The Reliability of Survey Measures
RESULTS Series

TYPES OF CORRELATIONS, MODELS, AND METHODS FOR HANDLING ATTRITION AND MISSING DATA IN RELIABILITY ESTIMATION

Paula A. Tufiș

University of Bucharest

Duane F. Alwin

Pennsylvania State University and the University of Michigan

JUNE 2023

Suggested citation: Tufiș, P.A. & Alwin, D.F. (2023). Types of Correlations, Models, and Methods for Handling Attrition and Missing Data in Reliability Estimation. *The Reliability of Survey Measures Results Series.*

The[1] present research proposes a strategy in which the reliability of measurement is analyzed with respect to the factors that contribute to high vs. low levels of precision in measurement. The approach entails the estimation of reliability for a large number of survey questions using longitudinal or panel designs, creating a dataset yielding estimates of reliability along with variables representing attributes of questions, e.g. questions about facts versus questions about subjective states, or questions in batteries vs. questions in series, or questions employing differing numbers of response categories.

As discussed in introductory and related documents (see Alwin, 2007), if one has three waves of panel data, separated in time, and if one can meet certain assumptions about the independence of errors, the 3-wave *quasi-Markov simplex model* (QMSM) can be employed to estimate measurement reliability. Under appropriate design conditions, results obtained from parameter estimation can be interpreted in terms of the reliability of measurement. When applied to panel data that meet the relevant design requirements, the results can be used to evaluate the relative quality of the data within the framework of classical reliability theory.

The model employed here may not always work, in the sense that it may not fit the data well, but over most trials, it has; and we believe our results can tell us something about the *relative* quality of representative measures of particular questions. To make use of this model, there are several issues that need to be considered before estimates can be obtained. In the following, we briefly cover three preliminary considerations, discussing what we learned from this methodology from the past 30+ years of research using these methods. Specifically, here we focus on three issues: (1) what correlations we should use for assessing relationships of multiple measurements

[1] Results described in this document are based on the six panel studies in the *Margins of Error* (2007) project, supplemented by more recent research using the Health and Retirement Study and the General Social Surveys. See Alwin (2007, pp. 130-147) and Alwin, Baumgartner and Beattie (2017).

over time using longitudinal designs? (2) what assumptions must be made to identify the 3-wave *quasi-Markov simplex model*, for example, the assumptions of equal reliabilities over time, or equal error variances, or neither? and (3) how to handle missing data due to attrition and other sources of missingness? We turn, first, to the types of correlations that should be analyzed, given the nature of the question and its response categories.

Type of Correlations

The estimation of reliability using the 3-wave *quasi-simplex model* is based on correlational data, that is, the correlations among a given variable measured at three separate waves. A basic question is, then, what correlations should one use? The original models written by Heise (1969) and Wiley and Wiley (1970) assumed continuous variables, and the model was applied to simple Pearson correlations, as well as related covariances in the case of the Wiley and Wiley (1970) estimates. Later expositions made convincing arguments that when the variables are not continuous, but are ordinal in nature (e.g., having 10 or fewer categories), it is more appropriate to use polychoric correlations, and in the case of true dichotomies, tetrachoric correlations (Muthén, 1984; Jöreskog, 1990, 1994; see also Alwin, 2007, pages 127-135). The latter estimates the correlation for a true underlying variable that is continuous.

In Table 1 we present estimates of reliability using several different methods, allowing comparisons of Pearson vs. polychoric correlational approaches, as well as comparisons of differing approaches to model assumptions. To summarize the estimates: column (1) presents estimates based on Heise's approach applied to Pearson correlations and based on a listwise data sample; column (2) contains estimates based on Pearson correlations using the Wiley assumptions of equal error variances based on a listwise sample; column (3) contains estimates using the Heise approach applied to polychoric correlations and based on a listwise data sample; column (4) contains estimates based on polychoric correlations using the Wiley assumptions of equal error variances based on a listwise sample approach. We return below to the comparison of the two different models of the error structure. [2]

