The Canadian Journal of Program Evaluation Vol. 10 No. 2 Pages 17–32
ISSN 0834-1516 Copyright © 1995 Canadian Evaluation Society

ON THE DIFFERENCE BETWEEN RELIABILITY OF MEASUREMENT AND PRECISION OF SURVEY INSTRUMENTS

Brian Evans
Corporate Review
Department of Canadian Heritage

Abstract: Despite the importance of assessing the reliability of evaluative measures, there is a confusing array of conceptual schemes and coefficients for calculating "the reliability" of an item or scale. In particular, reliability may be conceived of and estimated from a true-score model or a sampling precision perspective. The former model is associated with such estimates as parallel or alternate forms reliability, split-half reliability, and coefficient alpha, the latter with standard error, coefficient of variation, and confidence intervals for observed scores. This review clarifies the distinction between two models. The basic theoretical models for each approach are developed and illustrated using data from the author's work on measuring organizational climate. As a result, evaluators should be better able to judge the meaning of the reliability information provided in reports and to calculate reliability in situations requiring some assessment of the quality of their data.

Résumé: Malgré l'importance d'évaluer la fiabilité des mesures d'évaluation, il y a une série confuse de schémas conceptuels et de coefficients spécifiques de fiabilité dans le calcul de la *fiabilité* d'un élément ou d'une échelle. On peut concevoir et estimer la fiabilité à partir d'un modèle de valeurs vraies ou dans le contexte de la précision de l'échantillonnage. Le premier modèle est associé aux estimations telles que les versions parallèles ou alternatives de la fiabilité, la fiabilité moitié-moitié, et le coefficient alpha, le deuxième à l'écart type, le coefficient de variation, et les intervalles de confiance des scores observés. Ce survol précise la différence entre les deux modèles. Les modèles théoriques de base pour chaque approche sont développés et illustrés grâce à des données tirées des travaux de l'auteur sur la mesure du climat organisationnel. Les évaluateurs seront donc en meilleure position pour juger la signification des renseignements sur la fiabilité contenus dans les rapports et de calculer la fiabilité lorsqu'ils voudront obtenir une certaine évaluation de la qualité de leurs données.

Introductory textbooks on research methods and measurement, including those aimed at evaluators, routinely include discussions of the concept of reliability (Henderson, Morris, & Fitz-Gibbon, 1987; Morris, Fitz-Gibbon, & Lindheim, 1987; Posavac & Carey, 1985). Typically, the textbook model of reliability is based on classic psychometric theory that views an individual's observed response on a survey or in an interview as a result of that individual's underlying, true attitude or characteristic and of random measurement error. Reliability in this model refers to the extent to which an observed score or response is related to the individual's true attitude or characteristic. Within this tradition, different types of reliability (i.e., temporal stability, equivalence, internal consistency) and several estimators of the reliability of an item or scale (e.g., test-retest, split-half, coefficient alpha) are distinguished. As each of these approaches yields a different estimate of "the reliability" of an item or scale, clearly reliability even within the psychometric approach is not a simple concept. Rather, reliability may be thought of as a generic term applied to several specific estimators, all of which are concerned with consistency of measurement but differ in the sources of measurement error that are considered (Pedhazur & Schmelkin, 1991; Selltiz, Wrightsman, & Cook, 1976).

The terms *reliability* and *precision* have both been used by survey researchers to refer to consistency of response (Alreck & Settle, 1985; Bustros & Kelly, 1992; Grove, 1989; Satin & Shastry, 1988). Some authors (Dutka & Frankel, 1993) have suggested that the survey-sampling use of the term *precision* parallels the meaning of the psychometric concept of reliability. However, the psychometric concept of reliability and the survey-sampling concept of reliability and precision differ fundamentally in how survey error or lack of consistency is understood and estimated. The psychometric model is associated with estimates of parallel or alternate form reliability, test-retest reliability, split-half reliability, and coefficient alpha. The survey sampling approach is associated with estimates of the standard error, the coefficient of variation, and confidence intervals for observed scores. Consequently, the different approaches to reliability can lead to different conclusions about the quality of survey data (Grove, 1989).

This brief review is intended to help evaluators avoid potential confusion and make sense of competing concepts and estimates of reliability. The focus is on application of reliability in survey research. As a result, evaluators should be in a better position to judge the meaning of the reliability information provided in reports and to calculate reliability in situations requiring some assessment of the quality of their data.

