Psychological Bulletin
1979, Vol. 86, No. 2, 376-390

Interobserver Agreement, Reliability, and Generalizability of Data Collected in Observational Studies

Sandra K. Mitchell
Department of Maternal and Child Nursing
University of Washington

Research in developmental and educational psychology has come to rely less on conventional psychometric tests and more on records of behavior made by human observers in natural and quasi-natural settings. Three coefficients that purport to reflect the quality of data collected in these observational studies are discussed: the interobserver agreement percentage, the reliability coefficient, and the generalizability coefficient. It is concluded that although high interobserver agreement is desirable in observational studies, high agreement alone is not sufficient to insure the quality of the data that are collected. Evidence of the reliability or generalizability of the data should also be reported. Further advantages of generalizability designs are discussed.

Almost everyone engaged in research recognizes the need for reliable measuring instruments. Reliability is a central topic particularly for courses and textbooks concerned with the behavioral sciences. In spite of varying theoretical derivations, its definition is remarkably uniform: A reliable instrument is one with small errors of measurement, one that shows stability, consistency, and dependability of scores for individuals on the trait, characteristic, or behavior being assessed.

The preparation of this article was supported in part by Contract NO1-NU-14174 with the Division of Nursing, Bureau of Health Resources and Development, U.S. Public Health Service, Health Resources Administration, Department of Health, Education, and Welfare. The article is based in part on a dissertation submitted in partial fulfillment of the requirements for the PhD degree at the University of Washington.

I would like to thank Halbert B. Robinson and Kathryn E. Barnard for their support of the dissertation research, and Terence R. Mitchell and Nancy E. Jackson for their comments on an earlier draft of the article.

Requests for reprints should be sent to Sandra K. Mitchell, Department of Maternal and Child Nursing, Child Development and Mental Retardation Center, University of Washington, Seattle, Washington 98195.

Historically, the study of reliability has been linked to the study of individual differences and has been largely restricted to standardized tests of intelligence, achievement, and personality. These tests, however, are increasingly being replaced in developmental and educational psychology research by observations of subjects made in natural and quasi-natural settings. Although these observational studies vary widely in content and method, they all use human observers to record (and in some cases to summarize and abstract) the behavior of the subjects. Surprisingly, the reliability of these observational methods has not received the same attention as has the reliability of the more traditional methods (Johnson & Bolstad, 1973).

There are at least three different ways to think about the reliability of observational data. First, the researcher could focus on the extent to which two observers, working independently, agree on what behaviors are occurring. A coefficient that reflects the extent of this agreement has often been used to report reliability in observational studies. Second, the observational measure could be considered a special case of standardized psychological test, and the definitions of reliabil-

ity that come from classic psychometric theory (e.g., test–retest and alternate forms) could be used. Finally, an observational measure could be thought to provide data that are under the influence of a number of different aspects of the observation situation (e.g., different observers or different occasions), including individual differences among subjects. This third viewpoint was developed in Cronbach's (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) theory of generalizability. The purpose of this article is to examine the appropriateness and correct interpretation of these three coefficients when they are used to reflect the quality and dependability of data gathered in observational studies.

Observer Agreement

To insure that data collected by human observers are objective, researchers typically obtain and report coefficients that demonstrate that two or more observers watching the same behavior at the same time will record the same data. These coefficients are offered as the reliability of the instrument being used.

Interobserver Agreement Percentage

The most common index of the quality of the data collected in observational studies is the interobserver agreement percentage. In its simplest form, this coefficient is just what its name implies: the percentage of time units during which the records of two observers are in agreement about the record of behavior.

Two comments can be made about the use of observer agreement percentages. First, the majority of studies report data only in this fashion. Consider, for example, the field of developmental psychology, in which about one third of recently published research articles use observational techniques. In Volume 47 (1976) of *Child Development,* 33 full-length articles reported observational measures. Of these studies, just under half (49%) reported only observer agreement figures as indicants of the quality of their data. Similarly, of 21 observational studies reported in Volume 12 (1976) of *Developmental Psychology,* 57% reported observer agreement percentages only.