The first issue we address in Table 1 is the question of the type of correlation to use in the calculation of relationships between cross-time measures. For instance, when should one employ simple Pearson correlations among cross-time measures? Or should one instead consider tetrachoric correlations for dichotomous variables as is recommended in the statistical literature, and should one more profitably employ polychoric correlations for ordinal data more generally. Table 1 displays reliability estimates from several different approaches to estimating the reliability of individual survey items. This table is an extension of an earlier table (Alwin, 2007, p. 131) and illustrates some of the similarities and differences in results across different estimates, as well as by content areas. Note the numbers in Table 1 are presented for the total set of questions, as well as broken down by question content, operationalized here according to Alwin's (2007, pages 153-154) differentiation of facts (content that can be verified), vs. non-facts, which are largely subjective states), as well as differences among types of non-factual content, specifically, beliefs (statements about what is), attitudes (positive and negative sentiments toward a social object, values (statements about what should be), self-perceptions (beliefs about the self), self-assessments (evaluations of the self) and expectations (beliefs about future events or situations). With respect to content, the patterns across all methods indicate that facts can be more reliably measured than non-facts. We take up this issue in a separate document but suffice it to say at this point that content

[2] Note the results in Tables 1 and 2 pertain only to the six studies in the *Margins of Error* (2007) project. Similar results from the other studies replicate these findings.

is a major source of differences in our estimates of reliability for survey questions, which is therefore controlled by partitioning the sample of questions in Table 1.

Table 1. Estimates of reliability for survey measures by type of content, type of correlation, and type of model

	Reliability Estimates			
Content	Pearson/ Heise[1]	Pearson/ Wiley-Wiley[2]	Polychoric/ Heise[3]	Polychoric/ Wiley-Wiley[4]
Facts	0.7662	0.7678	0.7556	0.7580
	(72)	(72)	(37)	(37)
Beliefs	0.4792	0.4922	0.5854	0.5996
	(116)	(116)	(114)	(114)
Values	0.5547	0.5516	0.6614	0.6635
	(43)	(43)	(42)	(41)
Attitudes	0.6034	0.5942	0.6651	0.6662
	(76)	(76)	(74)	(72)
Self-Assessments	0.4966	0.5018	0.6359	0.6368
	(25)	(25)	(25)	(25)
Self-Perceptions	0.5049	0.5064	0.6259	0.6301
	(87)	(87)	(86)	(83)
Proxy Facts	0.8083	0.8200	1.0000	1.0000
	(7)	(7)	(1)	(1)
	0.5692	0.5718	0.6396	0.6456
Total	(426)	(426)	(379)	(373)

Note: number of measures in parentheses.
[1] Estimates based on Pearson correlations and assumption of equal reliabilities.
[2] Estimates based on Pearson-based covariance matrix and the assumption of equal error variances. The average of the three estimates is presented.
[3] Estimates based on polychoric correlations with thresholds constrained equal and the assumption of equal reliabilities. Excludes items with 16+ response categories.
[4] Estimates based on the polychoric-based asymptotic covariance matrix with thresholds constrained equal and the assumption of equal error variances. Excludes items with 16+ response categories and feeling thermometers. The average of the three estimates is presented.
Source: adapted from Alwin (2007).

The numbers in Table 1 – comparing columns 1 and 3 and comparing columns 2 with 4 – strongly support the strategy of using polychoric correlations for ordinal variables and Pearson correlations for interval variables. Reliabilities are consistently higher using polychoric correlations, especially for non-factual content, when the polychoric approach is used for ordinal variables. Based on extensive analysis of this issue, we recommend using a "hybrid" approach (see Alwin, 2007, page 147). We concluded that for continuous variables one should estimate reliability based on Pearson correlations, but if the variables are no more than ordinal, reliability estimation should be based on polychoric correlations. The "hybrid" estimates can then be combined for purposes of meta-analysis of reliability across questions.

Assumptions about Error Variance

The only assumption necessary to estimate the reliability of measurement using these 3-wave *quasi-simplex models* is the independence of errors. Design requirements for interpreting reliability estimates from these assumptions are stringent (see (Alwin, 2021). The Heise (1969) model computes reliability using the simple formula, reliability = COR(21)*COR(32) / COR(31). This is wave-2 reliability, and it is identified regardless of any other assumptions (see Alwin, 2007, pp. 109-110 for a discussion of the identification issue). If one wishes to identify all the parameters of this model, particularly the stability coefficients, then some further assumptions are needed to identify the model. But, it should be emphasized that the Heise estimate is simply the wave-2 reliability. It is completely identified given the assumption of measurement independence. This estimate can then be used to further identify the other parameters of the model, making assumptions about the error structures over waves.