THE PSYCHOMETRIC MODEL OF RELIABILITY

As noted above, the dominant model of reliability adopted by most textbooks on research methodology and measurement divides an individual's observed score on a survey or their response in an interview into true-score and random error components. This model is typically represented as:

$$X = T + E$$

where X is an observed score, T is an individual's true attitude, and E is random error in measurement. True scores may be thought of as the mean of an individual's scores given an infinite number of measurements. Random error includes the effects of interviewer behavior; the mode of data collection; the characteristics of questions, such as topic, wording, context, and response formats; and characteristics of the respondents (Alwin, 1989; Carmines & Zeller, 1979).

The classic psychometric model of reliability makes several assumptions about the relationship between true and error scores (Allen & Yen, 1979), including the following:

1. The expected value (i.e., the mean) of the observed scores is the expected value of the true scores ($\mathrm{E}(X) = T$).
2. The expected value of the errors is zero ($\mathrm{E}(E) = 0$).
3. The correlation of error and true score is zero $(E,T) = 0$.
4. The error of variable X is uncorrelated with the errors of other variables $(E_1, E_2) = 0)$.
5. The error of variable X is uncorrelated with the true score for another variable $(E_1, T_2) = 0)$.

Given these assumptions, it can be shown that the variance of an observed score equals the sum of true-score variance and error variance, that is:

$$\sigma^2_x = \sigma^2_t + \sigma^2_e$$

The traditional definition of reliability follows directly from this result. Reliability is the extent to which the observed variance (σ^2_x) of a survey response is true-score variance (σ^2_t) rather than random error variance (σ^2_e). In other words, the reliability (p) of x is simply the ratio of true-score variance to the observed variance:

$$p_x = \sigma^2_t / \sigma^2_x$$

If the ratio is one, then all observed-score variance is true-score variance and the item or scale is perfectly reliable. When the ratio equals zero, all the observed-score variance is error variance. Reliability coefficients may therefore be interpreted as the percentage of observed-score variance that is accounted for by variation in true scores.

Calculation of specific reliability coefficients requires additional assumptions to estimate what portion of observed variance is true-score variance and what portion is error variance. Different sets of assumptions have been used to derive the required statistics. Perhaps the oldest and among the most common are the assumptions of the parallel measurement model (Allen & Yen, 1979; Pedhazur & Schmelkin, 1991). If two measures of a construct have equal true-score variance and equal error variance, they are referred to as parallel measures. A less restrictive version of the model assumes that only the true-score variances are equal, in which case the measures are referred to as tau-equivalent. Given either of these sets of assumptions, it can be shown that the correlation between the measures is equal to the reliability of the measure. That is, the correlation is equal to the ratio of true-score to observed variance.

Within the psychometric tradition, the fact that under some conditions the correlation between measures is equal to the reliability of the measure is applied to derive several different reliability coefficients. One application focuses on the temporal stability of a set of scores. That is, the same or alternate measures of a single construct are administered on separate occasions and the results correlated to yield an estimate of the temporal stability or reliability of scores over time. When the same scale is administered on two occasions this estimate is referred to as a test-retest estimate of reliability.

A second application is based on the correlation between two items or scales administered at one time. These methods estimate the equivalence or internal consistency of items or scales rather than the temporal stability. Several methods exist for estimating internal consistency, including correlating the results of alternate forms administered on a single occasion and correlating the scores from a single scale that has been split into two parts (i.e., split-half estimates of reliability). In the latter case, the resulting correlation is adjusted to reflect the increased scale length, and resulting increase in reliability, that comes from using the total scale instead of the half-scales used to generate the correlation.

Although alternate forms and split-half reliability coefficients continue to be recommended in literature aimed at evaluators (see, e.g., Posavac & Carey, 1985), the most widely used measure of internal consistency in the social sciences is coefficient alpha (Pedhazur & Schmelkin, 1991). Unlike other estimates of reliability, which are based on the correlation between two measures of a construct, coefficient alpha is calculated from either the covariance or correlations between multiple (i.e., two or more) measures of a construct. Although coefficient alpha is distinguished from other estimators of both temporal stability and internal consistency by its use of all the information in an item variance-covariance matrix, it still rests on the basic true-score model. That is, it still partitions the observed variance into true and random error variance components.