Second, the amount of variability among the subjects in a study has very little impact, at least in theory, on the size of an interobserver agreement percentage. In actual practice, however, the degree of variability can make quite a difference: In very homogenous groups, observer agreement percentages are necessarily quite high because all scores given to all subjects are very close together. Thus, a measure that shows high agreement may, in some populations at least, have done a poor job of differentiating among subjects.

Other Problems With Observer Agreement

The interobserver agreement percentage has several important shortcomings. It is, first of all, insensitive to degrees of agreement, that is, it treats agreement as an all-or-none phenomenon, with no room for partial or incomplete agreement. In this sense, the percentage underestimates the actual extent of agreement between two observers. Second, some agreement between independent observers can be expected on the basis of chance alone. For many observational studies, especially those that use frequency counts of individual behaviors, the extent of this chance agreement is dependent on the rates at which the target behaviors occur. Behaviors with very high and very low frequencies can have extremely high chance levels of agreement (Johnson & Bolstad, 1973). In this sense, the percentage appears to overestimate the real agreement.

These difficulties in the use of observer agreement percentages are not unknown, and considerable effort has been expended to develop indices that overcome them. The alternative coefficients have been reviewed in detail elsewhere (Tinsley & Weiss, 1975). In general, they have been designed to overcome the mathematical shortcomings of the agreement percentage, such as chance levels of agreement.

A serious question remains, however, about the utility of these alternatives. In spite of their mathematical superiority, they deal with only one source of error (observer disagreement), and they deal with it without regard

to the magnitude of the individual differences. These alternatives may give a more accurate picture of the level of observer agreement in a study than does the simple percentage, but they do not otherwise describe the stability, consistency, and dependability of the data that have been collected.

Influences on Observer Agreement

Interobserver agreement has been experimentally studied as a phenomenon in its own right. In these experiments, observer accuracy, that is, agreement with a predetermined correct behavioral record, is the dependent variable.

Reid (1970) compared the accuracy of observers during overt and covert assessment of their reliability and found that they were significantly more accurate when they were aware that they were being checked. Romanczyk, Kent, Diament, and O'Leary (1973) found a similar drop from the overt to the covert assessment situation. They also found that observers recorded behavior differently depending on which of several researchers' records they thought their own records would be compared with. Taplin and Reid (1973) found that observer accuracy decreased between the end of a training period and the beginning of data collection and that it increased on days when "spot checks" were expected. Mash and McElwee (1974) reported that observers who had been trained to code "predictable" behavior (i.e., conversations with redundant information) showed a decline in accuracy when they later coded "unpredictable" behavior, whereas observers trained with unpredictable sequences showed no such decline in accuracy. Taken together, these studies imply that differences in experience, mental set, and training of observers can influence the accuracy with which behavioral records are made and scored.

Interobserver agreement has also been used as the dependent variable in studies that compare different methods of observation. McDowell (1973) found comparable observer agreement in time sampling and continuous recording of infant caretaking activities in an institution. Lytton (1973) found the interobserver agreement of ratings, home observations, and laboratory experiments to vary only slightly, but the amounts of time and effort necessary to achieve these levels were quite different. Mash and McElwee's (1974) rating system with only four categories was used more accurately than was an eight-category form. It appears, then, that there are differences in the interobserver agreement that can be expected from different methods of behavioral observation. These differences are reflected both in different levels of agreement for equal amounts of training and inequal levels of agreement for different amounts of training.

It is difficult, however, to put these agreement differences into perspective without knowledge of the overall variability among subjects in the studies. If between-subjects variability is low, then the reported differences in agreement due to observer instructions or observational methods may considerably influence the outcome of the study. If between-subjects variability is high, on the other hand, then these differences will probably have little influence. The studies that used agreement as a dependent variable have identified some problem areas in data collection, but they have not evaluated the relative importance of these problems.