There are two basic approaches to modeling the error structures using these 3-wave *quasi-simplex models*. One is the approach of Heise (1969), which simply assumes that reliability of measurement is a constant over waves of the 3-wave panel. This *equal reliabilities* approach requires no more than the correlational data referred to earlier, whether based on Pearson correlations or polychoric correlations. In both cases, the Heise model simply computes reliability using the simple formula, reliability = COR(21)*COR(32) / COR(31). Obviously, this model assumes a *simplex structure* to the data (hence the name "simplex model"), which means that the correlation COR(31) will be smaller than the correlations COR(21) and COR(32). If that assumption does not hold, this is the wrong model for the data, and one must resign oneself to the fact that the process being modeled is more complicated than this model assumes. Such results are rare, but when they occur, they usually suggest that there is something more complicated going on and one cannot make the assumption of "dynamic equilibrium" (see Alwin, 2007).

Using these methods, it is possible to estimate "Heise reliabilities" as given above. Note also that we currently use the M-*plus* definition of ordinal variables, that is, those where the number of response categories is 10 or less, although we have used other approaches in the past (see Alwin, 2007). For variables that are considered to be continuous, that is, those with response categories greater than 10, it is possible to entertain more than the "Heise reliabilities" and proceed further to estimate a separate reliability for each wave, based on the Wiley-Wiley approach. Again, this can be done once for the listwise sample and again for the FIML sample, as follows:

Wiley estimate (1) = [Var(1) – Var (e)] / Var(1)
Wiley estimate (2) = [Var(2) – Var (e)] / Var(2)
Wiley estimate (3) = [Var(3) – Var (e)] / Var(3)

where Var (e) = Var(2) – [HeiseReliability * Var(2)].

Table 1 provides two comparisons between the Heise and Wiley-Wiley models. One involves a comparison of the first two columns of the table (labelled Pearson/Heise and Pearson/Wiley-Wiley). There are virtually no differences between the two sets of results, and one would be hard-pressed to argue for one over the other—they are essentially the same model. A second comparison involves the case of ordinal measures, in contrast to continuous measures. It is more difficult to obtain a covariance matrix among ordinal variables (Jöreskog, 1990, 1994). Some approaches have been taken to obtaining an asymptotic covariance matrix for ordinal variables, but these are not universally accepted approaches. We have applied them in the present project, and they do not produce substantially different estimates than the correlational approaches. The comparison of columns 3 and 4 (labelled Polychoric/Heise and Polychoric/Wiley-Wiley) in Table 1 reveals few differences between the two sets of results.

If one is satisfied with the Heise model, then one can proceed with the results and easily analyze differences among survey questions in their levels of reliability. Still, there were some serious issues raised in the paper by Wiley and Wiley (1970), in which they clarified the fact that Heise's *equal reliabilities* assumption may be sufficient to identify the 3-wave model, but it was not a necessary set of constraints. They showed that the assumption of *equal error variances* was an alternative, less restrictive model, and using a covariance matrix, rather than a correlation matrix, one could obtain estimates that would permit a different interpretation of reliability at each wave of the panel. While true, there is considerable debate about whether this is a desirable alternative, especially given the possibility that the measurement properties of a questionnaire may vary over time. In such cases, one may reasonably question whether the simplex model is the correct model.

In general, we have found that these separate wave-specific reliability estimates are not very different, and can easily be disregarded, but this set of operations can provide additional insight into whether the Heise (1969) model is appropriate for the data. Examples can be provided, e.g. see estimates for the reliability of reports of income in Alwin, Zeiser and Gensimore (2014), in which wave-specific reliabilities are presented. In other words, it appears there is really little to be gained in obtaining Wiley-Wiley-type estimates for ordinal data, so in practice we rely solely on Heise estimates in such cases. On the other hand, when possible, we routinely obtain Wiley-

Wiley wave-specific estimates of reliability in the case of continuous variables, but these estimates are generally no better than the Heise estimates. Recall that the middle wave reliabilities in the Wiley-Wiley approach are identical to the Heise estimates (see Table 1 above), which in part explains the convergence of the two sets of results. When covariance information exists, as in the case of continuous variables, such as age, or years of schooling, or income, it is easily possible to estimate wave-specific reliabilities using the Wiley-Wiley approach, and even though there are very few differences between the Heise and Wiley-Wiley approach, we recommend that one should estimate both models because to do so can be informative. Of course, as previously noted, the reliability of wave-2 of the Wiley-Wiley model is always going to be equal to the Heise reliability estimate, so the question revolves around whether or not the wave-1 and wave-3 reliabilities are appreciably different. Such occurrences are typically rare. Nevertheless, we recommend computing these separate reliabilities when possible, in order to assess one's comfort level with the Heise approach. In fact, although no one ever does, the assumption is testable, but one needs more than three waves of data, or a multiple group approach (see Alwin, 2007). Cernat et al. (2021) provide an example using several waves of the British Household Panel Study.