In the parallel measurement model, the idea that a particular reliability estimate may not represent the true reliability of an item or scale arises from the failure of actual measures to meet the assumption of the theoretical measurement model. As noted above, the use of correlations as estimates of the reliability of measures is predicated on the assumption that the measures are at least tau-equivalent. If measures are tau-equivalent, then the correlation between them is an exact specification, not an estimate, of the scale reliability. In the absence of tau-equivalent measures, correlations are biased estimates of item or scale reliability. In estimating the stability of scores, failure to meet the assumptions of the measurement model can lead to either over- or underestimation of the true reliability of measure. Estimating the internal consistency of a set of items while failing to meet the assumptions of the measurement model will lead to underestimating the reliability of a scale. Consequently, coefficient alpha is sometimes referred to as the lower bound for the reliability of a scale.

RELIABILITY AND PRECISION OF ESTIMATES

Although most discussions of reliability derive from the psychometric tradition, some authors have used the term *reliability* in a way that is synonymous with *measures of precision* of survey estimates.

Alreck and Settle (1985), for example, discuss the standard error and the range of the confidence interval associated with a statistical estimate as measures of reliability. Satin and Shastry (1988) refer to the precision or reliability of sample results as expressed in sampling variance, the standard error, and the coefficient of variation. Bustros and Kelly (1992), in their guide to employee surveys in the Canadian federal government, discuss the level of reliability in terms of the desired precision of the estimates, the margin of error that will be tolerated, and the coefficient of variation.

To understand the sampling precision approach to reliability, it is helpful to recall how terms such as *variance*, *standard error*, *coefficient of variation*, and *confidence interval* are used. An essential distinction concerns variability of the elements in a sample or population, and variability among estimates from repeated samples. The terms *variance* and *standard deviation* refer to measures of variability among elements in a sample or population. The term *standard error* is used to refer to the standard deviation of estimates from repeated samples (Frankel, 1983).

For elements in a sample, the standard deviation is the square root of the variance. The standard error for a simple random sample is the standard deviation divided by the square root of the number of elements in the sample:[1]

$$S_e = \frac{S_d}{\sqrt{n}}$$

One characteristic of these measures of variability (i.e., variance, standard deviation, standard error) is that the absolute magnitude of the estimate will depend on the magnitude of the original scale units. Therefore, the variability or, conversely, the precision of items measured on different scales is not directly comparable.

The coefficient of variation (CV) creates a scale-free measure of variability that permits direct comparisons of the variability of concepts measured in different units. Weisberg (1992) gives the formula for the CV as:

$$CV = \frac{S_d}{\bar{x}}$$

Several variations of this basic formula are found in the literature. For example, in Proc. Univariate of the SAS-PC program, the Weisberg formula is multiplied by 100. More fundamentally, Satin and Shastry (1988) include the square root of the number of elements in the sample as part of the denominator in their version of the formula. This, of course, changes the formula into an estimator for a sampling distribution rather than an estimator for a single sample.

The precision of point estimates (i.e., individual scores, group means or proportions) can also be assessed with confidence intervals. A confidence interval expresses the level of confidence that the true value for the population lies within a specific range of values (Satin & Shastry, 1988). In the sampling precision approach confidence intervals are constructed based on the formula (Frankel, 1983; Satin & Shastry, 1988)

$$X = \pm \mathrm{Z_a\ S_e}$$

where $\mathrm{Z_a}$ is the standard normal deviate used to represent a given confidence level (i.e., 1.0 for a 68% confidence interval, 1.25 for a 90% confidence interval, 1.96 for 95% confidence, and 2.58 for 99% confidence), and $\mathrm{S_e}$ refers to the standard error of the estimate as defined above. It can readily be seen from this that for a fixed level of error, greater confidence in a point estimate will be associated with larger confidence intervals. That is, a 95% confidence interval will be associated with a greater range of values than a 68% confidence interval.

