An Approach to Estimating Reliability of Survey Measurement Using Panel Data

Duane F. Alwin

Introduction

In this document we present a non-technical discussion of our approach to estimating the reliability of survey measurement for specific survey questions. The estimation of reliability of survey data one needs, first, a model that specifies the linkage between true and observed variables; second, a research design that permits the estimation of parameters of such a model; and third, an interpretation of these parameters that is consistent with the concept of reliability (see Alwin and Krosnick, 1991). As we discuss in the following, this study meets these requirements, and here we discuss the rationale and background for thinking about measurement error using longitudinal data and the design requirements employed.

The estimates of reliability in this data base are based on the longitudinal analysis of panel survey data. The approach is rooted in classical ideas about measurement error, and it capitalizes on the availability of statistical models that separate measurement error from the measurement of change or stability in the underlying variables of interest. As an example, an individual's true-score on some variable, say income level, or years of schooling, or any variable that can be thought of as basically continuous, may change over time, and measurement error can contribute to the observed change. Our approach uses three waves of panel survey data to separate the amount of true change from the amount of random measurement error.

Our study design for estimating the reliability of survey questions requires (1) the use of large-scale panel studies that are representative of known populations, (2) with a minimum of three waves of measurement, and (3) separated by two-year re-interview intervals. In order to estimate the reliability of a particular survey question, it was necessary that questions were selected for use only if they were exactly replicated, that is, using the exact wording, response categories, mode of interviewing, etc.) across the three waves. It was also assumed that the question assessed an underlying variable that could be expressed as continuous (rather than categoric) in nature.

One of the main advantages of the re-interview or panel design using long (2 year) re-interview intervals is that under appropriate circumstances it is possible to eliminate the confounding of the systematic and random error components. In the panel design, by definition, measurement is repeated, and memory, or other systematic sources of error, must be ruled out. So, while this overcomes one limitation of cross-sectional surveys, namely the failure to meet the assumption of the independence of errors, it presents problems if respondents can remember what they said in a previous interview and are motivated to provide consistent responses (Moser and Kalton, 1972). Given the difficulty of estimating memory functions, estimation of reliability from re-interview designs makes sense only if one can rule out memory as a factor in the covariance of measures over time, and thus, the occasions of measurement must be separated by sufficient periods of time to rule out the operation of memory. In cases where the re-

measurement interval is insufficiently large to permit appropriate estimation of the reliability of the data, the estimate of the amount of reliability will most likely be inflated (see Alwin, 1989; 1992; Alwin and Krosnick, 1991), and the results of these studies suggest that longer re-measurement intervals, such as those employed here, are highly desirable.

The Quasi-Simplex Model

If one has three waves of panel data, and if certain other assumptions are met, you can use the 3-wave *quasi-simplex model* to obtain estimates of the reliability of measurement for individual survey questions. This model does not always work, but over the vast majority of trials, it does, and the results can tell you something about the *relative* quality of your data for a particular question. The rationale and justification for this approach, as well as the tools used for analysis are due to Coleman, Wiggins, Jöreskog, Heise and others (see the references given below).

This strategy throws out all the traditional approaches to reliability estimation based on composite measures and the use of tools such as Cronbach's alpha and related approaches to evaluate reliability. The focus is on the single survey question rather than the composite measure. This approach also rejects the simple test-retest approach by employing 3 waves of data, rather than two, which permits the model to account for both unreliability and true change in latent variable of interest. In the words of Dave Heise, the purpose is to "separate unreliability from true change."

I wrote a book on the subject (see D.F. Alwin, *Margins of Error—A Study of Reliability in Survey Measurement*, John Wiley & Sons, 2007), plus a dozen additional articles that employ this method, but I would advise that you do not use that book to try to figure out how to go about estimating reliability of measurement. In that monograph we tried a dozen different approaches, and although what we learned regarding how to go about estimating reliability can be captured from a detailed and thorough reading (and understanding) of that book, it is easy to get lost, if one does not already understand certain basic principles regarding reliability estimation.

