1. Introduction Definition of Measurement

The notion that measurement is crucial to science seems a commonplace and unexceptional observation. Most book-length treatments of the philosophy of science include a discussion of the topic. And books focusing on research methods invariably have a chapter dealing with the problems associated with measurement. Yet, the widespread acknowledgement of the importance of good measurement has not—until quite recently —led to the development of systematic and general approaches to measurement in the social sciences. Quite the contrary, historically, measurement has been more of an abstract, almost ritualistic concern instead of being an integral and central aspect of the social sciences. The coexistence of this asymmetric condition of ritualistic concern but lack of systematic attention with regard to measurement may be partially attributable to the way in which this term is most commonly defined. The most popular definition of measurement is that provided by Stevens more than 25 years ago. "Measurement," Stevens wrote, "is the assignment of numbers to objects or events according to rules" (1951: 22). The problem with this definition, from the point of view of the social scientist, is that, strictly speaking, many of the phenomena to be masured are neither objects nor events. Rather, the phenomena to be measured are typically too abstract to be adequately characterized as either objects or events.

Thus, for example, phenomena such as political efficacy, alienation, gross national product, and cognitive dissonance are too abstract to be considered "things that can be seen or touched" (the definition of an object) or merely as a "result, consequence, or outcome" (the definition of an event). In other words, Stevens's classical definition of measurement is much more appropriate for the physical than the social sciences. Indeed, it may have inadvertently impeded efforts to focus systematically on measurement in social research. 1 A definition of measurement that is more relevant to the social sciences is that suggested by Blalock's observation that: Sociological theorists often use concepts that are formulated at rather high levels of abstraction. These are quite different from the variables that are the stock-in-trade of empirical sociologists. … The problem of bridging the gap between theory and research is then seen as one of measurement error [1968: 6; 12]. In other words, measurement is most usefully viewed as the "process of linking abstract concepts to empirical indicants" (Zeller and Carmines, forthcoming), as a process involving an "explicit, organized plan for classifying (and often quantifying) the particular sense data at hand—the indicants—in terms of the general concept in the researcher's mind" (Riley, 1963: 23). This definition makes it clear that measurement is a process involving both theoretical as well as empirical considerations. From an empirical standpoint, the focus is on the observable response— whether it takes the form of a mark on a self-administered questionnaire, the behavior recorded in an observational study, or the answer given to an interviewer. Theoretically, interest lies in the underlying unobservable (and directly unmeasurable) concept that is represented by the response. Thus, using the above examples, the "mark" may represent one's level of self-esteem, the "behavior" may indicate one's level of personal integration during a conflict situation, and the "answer" may signify one's attitude toward President Carter. Measurement focuses on the crucial relationship between the empirically grounded indicator(s)—that is, the observable response—and the underlying unobservable concept(s). When this

4. Assessing Reliability

In this chapter we discuss the four basic methods for estimating the reliability of empirical measurements. These are the retest method, the alternative-form method, the split-halves method, and the internal consistency method. This chapter also discusses how reliability estimates can be used to "correct" correlations for unreliability due to random measurement error. Finally, we briefly evaluate the strengths and weaknesses various methods for assessing reliability.

Retest Method

One of the easiest ways to estimate the reliability of empirical measurements is by the retest method in which the same test is given to the same people after a period of time. 14 One then obtains the correlation between scores on the two administrations of the same test. The retest method is diagramed in Figure 1. It is presumed that responses to the test will correlate across time because they reflect the same true variable, t. The equations for the two tests may be written as follows:

But recalling that the definition of parallel measurements specifies that t = t and s e12 = se22* and that by the assumptions of classical test theory r(e1,t2) = 0, and r (e1,e2) = 0, it can be shown that following exactly the same logic used to show that the correlation between parallel measures equals the reliability coefficient (see the derivation of Equation 10 above). That is, the reliability is equal to the correlation between the scores on the same test obtained at two points in time.