In contrast to the survey sampling literature, the typical formula for calculating confidence intervals in the psychometric literature is

$$X = \pm Z_a S_d \sqrt{(1 - r_{xx})}$$

where the term

$$(S_d)\sqrt{(1 - r_{xx})}$$

is referred to as the standard error of measurement (Allen & Yen, 1979; Pedhazur & Schmelkin, 1991). In this term $\mathrm{S_d}$ refers to the sample standard deviation and r_{xx} refers to the reliability coefficient for the measure. Given the assumptions of classical psychometric theory, it is possible to demonstrate algebraically that the standard error of measurement is equivalent to the error variance term in the classic model of reliability (Allen & Yen, 1979). Clearly, in the psychometric approach the standard error of measurement replaces the standard error of the estimate used for constructing confidence intervals in sampling precision approach. The two approaches differ in the term used to represent overall variability (i.e., either S_e or S_d) and in the inclusion of the term

$$\sqrt{(1 - r_{xx})}$$

in the psychometric formula. This latter term serves to reduce the total sample variability represented by S_d, and by extension, to reduce the estimated confidence interval. It should be clear from this that in psychometric theory the reliability of an item or scale is a factor that influences the precision of an estimate. In contrast, discussions of reliability in the sampling precision perspective have treated reliability as equivalent to the precision of the estimate.

The use of the different estimates of overall variability in the two formulas results from the differential emphasis in the sampling precision approach and psychometric theory on individual-level and group-level statistics. Texts on sampling typically show the formula for confidence intervals for group-level statistics (i.e., means, proportions) given repeated samples. There is little interest in the precision of individual-level statistics (i.e., individual scores) (Frankel, 1983; Satin & Shastry, 1988). In contrast, psychometric theory developed in a context of measuring individual abilities and aptitudes. In this context, the measurement specialist was often interested in the precision of individual scores for purposes of selection or placement given a single sample (Stanley, 1971). This differential emphasis accounts for the different estimates of overall variability in the formula for calculating confidence intervals. Confidence intervals for individual scores based on the sample S_d will be wider than confidence intervals for group statistics based on the S_e. That is, we are more likely to have confidence in estimates of group statistics that reflect many individual scores than in the estimate of a single score.

Variations of the standard error of measurement for use in calculating confidence intervals for group-level statistics have been proposed (Brown, 1976, cited in Posavac & Carey, 1985, p. 59) where the S_e replaces the S_d term found in the classic formula. This results in a reduced confidence interval compared with the classic formula. It also results in a reduced confidence interval compared with the sampling precision approach, as only part of the overall variability represented by S_e is used to construct the confidence interval.

The psychometric approach to confidence intervals differs from the sampling precision approach in one other important respect. Discussions of confidence intervals from the psychometric perspective frequently focus on the desirability of constructing intervals that are symmetric around true scores rather than observed scores (Pedhazur & Schmelkin, 1991; Stanley, 1971). Calculation of individual true scores is possible given knowledge of the reliability of an item or scale.

Individual observed scores, especially those at the extremes of the distribution, tend to be biased estimates of true scores. In consequence they will show regression toward the observed mean because of measurement error (Stanley, 1971). A similar problem occurs when one is investigating change at either the individual or group level. When extreme scoring groups are retested, error in measurement can affect the average gain or loss score for the groups. Imagine a comparison of those who received a particular program intervention and those who did not; the relative gain or loss for the two groups may be distorted because of regression toward the mean owing to measurement error.

Although individual observed scores may be biased estimates of individual true scores, and changes in observed scores may be biased because of measurement error, in classical true-score theory the observed mean is an unbiased estimated of the mean of the true scores (i.e., ($E(X) = T$). Therefore, a confidence interval for the group mean calculated using the appropriate standard error of measurement would be symmetrical around both the observed and true-score mean, as these are the same value.

This latter point is true given the assumptions of the parallel measurement model. Alternative models for deriving estimates of reliability within the psychometric tradition, such as the domain sampling model (Nunnally, 1978), view scales as composed of samples of items from a hypothetical domain of all possible items measuring a construct. In this model, the observed means of two measures of a construct can differ because of the differential sampling of items from the domain of all possible items. Therefore, a sampling distribution of observed means is possible. The mean of this sampling distribution is the population true-score mean. How precisely a particular mean represents the true mean is given by the standard error of the estimate as defined above. Lack of precision in this case results from the sampling of items rather than people.

The domain sampling model also allows for the possibility that estimates of item or scale reliability may be more or less precise. In the classic parallel test model, reliability is specified exactly by the correlation between two parallel measures. In contrast, the domain sampling model derives the reliability of a scale from the average correlation between items on a scale. Variations in individual correlations between pairs of items will lead to some imprecision in estimating the "true" reliability of a scale from a sample of items. It is possible to estimate the standard error associated with the average observed correlation and, by extension, to estimate the error associated with estimates of the reliability of a scale (Nunnally, 1978). The fact that the reliability of a measure may be more or less precise again highlights the difference between the psychometric approach to reliability and the sampling precision approach. In the former case, reliability is something that is more or less precisely estimated; in the latter case, reliability and precision are the same thing.