Handling Attrition and Missing Data

In addition to the above considerations, a final issue that arises in the use of 3-wave *quasi-simplex models* to estimate the reliability of survey measures is how one must deal with the issue of missing data, given the often-high levels of attrition in longitudinal studies. In Table 2 we present a comparison of three approaches to the problem of missing data. First, is a simple listwise approach, including information for cases only if they have complete data at all three waves. This approach—used almost exclusively in the monograph mentioned earlier (see Alwin, 2007) was "listwise" data present, that is, using only those cases that had data present in all three waves of the panel. A second approach is Allison's (1987) multiple-group pattern-mixture in which the problem of incomplete data due to attrition is dealt with by specifying a model in which multiple-group sub-models are formulated for different patterns of incomplete data. A third approach is to apply the Full-Information Maximum-Likelihood (FMIL) approach advocated by Wothke (2000) or in the case of ordinal variables, using the weighted least squares mean- and variance-adjusted (WLSMV) estimation (Asparouhov and Muthén, 2010) to handle missing data due to attrition and other causes (see Alwin, 2007, pages 137-146). This approach is statistically justified but can be misleading when there is not much data present across waves of the survey. Therefore, before using such an approach to estimate reliability of measurement, it is important to assess the extent of missing data. One useful indicator to evaluate is the "proportion of data present" across waves—this is a set of percentage figures routinely produced by software such as M-plus—which gives one an idea of how many cases have data across waves of the panel (see below).

Table 2 presents three estimates of reliability based on different approaches to handling the problem, as follows: Column (1) presents estimates based on Heise's approach applied to Pearson correlations and based on a listwise data sample; column (2) is the reliability estimate based on the Allison approach using Pearson correlations, and column (3) presents the estimate of reliability, again using Pearson correlational data based on the FIML/WLSMV approach to handling missing data. In general, we recommend estimating reliability of measurement in more than one way, using both listwise and FIML/WLSMV estimates, and examine the differences. In our experience, one should only be concerned about the disparity between the two sets of results when the proportion of data present across waves is less than 20 percent.

Table 2. Estimates of reliability for survey measures by type of content and type of approach to missing data

	Heise Reliability Estimates (based on Pearson correlations)		
Content	Listwise Present[1]	Allison Method[2]	FIML Method[3]
Facts	0.7662	0.7267	0.7542
	(72)	(62)	(71)
Beliefs	0.4792	0.4674	0.4762
	(116)	(114)	(116)
Values	0.5547	0.5360	0.5385
	(43)	(43)	(43)
Attitudes	0.6034	0.5577	0.5522
	(76)	(76)	(76)
Self-Assessments	0.4966	0.4946	0.4923
	(25)	(25)	(25)
Self-Perceptions	0.5049	0.4946	0.5000
	(87)	(87)	(87)
Proxy Facts	0.8083	0.8170	0.8213
	(7)	(7)	(7)
	0.5692	0.5432	0.5540
Total	(426)	(414)	(425)

Note: number of measures in parentheses.
[1] *Estimates based on Pearson correlations and assumption of equal reliabilities.*
[2] *Estimates based on Pearson correlations and the assumption of equal reliabilities using Allison's multiple-*group *approach to missing data, using AMOS program.*
[3] *Estimates based on Pearson correlations and the assumption of equal reliabilities using Full-Information Maximum-Likelihood (FIML), using AMOS program.*
Source: adapted from Alwin (2007).

The results in Table 2 allow us to compare the average reliability estimates for common measures using the three strategies—the listwise, Allison and FIML/WLSMV approaches. The first column on the left employs the listwise approach, and the two columns on the right of the table employ estimates from the Allison multiple-group approach and from the FIML approach. Our first finding is that the Allison and FIML/WLSMV approaches are hardly different. Secondly, the Allison and FIML approaches to incomplete data provide estimates that are trivially different from those based on listwise samples. There may be some differences in the nature of the reliability estimates in the extreme cases, where little data are present across waves, but our experience indicates that estimates based on listwise and FIML approaches yield sufficiently similar estimates.

Results Regarding Missing Data from the GSS Data

In addition to the data from the *Margins of Error* study, with regard to handing missing data, we develop comparisons of two methods for estimating reliability taking missing data into account—the listwise approach and the FIML approach using General Social Surveys (GSS) data. The data is sourced from three panel studies, conducted by the GSS in 2006-08-10, 2008-10-12, and 2010-12-14. We examine the reliability for approximately 200 questions in each of the three GSS panels.