Figure 1: A Schematic Representation of the Retest Method for Estimating Reliability

one obtains exactly the same results on the two administrations of the test, then the retest reliability coefficient will be 1.00. But, invariably, the correlation of measurements across time will be less than perfect. This occurs because of the instability of measures taken at multiple points in time. For example, a person may respond differently to a set of indicators used to measure self -esteem from one time to another because "the respondent may be temporarily distracted, misunderstand the meaning of an item," feel uncomfortable due to someone else being present, and so forth (Bohrnstedt, 1970: 85). All of these conditions reduce the reliability of empirical measurements.

While test-retest correlations represent an intuitively appealing procedure by which to assess reliability, they are not without serious problems and limitations. Perhaps most important, researchers are often only able to obtain a measure of a phenomenon at a single point in time. Not only can it be unduly expensive to obtain measurements at multiple points in time but it can be impractical as well. Even if test-retest correlations can be computed, their interpretation is not necessarily straightforward. A low test-retest correlation may not indicate that the reliability of the test is low but may, instead, signify that the underlying theoretical concept itself has changed. For example, one's attitude toward capital punishment may be very different before and after the person has viewed an execution. But true change is interpreted as measurement instability in the assessment of retest reliability. Moreover, the longer the time interval between measurements, the more likely that the concept has changed. In other words, a naive interpretation of test-retest correlations can drastically underestimate the degree of reliability in measurements over time by interpreting true change as measurement instability. 15 A second problem that affects test-retest correlations and also leads to deflated reliability estimates is reactivity. Reactivity refers to the fact that sometimes the very process of measuring a phenomenon can induce change in the phenomenon itself. Thus, in measuring a person's

attitude at time 1, the person can be sensitized to the subject under investigation and demonstrate a change at time 2, which is due solely to the earlier measurement. For example, if a person is interviewed about the likelihood of voting in an approaching election at time 1, the person might decide to vote (at time 2) and cast a ballot (at time 3) merely because he or she has been sensitized to the election. In this case, the test-retest correlation will be lower than it would be otherwise because of reactivity. While the test-retest correlations can certainly underestimate the reliability of empirical measurements, the more typical problem is overestimation due to memory. For example, the person's memory of his responses during the first interview situation is quite likely to influence the responses which he gives in the second interview. In other words, if the time interval between measurements is relatively short, the subjects will remember their earliest responses and will appear more consistent than they actually are. Memory effects lead to inflated reliability estimates. In fact, Nunally believes that "during the two-week's to one-month's time in which it is advisable to complete both testings, memory is likely to be a strong factor, thus, the retest method will often provide a substantial overestimate of what would be obtained from the alternative-form method" (1964: 85). Alternative-Form Method The alternative-form method is used extensively in education to estimate the reliability of all types of tests. In some ways, it is similar to the retest method in that it also requires two testing situations with the same people. However, it differs from the retest method in one very important regard: The same test is not given on the second testing but an alternative form of the same test is administered. These two forms of the test are intended to measure the same thing. Thus, for example, the two tests might focus on arithmetical operations with each containing 25 problems that are at approximately the same level of difficulty. Indeed, the two forms should not differ from each other in any systematic way. One way to help insure this is to use random procedures to select items for the different forms of the test. The correlation between the alternative forms provides the estimate of reliability. It is recommended that the two forms be administered about two weeks apart, thus allowing for day - to-day fluctuations in the person to occur (Nunnally, 1964). The alternative-form method for assessing reliability is obviously superior to the simple retest method, primarily because it reduces the extent to which individuals' memory can inflate the reliability estimate. However, like the retest method, the alternative-form method when used for only two testing administrations does not allow one to distinguish true change from unreliability of the measure. For this reason, the results of alternative-form reliability studies are easier to interpret if the phenomenon being measured is relatively enduring, as opposed to being subject to rapid and radical alteration. The basic limitation of the alternative-form method of assessing reliability is the practical difficulty of constructing alternative forms that are parallel. It is often difficult to construct one form of a test much less two forms that display the properties of parallel measurements.