A NUMERICAL EXAMPLE

Calculation of the various estimates of variability is demonstrated with a numerical example drawn from the author's work on measuring organizational climate. Two variables, overall job satisfaction and days absent from work, were chosen to illustrate the properties of the various statistics. Data were collected by mail survey from 1,585 employees (63%) of a Canadian federal government department. Overall job satisfaction was measured by three items assessing the degree to which the respondents found their job interesting and felt satisfied with their job. Each item was scored on a five-point scale, and scores on each item were summed to create an overall job satisfaction score ranging from three to fifteen. A single question, "How many days of sick leave have you used over the last twelve months?", was used to assess days absent from work, with a range of 0 to 120 days absent reported. A total of 1,575 respondents provided data on all four items, the results of which are shown in Table 1. The reliability of the job satisfaction scale reported in Table 1 is the coefficient alpha referred to above. Given that a single item was used to measure days absent, it was not possible to actually calculate the reliability of this measure. For purposes of illustration, the reliability of this item was set to the same value as the reliability of the job satisfaction scale.[2]

As noted above, variability estimates such as variance, standard deviation, and standard error differ in magnitude when the original

Table 1
Summary Measures of Variability and Error for Job Satisfaction and Days Absent From Work

Statistic		Job Satisfaction	Days Absent
Mean		11.03	5.43
Reliability		.85	.85
Variance		6.99	68.85
Standard dev.		2.64	8.30
CV	$CV = \frac{S_d}{\bar{x}}$	.24	1.53
Standard error of mean	$S_e = \frac{S_d}{\sqrt{n}}$	.07	.21
SEM–s_d	$(S_d)\sqrt{(1-r_{xx})}$	1.02	3.21
SEM–S_e	$(S_e)\sqrt{(1-r_{xx})}$	.03	.08
95% CI—— S_e		10.08 to 11.67	4.99 to 5.81
95% CI—— SEM–S_d		9.04 to 13.02	2.19 to 7.59
95% CI—— SEM–S_e		10.98 to 11.09	5.24 to 5.56

CV = coefficient of variation
SEM = standard error of measurment
CI = confidence interval

units of measurement are different and are not directly comparable for the two variables. However, the CV is a scale-free measure of variation and can be directly compared for the two variables. In scale-free units, respondents varied less in overall job satisfaction than in days absent from work. In other words, in this sample mean job satisfaction is estimated more precisely than mean days absent from work.

Comparisons of the standard error of the mean (S_e) and the standard error of the measure for individual (SEM–S_d) and group (SEM–S_e) scores show that the classic psychometric formula SEM-S_d results in the largest error value for each variable, followed by the S_e and then the SEM–S_e. This is to be expected given the fact that SEM–S_d uses the S_d as a measure of total variability in the sample, whereas the S_e uses only a fraction of the S_d as a measure of total variability and the SEM–S_e uses only a portion of S_e as a measure of variability. Given this set of relationships, it follows that the CI–S_e calculated from the traditional sampling precision formula is narrower than the CI based on the SEM–S_d, but wider than the CI based on the SEM–S_e.

It should be noted that the application and interpretation of these CIs is different in each case. The interval based on the SEM–S_d is applied to individual scores. In this case, it was applied to the hypothetical individual score that was identical to the group mean. In this case, 95% of the intervals constructed using this procedure will include the individual's score. Both the intervals based on the S_e and those based on the SEM–S_e apply to the group mean itself and are correspondingly shorter than the interval for the hypothetical individual score.

In general, CIs constructed using the SEM treat only part of the total variability as measurement error. If the reliability of the measure is perfect (i.e., 1), the confidence interval reduces to zero and there is no error of measurement around the observed value. If the item or scale is totally unreliable (i.e., 0), all the observed variability, whether measured by the S_d or the S_e, is treated as error variance. Therefore, when reliability is equal to zero, the psychometric and sampling precision approaches to constructing confidence intervals will yield identical results for group-level statistics. When reliability is greater than zero but less than one, the confidence interval for the group mean calculated by the psychometric approach will be smaller than that calculated by the sampling precision approach. The interval becomes progressively smaller as the reliability of the item or scale increases.