The purpose of this brief note is to clarify how one can apply a simple set of principles to estimating the reliability of survey measures, assuming you have three waves of data, and certain other assumptions can be met. In other words, we have made progress in terms of how we recommend people use this approach, and we have boiled things down to a simple set of principles. In the following I briefly discuss what we learned about how to approach the estimation of reliability in 3-wave panel data, focusing on three issues: (1) what correlations we should use, (2) whether one assumes equal reliabilities over time, or equal error variances, and (3) how to handle missing data.

First, 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 you 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 (and related covariances). 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. The latter correlations estimate the correlation for a true underlying variable that is continuous.

Based on extensive analysis of this issue (see Alwin, 2007), we recommended using a "hybrid" approach. 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.

Second, there are two basic approaches to modeling the error structures using these 3-wave *quasi-simplex models*. One is the approach of Dave Heise (ASR, 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 him or herself to the fact that the process being modeled is more complicated that this model supposes. Such results are rare, but when they occur, it is usually a tip that there is something more complicated going on and one cannot make the assumption of "dynamic equilibrium" (see Alwin, 2007).

If one is satisfied with the Heise model, and most everything I have seen leads us to prefer this approach, then one can proceed with the results and analyze differences among survey questions in their levels of reliability. Still, there were some serious issues raised in the paper by David and James Wiley (ASR 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.

When covariance data exist, 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 we recommend that one should do this because it can be informative. Of course, the reliability of wave-2 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. In our experience, they rarely are, but still, we recommend computing these separate reliabilities when you can 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 another wave of data, or a multiple group approach (see Alwin, 2007).

In the case of ordinal measures, in contrast to continuous measures, it is more difficult to obtain a covariance matrix among the variables. Some approaches have been taken to obtaining an asymptotic covariance matrix for ordinal variables, but these are not universally-accepted approaches. Although we have tried them, they do not produce substantially different estimates than the correlational approaches, so we recommend that for ordinal measures, one simply rely on polychoric correlations and the Heise estimates. In other words, there is really little to be gained to obtain Wiley-Wiley-type estimates for ordinal data, so in practice we rely solely on Heise estimates in this case. On the other hand, we routinely obtain Wiley-Wiley wave-specific estimates of reliability in the case of continuous variables.

Missing Data

A final issue that arises in the use of 3-wave *quasi-simplex models* to estimate the reliability of survey measures is how to handle missing data. As everyone knows, attrition is a perennial problem in the implementation of panel surveys. One 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 wave of the panel. An alternative explored in that monograph. however, was *full-information maximum-likelihood* (FIML), which estimates the correlations (either Pearson or polychoric) using all information present. 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 wave—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.

Estimates of reliability were obtained from both list-wise and full-samples, the latter using weighted least squares mean- and variance-adjusted (WLSMV) (Asparouhov and Muthén, 2010) or full-information maximum-likelihood (FIML) (Allison, 1987; Wothke, 2000) to handle missing data due to attrition and other causes. Consistent with prior studies, results indicate (data not shown) that listwise and WLSMV/FIML estimates were virtually identical, suggesting an MCAR pattern to attrition and missing data (see also, Alwin, Beattie and Baumgartner, 2015). There appears to be a very slight tendency for the listwise estimates to be higher, but this result is not statistically significant. Due to the identical nature of these obtained results for the remainder of the paper we present only one set of estimates, specifically the estimates based on listwise data.

We recommend estimating reliability of measurement both ways, using both listwise and FIML estimates. 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. 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 to alleviate any concerns about substantial differences. Still, we recommend doing one's analysis both ways.

To summarize, we recommend that one create a data base containing the following information:

PDP (11)
PDP (21)
PDP (22)
PDP (31)
PDP (32)
PDP (33)

The above numbers are readily obtained from Mplus and will help assist one in deciding how to interpret the differences between "listwise" and "FIML" estimates.

Then, separately for "listswise" and "FIML" approaches, we recommend that the following information be assembled, which will permit the estimation of Heise reliabilities. Note that for ordinal variables, these correlations will be "polychoric" correlations, and for continuous variables they will be "Pearson" correiatons:

COR(21)
COR(31)
COR(32)

Using these correlations, it is possible to estimate "Heise reliabilities" – once for the listwise approach and once for the FIML approach – as given above. Note also that we currently use the Mplus 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 go 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)]

In general, we have found that these separate wave-specific reliability estimates are not very different, but this set of operations can provide additional insight into whether one is using the correct model for the data. Examples can be provided, e.g. see estimates for the reliability of income in Alwin, Zeiser and Gensimore (SMR, 2014), in which wave-specific reliabilities are presented.

