~~~~~~