In Table 3 we present estimates of reliability for GSS non-redundant self- and proxy-reports by GSS panel, question content and the approach taken to missing data. Our data analysis employed a "hybrid" approach, using polychoric correlational methods for ordinal variables, and Pearson correlations for interval or continuous variables. We estimated Heise reliabilities using both a listwise-present approach and a full-sample approach. In the full sample approach, we employed weighted least squares means-and-variance adjusted (WLSMV) estimation for ordinal variables (variables with 10 or less categories) and full-information maximum-likelihood (FIML) for interval variables (variables with 11 or more categories). Results in table 3 reinforce our previous conclusion: both methods of treating missing data lead to highly similar results.

Table 3. Estimates of Reliability for GSS Non-redundant Self- and Proxy-Reports by GSS panel, question content and approach to incomplete data

	2006 GSS Panel			2008 GSS Panel			2010 GSS Panel		
Content	Measures	FIML/ WLSMV	Listwise	Measures	FIML/ WLSMV	Listwise	Measures	FIML/ WLSMV	Listwise
Facts	35	0.845	0.841	31	0.852	0.853	31	0.860	0.861
Non-facts	173	0.662	0.672	171	0.651	0.657	168	0.667	0.678
Beliefs	64	0.643	0.653	63	0.624	0.629	60	0.662	0.674
Values	42	0.689	0.694	42	0.657	0.672	42	0.664	0.671
Attitudes	35	0.664	0.679	35	0.684	0.686	35	0.666	0.682
Self-Assessments	12	0.643	0.654	12	0.645	0.649	12	0.669	0.682
Self-Perceptions	14	0.732	0.744	13	0.741	0.744	13	0.756	0.764
Expectations	6	0.523	0.560	6	0.523	0.506	6	0.550	0.560
Total	208	0.692	0.701	202	0.682	0.687	199	0.697	0.707
<u>Comparisons</u>									
<u>All content</u>									
F-ratio		11.271	9.203		11.409	10.885		10.090	8.885
p-value		0.000	0.000		0.000	0.000		0.000	0.000
<u>Facts vs. Non-facts</u>									
F-ratio		52.092	43.409		51.281	46.921		49.595	43.066
p-value		0.000	0.000		0.000	0.000		0.000	0.000
<u>Within Nonfacts</u>									
F-ratio		2.397	1.914		2.725	2.956		1.761	1.712
p-value		0.039	0.095		0.021	0.014		0.124	0.135

Across all three GSS panels, the Listwise approach generally results in slightly higher reliability estimates, compared to the FIML/WLSMV approach. All categories of content show the same trend, with isolated exceptions (facts in 2006 and expectations in 2008). The total reliability estimate is also slightly higher using the listwise approach. All differences in the size of reliability estimates are negligeable. The hierarchy in terms of reliability is the same when comparing the Listwise approach and the FIML/WLSMV approach across all three GSS Panels. The hierarchy, from highest to lowest reliability, is consistently: facts, self-perceptions, values, attitudes, self-assessments, beliefs, and finally expectations. Both approaches to missing data are sensitive to identifying differences or lack of differences in reliability between different categories of content. For example, both listwise and FILM/WLSMV identify statistically significant differences between the reliabilities of facts versus non-facts and for reliabilities across all categories of content. At the same time, both approaches point to non-statistically significant differences (at the $p = 0.001$ level of significance) within categories of non-facts.

Other Considerations

We recommend the creation of a data-base containing several approaches to estimating reliability, and including the following information on the proportion of data present (PDP): PDP (11), PDP (21), PDP (22), PDP (31), PDP (32), and PDP (33). PDP (11) refers to the data present at wave-1, PDP (21) refers to the data present at both waves 1 and 2. These numbers are readily obtained from contemporary SEM software, such as M-*plus* (Muthén and Muthén, 2004), and will help assist one in deciding how to interpret the differences between “listwise” and WLSMV/FIML estimates. Then, separately for “listwise” and “FIML” approaches, we recommend that the following information be assembled, which will permit the estimation of both Heise and Wiley-Wiley reliabilities: COR(21), COR(31), and COR(32). Note that for ordinal variables, these correlations will be "polychoric" correlations, and for continuous variables they will be "Pearson" correlations.