Psychometric Theory of Reliability

The classical method of determining the reliability of a test is for the researcher to obtain two scores for a group of subjects on the test. These two scores may come from two separate scorings of the instrument, from administration of two parts or forms of the instrument to the subjects, or from two administrations of the same instrument to the subjects. The correlation between the two scores is the reliability of the instrument.

The central theoretical concept that underlies this psychometric view of reliability is that every test score is composed of two parts: a *true score,* which reflects the presence or extent of some trait, characteristic, or behavior, plus an *error score,* which is random and independent of the true score (Nunnally, 1967). The proportion of variance accounted

for by each of these parts is estimated from the correlation between the two scores obtained on the instrument.

The variance attributable to individual differences is usually given the same interpretation, regardless of how the two scores used to compute it were obtained. It reflects stable differences among individuals—the true score part of the data. The variance that is attributable to measurement error, however, is subject to varying interpretations, depending on how the two scores were obtained.

The error always includes, of course, the real error—those random fluctuations of the myriad factors that may affect the behavior being measured. These include such variables as the health or mental state of the subjects, the lighting or temperature in the testing room, and so forth. But the error also includes other sources of variation, depending on the method used to obtain the two scores. These other sources include differences within and between scorers, differences between different sections or forms of a test, and changes in subjects' behavior between two administrations of a test.

Thus, the three most common procedures for determining reliability involve (a) obtaining two separate scorings of the same instrument (intrascorer or interscorer reliability), (b) obtaining scores on two parts of the same instrument or on two very similar instruments (split-half or alternate-forms reliability), and (c) obtaining two scores from two separate administrations of the same instrument (test–retest reliability).

The researcher who wishes to use one of these classical methods of determining reliability in an observational study must somehow make his or her observations fit into the same general pattern as psychological tests. However, instead of one test with many items, all intended to measure the same trait, characteristic, or behavior, the observational researcher has a tool with a relatively small number of categories, with each category intended to measure a different trait, characteristic, or behavior. For each of these categories, which are generally mutually exclusive, data are usually collected during many distinct time units.

The most satisfactory way of making such data fit into the classic pattern seems to be to consider each mutually exclusive category (or type of behavior) a separate test with its own reliability. Each time unit is considered to be an item, since all time units are intended to measure the same trait, behavior, or characteristic. For example, a behavioral code might record a child's proximity to the teacher during each 10-sec unit of an observation. Each 10-sec unit would be an item in a test that measured proximity. If the measure consisted of a single summary score (such as in a rating), then there would be, in effect, no individual items at all, just one score.

This analogy between test reliability and the reliability of observational data can be extended to apply to each of the traditional ways of obtaining scores: intrascorer or interscorer reliability, split-half or alternate-forms reliability, and test–retest reliability.

Intrascorer or Interscorer Reliability

A clinical psychologist interested in self-directed aggression might listen twice to tape recordings of patients' responses to a projective test, each time counting the number of self-destructive statements. The correlation between the two counts for the group of patients would be the intrascorer reliability of the self-destructiveness score. The true score implied by this correlation would reflect real differences in self-destructiveness among the patients. The error would include not only the random error but also any inconsistencies in the psychologist's use of the self-destructiveness scale.

In actual practice, it is more likely that two psychologists would listen to the tape recordings. The correlation between their separate counts would be an interscorer reliability coefficient. The true score would again reflect real differences, but the error would reflect differences between the psychologists in their use of the scale, along with random error.

A similar situation exists, of course, when two or more observers record the behavior of subjects in other natural and quasi-natural settings. The correlation between the scores of two observers who kept track of how much

individual attention each child received from the teacher would be an interobserver reliability coefficient. This coefficient, once again, should not be confused with the observer agreement percentage.

Split-Half or Alternate-Forms Reliability

One way of determining the reliability of a standardized test is to compare scores on two subdivisions of the test (odd- and even-numbered items, frequently) or scores on two very similar versions of the test. In an observational study, the corresponding comparisons would be between subdivisions of one observation (e.g., odd- and even-numbered minutes during a tennis lesson), or between two very similar observations (first and second halves of a lesson, perhaps). This is an example of how time units can be considered analogous to test items.