It should be clear by this point that estimates of variability or, conversely, precision (i.e., coefficient of variation, standard error, and confidence intervals for observed scores) that are identified as measures of reliability by those writing from a sampling precision perspective (Alreck & Settle, 1985; Bustros & Kelly, 1992; Satin & Shastry, 1988) are not measuring reliability as defined in the psychometric model. None of these statistics partitions the observed variance into true and error score components, which is the hallmark of the psychometric approach. Instead, they simply express the observed variability in different units.

The different approaches to the concept of reliability in the two traditions are a result of more fundamental differences in how each tradition has dealt with problems of survey error (Grove, 1989). In survey-sampling approaches to error a distinction is often drawn between errors that contribute to *sampling bias* and those that contribute to *sampling variance* (Alwin, 1991). Reliability in the traditional psychometric sense is concerned with the effect of measurement error on variance. Bias is not a meaningful concept in the classic true-score model (Bohrnstedt, 1983). Instead, psychometric theory has dealt with the concept of bias or systematic error under the rubric of validity. Unlike sampling theory, psychometric theory has rarely tried to quantify the degree of bias. The concept of precision in the survey-sampling perspective is also concerned with sampling variability. However, it also considers all the sources of potential variability, including such nonmeasurement sources of error as coverage, nonresponse, and sampling (Grove, 1989). As a result, the sampling approach produces larger estimates of error than the psychometric approach when computing the errors associated with group-level statistics.

The two approaches have also differed in their use of population models of survey error. Both Alwin (1989; 1991) and Grove (1989) have noted the use of a population model for measurement error in the psychometric approach. That is, random measurement error is assumed for each element in the population. Reliability is therefore a joint characteristic of measurement instruments and populations. Nonmeasurement errors can affect the accuracy of sample characteristics (e.g., means and variances) in this perspective. However, these latter types of error are characteristics of samples and not populations. They should be thought of as logically subordinate to measurement errors. In contrast, all types of errors from the sampling precision perspective are characteristic of samples and not of populations.

CONCLUSION

Reliability is a complex construct involving multiple meanings and methods of estimation. The psychometric approach to reliability is based on the partitioning of the observed item or scale variance into error and true-variance components. Within the psychometric tradition this basic model is applied in different situations to estimate different types of reliability (i.e., temporal stability and internal consistency). The topic acquires additional complexity because of the use of the term *reliability* in the survey-sampling literature. Unlike in psychometric theory, where reliability is concerned with the partitioning of observed variance, in the sampling precision approach all observed variance around a point estimate is treated as error variance. This difference in basic approach means that statistics associated with the survey-sampling approach, such as variance, standard error, the coefficient of variance, and the confidence intervals for observed scores, are not synonymous with reliability coefficients such as alternate form, split-half, and alpha calculated from within the psychometric approach.

In general, the psychometric concept of reliability is not routinely applied in assessing the quality of survey data (Alwin, 1991; Bohrnstedt, 1983). The reason for this seems straightforward. Reliability, as defined in the psychometric tradition, is not estimable in most survey situations that consist of cross-sectional designs and single items measuring various constructs. In addition, Schuman and Presser (1981) suggest that survey researchers have avoided the psychometric approach because it focuses attention on abstract constructs rather than specific attitudes or behavior. However, others have shown that the classic psychometric model of reliability may be usefully applied to estimation of the specific attitudes, behaviors, and characteristics of populations (Alwin, 1989; Bollen, 1989). In addition, several survey researchers have begun to estimate the reliability of single items on surveys using structural equation modeling techniques (see, e.g., Alwin, 1989; Alwin & Krosnick, 1991; Bohrnstedt, Mohler, & Muller, 1987).

Whether an evaluator will adhere to a sampling or psychometric theory approach to reliability likely depends on professional background. The psychometric concept of reliability is strongly associated with the discipline of psychology (Grove, 1989; Nunnally, 1978; Stanley, 1971). Evaluators who are trained in this and related traditions (e.g., education) will almost certainly conceive of and report reliability in terms of the psychometric model. Whether evaluators from other social science, business, or statistical backgrounds will prefer a psychometric or the sampling precision model of reliability depends on their particular professional training. The experience of the author is that most but not all evaluators have some exposure to the psychometric concept of reliability. Furthermore, exposure to the psychometric model bears little relationship to an evaluator's methodological or statistical sophistication. Some sophisticated researchers are not familiar with the psychometric concept of reliability. Conversely, many evaluators are unfamiliar with the use of the term *reliability* to refer to sampling precision. In any case, referring to estimates of error variance that are due to both sampling and measurement errors as measures of reliability only adds confusion to an already confusing topic. The term *reliability* should be reserved for the variance partitioning approach to error and the term *precision* for the survey-sampling approach to error.