Split-Halves Method

Both the retest and the alternative-form methods for assessing reliability require two test administrations with the same group of people. In contrast, the split-halves method can be conducted on one occasion. Specifically, the total set of items is divided into halves and the scores on the halves are correlated to obtain an estimate of reliability. The halves can be considered approximations to alternative forms. As a practical example, let us assume that a teacher has administered a six-word spelling test to his students and would like to

determine the reliability of the total test. He should divide the test into halves, determine the number of words that each student has spelled correctly in each half, and obtain the correlation between these scores. But as we have determined previously, this correlation would be the reliability for each half of the test rather than the total test. Therefore, a statistical correction must be made so that the teacher can obtain an estimate of the reliability of the six-word test, not just the three-word half tests. This "statistical correction" is known as the Spearman-Brown prophecy formula, derived independently by Spearman (1910) and Brown (1910). In particular, since the total test is twice as long as each half, the appropriate Spearman-Brown prophecy formula is: where r xx" is the reliability coefficient for the whole test and r xx' is the split-half correlation. Thus, if the correlation between the halves is .75, the reliability for the total test is: r xx" = [(2) (.75)]/(1 + .75) = 1.50/1.75 = .857.

Internal Consistency Method

We noted above that an important limitation of the split-halves method of assessing reliability is that reliability coefficients obtained from different ways of subdividing the total set of items would not be the same. For example, it is quite possible that the correlation between the first and second halves of the test would be different from the correlation between odd and even items. However, there are methods for estimating reliability that do not require either the splitting or repeating of items. Instead, these techniques require only a single test administration and provide a unique estimate of reliability for the given test administration. As a group, these coefficients are referred to as measures of internal consistency. By far the most popular of these reliability estimates is given by Cronbach's alpha (Cronbach, 1951), which can be expressed as follows:

where N is equal to the number of items; Ss 2 (Y i ) is equal to the sum of item variances; and s 2x * is equal to the variance of the total composite. If one is working with the correlation matrix rather than the variance-covariance matrix, then alpha reduces to the following expression: where N is again equal to the number of items and p* is equal to the mean interitem correlation. To take a hypothetical example applying Equation 20, if the average intercorrelation of a six-item scale is .5, then the alpha for the scale would be: To give an example of how alpha is calculated, consider the 10-item self -esteem scale developed by Rosenberg (1965). The intercorrelations among the items for a sample of adolescents are presented in Table 3 (for further discussion of these data see the appendix). To find the mean interitem correlation we first sum the 45 correlations in Table 3: .185 + .451 + .048 + … + .233 = 14.487. Then we divide this sum by 45: 14.487/45 = .32. Now we use this mean interitem correlation of .32 to calculate alpha as follows: From Equation 20 it is not difficult to see that alpha varies between .00 and 1.00, taking on these limits when the average interitem correlations are zero and unity, respectively. The interpretation of Cronbach's alpha is closely related to that given for reliability estimates based on the split- halves method. Specifically, coefficient alpha for a test having 2N items is equal to the average value of the alpha coefficients obtained for all possible combinations of items into two half-tests (Novick and Lewis, 1967). Alternatively, alpha can be considered a unique estimate of the expected correlation of one test with an alternative form containing the same number of items. Nunnally (1978) has demonstrated that coefficient alpha can also be derived as the expected correlation between an actual test and a hypothetical alternative form of the same length, one that

may never be constructed. Novick and Lewis (1967) have proven that, in general, alpha is a lower bound to the reliability of an unweighted scale of N items, that is, r x ³ a. It is equal to the reliability if the items are parallel. Thus, the reliability of a scale can never be lower than alpha even if the items depart substantially from being parallel measurements. In other words, in most situations, alpha provides a conservative estimate of a measure's reliability. Equation 20 also makes clear that the value of alpha depends on the average interitem correlation and the number of items in the scale. Specifically, as the average correlation among items increases and as the number of items increases, the value of alpha increases. This can be seen by examining Table 1 which shows the value of alpha given a range in the number of items from 2 to 10 and a range in the average interitem correlation from .0 to 1.0. For example,