Conclusions

This document has reviewed several preliminary considerations in the use of the 3-wave *quasi-Markov simplex model* (QMSM) to assess question-specific reliabilities. Specifically we have addressed three issues: (1) what correlations we should use for assessing relationships of multiple measurements over time using longitudinal designs; (2) what assumptions must be made to identify the 3-wave *quasi-Markov simplex model*, for example, the assumptions of equal reliabilities over time, or equal error variances, or neither; and (3) how to handle missing data due to attrition and other sources of missingness. There are a number of additional issues that can be addressed, which are not covered here. For example, one can examine the question of the number of waves to include, and the tradeoffs between attrition and the ability to relax the restrictive assumptions of the quasi-Markov simplex model. The addition of $P > 3$ waves provides certain advantages, namely degrees of freedom available to test the model and its assumptions (see Cernat et al., 2021). We do not consider this issue here.

There are three main conclusions from the research so far. First, it is recommended that a "hybrid" approach be used with regard to the issue correlational data. Correlations based on traditional Pearson formulas are less well suited to variables that are ordinal (having ten or fewer response categories), and for those we recommend the use of polychoric correlations/covariances methods. Second, there seems to be very little difference between reliability estimates from models that identify the structure of errors over time in terms of equal reliabilities versus equal error variances over time. Third, with regard to the handling of missing data, the Allison and FIML approaches to incomplete data provide estimates that are trivially different from those based on listwise samples.

References

Alwin, Duane F. 2007. *Margins of Error: A Study of Reliability in Survey Measurement*. Hoboken NJ: John Wiley & Sons. [Wiley Series in Survey Methodology.

Alwin, Duane F. 2021. Developing Reliable Measures—An Approach to Evaluating the quality of Survey Measurement Using Longitudinal Designs. Pp. 113-154 in A. Cernat and J.W. Sakshaug (Eds.), *Measurement Error in Longitudinal Data.* Oxford UK: Oxford University Press.

Alwin, Duane F., Kristina Zeiser and Don Gensimore. 2014. Reliability of Self-reports of Financial Data in Surveys: Results from the Health and Retirement Study. *Sociological Methods & Research*, 43(1):98-136.

Alwin, Duane F., Erin M. Baumgartner, and Brett A. Beattie. 2018. Number of Response Categories and Reliability in Attitude Measurement. *Journal of Survey Statistics and Methodology*, 6(2):212-239.Cernat, Alexandru and Joseph W. Sakshaug (Eds.). 2021. *Measurement Error in Longitudinal Data.* Oxford UK: Oxford University Press.

Asparouhov, Tihomir, and Bength O. Muthén, B. 2010. Weighted Least Squares Estimation with Missing Data. Mplus Technical Appendix. Retrieved from https://www.statmodel.com/.

Cernat, Alexandru, Peter Lugtig, Nicole Watson, and S.C. Noah Uhrig. 2021. Assessing and Relaxing Assumptions in Quasi-Simplex Models. Pp. 155-171 in A. Cernat and J.W. Sakshaug (Eds.), *Measurement Error in Longitudinal Data.* Oxford UK: Oxford University Press.

Heise, David R. 1969. Separating Reliability and Stability in Test-retest Correlation. *American Sociological Review* 34:93-191.

Jöreskog, Karl G. 1970. Estimating and Testing of Simplex Models. *British Journal of Mathematical and Statistical Psychology* 23:121-145.

Jöreskog, Karl G. 1990. New Developments in LISREL: Analysis of Ordinal Variables Using Polychoric Correlations and Weighted Least Squares. *Quality and Quantity* 24:387-404.

Jöreskog, Karl G. 1994. On the Estimation of Polychoric Correlations and Their Asymptotic Covariance Matrix. *Psychometrika* 59:381-389.

Muthén, Bengt O. 1984. A General Structural Equation Model With Dichotomous, Ordered Categorical, and Continuous Latent Variable Indicators. *Psychometrika* 49:115-132.

Muthén, Linda.K. & Bengt O. Muthén. 2001-2004. *M-Plus—The Comprehensive Modeling Program for Applied Researchers*. *User's Guide*. Version 2.0. Los Angeles: Muthén & Muthén.

Wiley, David E. and James A. Wiley. 1970. The Estimation of Measurement Error in Panel Data. *American Sociological Review* 35:112-117.

Wothke, W. 2000. Longitudinal and Multigroup Modeling with Missing Data. Pp. 219-240 in T.D. Little, K.U. Schnabel and J. Baumert (Eds.), *Modeling Longitudinal and Multilevel Data: Practical Issues, Applied Approaches and Specific Examples*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.