NOTES

1. Writers concerned with sampling theory typically include a term correcting for sampling from finite populations in formulas for estimating variance and standard errors (Frankel, 1983; Satin & Shastry, 1988). Not including this term in the formulas simplifies the presentation and does not affect the point being made. Also note that the formula for standard error changes under stratified and cluster sampling designs.

2. Given at least three waves of data collection it is possible to estimate the reliability of single items on surveys. Alwin and his colleagues (Alwin, 1989; Alwin & Krosnick, 1991) have reported several studies of the reliability of survey items using this procedure. Typically, they find reliabilities of .8 to .9 for demographic items. Therefore, this estimate may in fact be a reasonable approximation of the reliability of the item.

REFERENCES

Allen, M.J., & Yen, W.M. (1979). *Introduction to measurement theory*. Belmont, CA: Wadsworth.

Alreck, P.L., & Settle, R.B. (1985). *The survey research handbook*. Homewood, IL: Irwin.

Alwin, D.F. (1989). Problems in the estimation and interpretation of the reliability of survey data. *Quality and Quantity, 23*, 277–331.

Alwin, D.F. (1991). Research on survey quality. *Sociological Methods and Research, 20*, 3–29.

Alwin, D.F., & Krosnick, J.A. (1991). The reliability of survey attitude measurement. *Sociological Methods and Research, 20*, 139–181.

Bohrnstedt, G.W. (1983). Measurement. In P.H. Rossi, J.D. Wright, & A.B. Anderson (Eds.), *Handbook of survey research* (pp. 70–121). San Diego: Academic Press.

Bohrnstedt, G.W., Mohler, P.P., & Muller, W. (Eds.). (1987, February). *Sociological Methods and Research* (special issue), *15*(3).

Bollen, K. (1989). *Structural equations with latent variables*. New York: Wiley.

Bustros, J., & Kelly, K. (1992). *Guide to conducting an employee opinion survey in the federal public service*. Ottawa: Human Resources Development Branch, Treasury Board Secretariat.

Carmines, E.G., & Zeller, R.A. (1979). *Reliability and validity assessment* (Sage University Paper Series on Quantitative Applications in the Social Sciences 07–017). Newbury Park, CA: Sage.

Dutka, S., & Frankel, L.R. (1993). Measurement error in organizational surveys. In P. Rosenfeld, J.E. Edwards, & M.D. Thomas (Eds.), *Improving organizational surveys*. Newbury Park, CA: Sage.

Frankel, M. (1983). Sampling theory. In P.H. Rossi, J.D. Wright, & A.B. Anderson (Eds.), *Handbook of survey research*. San Diego: Academic Press.

Grove, R.M. (1989). *Survey errors and survey costs*. New York: Wiley.

Henderson, M.E., Morris, L.L., & Fitz-Gibbon, C.T. (1987). *How to measure attitudes*. Newbury Park, CA: Sage.

Morris, L.L., Fitz-Gibbon, C.T., & Lindheim, E. (1987). *How to measure performance and use tests*. Newbury Park, CA: Sage.

Nunnally, J.C., (1978). *Psychometric theory*. 2nd ed. New York: McGraw-Hill.

Pedhazur, E.J., & Schmelkin, L.P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

Posavac, E.J., & Carey, R.G. (1985). *Program evaluation: Methods and case studies*. Englewood Cliffs, NJ: Prentice-Hall.

Satin, A., & Shastry, W. (1988). *Survey sampling: A non-mathematical guide*. Ottawa: Supply & Services.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments in question form, wording and context*. New York: Academic Press.

Selltiz, C., Wrightsman, L.S., & Cook, S.W. (1976). *Research methods in social relations*. 3rd ed. New York: Holt, Reinhart & Winston.

Stanley, J.C. (1971). Reliability. In R.L. Thorndike (Ed.), *Educational measurement*. 2nd ed. Washington: American Council on Education.

Weisberg, H.F. (1992). *Central tendency and variability*. (Sage University Paper Series on Quantitative Applications in the Social Sciences 07–083). Newbury Park, CA: Sage.