KR20 Cronbach's alpha is a generalization of a coefficient introduced by Kuder and Richardson (1937) to estimate the reliability of scales composed of dichotomously — scored items. Dichotomous items are scored one or zero depending on whether the respondent does or does not possess the particular characteristic under investigation. Thus, for the items making up a spelling test, a score of 1 would be given when the students spelled a particular word correctly but zero if the word is spelled incorrectly. To determine the reliability of scales composed of dichotomously scored items, one uses the following Kuder-Richardson formula number 20 (symbolized KR20): where N is the number of dichotomous items; p i is the proportion responding "positively" to the i th item; q i is equal to 1 - p i ; and s 2x * is equal to the variance of the total composite. Since KR20 is simply a special case of alpha, it has the same interpretation as alpha; that is, it is an estimate of the expected correlation between one test and a hypothetical alternative form containing the same number of items. Correction for Attenuation Whatever particular method is used to obtain an estimate of reliability, one of its important uses is to "correct" correlations for unreliability due to random measurement error. That is, if we can estimate the reliability of each variable, then we can use these estimates to determine what the correlation between the two variables would be if they were made perfectly reliable. The appropriate formula is as follows: where r xtyt is the correlation corrected for attenuation; r xiyj is the observed correlation; r xx' is the reliability of X; and r yy' is the reliability of Y. For example, if the observed correlation between twovariables was .2 and the reliability of each variable was .5, then the correlation corrected for attentuation would be: This means that the correlation between these two variables would be .4 if both were perfectly reliable (measured without random error). Table 2 illustrates the behavior of the correlation coefficient under varying conditions of correction for attenuation. Table 2A shows the value of the correlation corrected for attenuation given that the observed correlation is .3 with varying reliabilities of X and Y. As an example, when the reliabilities of X and Y are .4, respectively, the corrected correlation is .75. When the reliabilities of X and Y are 1.0, respectively, the corrected correlation is equal to the observed correlation of .3. Table 2B presents similar calculations when the observed correlation is .5. Examining sections A and B of Table 2 it is clear that the higher the reliabilities of the variables, the less the corrected correlation differs from the observed correlation. Table 2C presents the value of the correlation that one will observe when the correlation between X t and Y t is .5 under varying conditions of reliability. If the reliabilities of X and Y are .8, respectively, the observed value of a

theoretical .5 correlation is .4. Table 2D presents similar calculations when the correlation between X t and Y t is .7. For example, even if the theoretical correlation between X t and Y t is .7, the observed correlation will be only .14 if the reliabilities are quite low (.2). Thus, one must be careful not to conclude that the theoretical correlations are low simply because their observed counterparts are low; it may instead be the case that the measures are quite unreliable. Conclusion This chapter has discussed four methods for assessing the reliability of empirical measurements. For reasons mentioned in the chapter, neither the retest method nor the split-halves approach is recommended for estimating reliability. The major defect of the retest method is that experience in the first testing usually will influence responses in the second testing. The major problem with the split-halves approach is that the correlation between the halves will differ somewhat depending on how the total number of items is divided into halves. As Nunnally argues, "it is best to think of the corrected correlation between any two halves of a test as being an estimate of coefficient alpha. Then it is much more sensible to employ coefficient alpha than any split-half method" (1978: 233). In contrast, the alternative-form method and coefficient alpha provide excellent techniques for assessing reliability. The practical limitation of using the alternative-form method is that it can be quite difficult to construct alternative forms of a test that are parallel. One recommended way of overcoming this limitation is by randomly dividing a large collection of items in half to form two randomly parallel tests. In sum, if it is possible to have two test administrations, then the correlation between alternative forms of the same test provides a very useful way to assess reliability. Coefficient alpha should be computed for any multiple-item scale. It is particularly easy to use because it requires only a single test administration. Moreover, it is a very general reliability coefficient, encompassing both the Spearman-Brown prophecy formula as well as the Kuder-Richardson 20. Finally, as we have seen, alpha is easy to compute, especially if one is working with a correlation matrix (for further details on the computation of alpha see Bohrnstedt, 1969). The minimal effort that is required to compute alpha is more than repaid by the substantial information that it conveys about the reliability of a scale. What is a satisfactory level of reliability? Unfortunately, it is difficult to specify a single level that should apply in all situations. As a general rule, we believe that reliabilities should not be below .80 for widely used scales. At that level, correlations are attenuated very little by random measurement error. At the same time, it is often too costly in terms of time and money to try to obtain a higher reliability coefficient. But the most important thing to remember is to report the reliability of the scale and how it was calculated. Then other researchers can determine for themselves whether it is adequate for any particular purpose.

TABLE 2 Examples of Correction for Attenuation