One can obtain listwise and FIML correlations (both Pearson and polychoric) from a number of different software platforms—we choose M-plus mainly because that is where we can

easily obtain the proportion of data present across waves, as well as everything else we need. It will not produce an asymptotic covariance matrix for ordinal variables, but that should not get in the way, for the reasons given above.

Our study design requires the use of large-scale panel studies that are representative of known populations, with a minimum of three waves of measurement separated by two-year re-interview intervals. Questions were selected for use if they were exactly replicated (exact wording, response categories, mode of interviewing, etc.) across the three waves, and if the underlying variable measured was continuous (rather than categoric) in nature.

As we noted earlier, one of the main advantages of the re-interview or panel design using re-interview intervals of at least 2-years is that under such circumstances it is possible to eliminate the confounding of the systematic and random error components. Given the difficulty of estimating memory functions in survey research, estimation of reliability from re-interview designs makes sense only if one can rule out memory as a factor in the covariance of measures over time, and thus, the occasions of measurement must be separated by sufficient periods of time to rule out the operation of memory. In cases where the re-measurement interval is insufficiently large to permit appropriate estimation of the reliability of the data, the estimate of the amount of reliability will most likely be inflated (see Alwin, 1989; 1992; Alwin and Krosnick, 1991), and the results of these studies suggest that longer re-measurement intervals, such as those employed here, are highly desirable.

References

Alwin, Duane F. 2005. Reliability. In K. Kempf-Leonard and others (Eds.), *Encyclopedia of Social Measurement*. New York: Academic Press.

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

Alwin, Duane F. 2016. Survey Data Quality and Measurement Precision. Pp. 527-557 in Christof Wolf, Dominique Joye, Tom W. Smith, and Yang-chih Fu (Eds.), *The SAGE Handbook of Survey Methodology*. London: SAGE International Publishers.

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

Coleman, James S. 1964. *Models of Change and Response Uncertainty.* Englewood Cliffs, NJ: Prentice-Hall.

Coleman, James S. 1968. The Mathematical Study of Change. In H.M. Blalock, Jr. and A.B. Blalock (Eds.) *Methodology in Social Research* (pp. 428-478). New York: McGraw-Hill.

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.

Moser, C.A. and Graham Kalton. 1972. Survey Methods in Social Investigation. 2nd Edition. New York: Basic Books.

Wiggins, Lee M. 1973. *Panel Analysis: Latent Probability Models for Attitude and Behavior Processes*. New York: Elsevier Scientific Publishing Company.

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. In T.D. Little, K.U. Schnabel and J. Baumert (Eds.), *Modeling Longitudinal and Multilevel Data: Practical Issues, Applied Approaches and Specific Examples* (pp. 219-240). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Further Reading

Alwin, Duane F. 1989. Problems in the Estimation and Interpretation of the Reliability of Survey Data. *Quality and Quantity* 23:277-331.
Alwin, Duane F. 1992. Information Transmission in the Survey Interview: Number of Response Categories and the Reliability of Attitude Measurement. In P.V. Marsden (Ed.), *Sociological Methodology 1992* (Pp. 83-118). Washington D.C.: American Sociological Association.
Alwin, Duane F. and Brett A. Beattie. 2016. The Kiss Principle in Survey Measurement—Question Length and Data Quality. *Sociological Methodology*, vol. 46, Forthcoming.
Alwin, Duane F. and Jon A. Krosnick. 1991. The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes. *Sociological Methods and Research* 20:139-181.
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 and Research* 43:98-136.
Alwin, Duane F., Brett A. Beattie, and Erin M. Baumgartner. 2015. Assessing the Reliability of Measurement in the General Social Survey: The Content and Context of the GSS Survey Questions. Paper presented at the session on "Measurement Error and Questionnaire Design," the 70th annual conference of the American Association for Public Opinion Research. May.