Just as with interobserver reliability, the true score component of the variance in split-half or alternate-forms reliability reflects consistent individual differences among subjects. The error component, however, has a different interpretation. Along with random fluctuations in the behavior of the subjects, real differences in subject behavior between the two observed subdivisions are included as part of the error.

Test–Retest Reliability

Perhaps the most straightforward way to obtain two scores in a reliability study is to administer the same instrument at two different times. An observer might, for example, visit classrooms on different days to record the teacher's use of a particular instructional technique. As before, the true score is assumed to reflect some stable trait, characteristic, or behavior. In this case, the error includes not only random fluctuations of subject behavior but also whatever real changes in subject behavior have occurred between the two administrations of the test.

It is interesting to note that there is little difference between alternate-forms and test–retest reliability for observational measures. Since time units serve as items, observations made on different days can be considered either as alternate-forms or as test–retest conditions, depending on the situation.

Use of Reliability Coefficients

Three comments apply to all of these versions of the reliability coefficient. First, although the examples given are from hypothetical observational studies, real observational studies do not make use of all of the possible coefficients. Interobserver reliability or agreement is reported to the virtual exclusion of split-half and test–retest coefficients. Once again, developmental psychology can serve as an example. In Volume 47 (1976) of *Child Development,* 49% of the full-length research articles that used observational methods reported only observer agreement, and 39% more reported interobserver correlation coefficients. Only three of the studies (12%) reported a reliability coefficient that reflected the stability of subject behavior over time, that is, a split-half or test–retest reliability. Similarly, in Volume 12 (1976) of *Developmental Psychology,* 57% of the studies reported agreement only, 38% reported observer reliability coefficients, and only one study used a measure based on more than one sample of behavior per subject.

Second, although this discussion has emphasized the sources of error in these coefficients, the variance of true scores is as important in determining the size of the reliability coefficient as the variance of error scores is. Recall that the reliability of a test score can be expressed

$$\frac{\text{true score variance}}{\text{true score variance} + \text{error variance}}.$$

For a given level of error variance, then, an instrument will have a lower reliability when it is used on a homogenous group of subjects (low true score variance) than it will when it is used on a more heterogenous group (high true score variance). For instance, if the error variance is 10 and the true score variance is also 10, the reliability of the instrument is 10/20 or .50. But if the error variance is 10 and the true score variance is 40, the reliability of the instrument is 40/50 or .80. This is

in contrast to the observer agreement percentage, which is highest for homogenous groups of subjects.

Third, it should be repeated that reliability and observer agreement are not the same. It is possible, as illustrated by Tinsley and Weiss (1975), to have high interobserver agreement and a low reliability (correlation) coefficient, and vice versa. For instance, two observers might have perfect agreement about the color of shoes worn by children in a classroom, but if all the children wore red shoes, shoe color would not differentiate among the children. On the other hand, there might be a high correlation between two observers' records of the duration of a teacher's attention to a partieular youngster, but if one observer's watch ran slower than the other's watch, they would probably never agree on the actual duration of the attention.

The differences between agreement and reliability are based on the way the two indices are defined. Reliability coefficients partition the variance of a set of scores into a true score (individual differences) and an error component. The error component may include random fluctuations in the behavior of subjects, inconsistencies in the use of the scale, differences among observers, and so forth. Interobserver agreement percentages, on the other hand, carry no information at all about individual differences among subjects and contain information about only one of the possible sources of error—differences among observers. In other words, a reliability coefficient reflects the relative magnitude of all error with respect to true score variability, whereas an agreement percentage reflects the absolute magnitude of just one kind of error.

All in all, there is no perfect reliability coefficient, nor is there one that is even generally best. Coefficients that use two scorings of the same instrument (interobserver and intraobserver reliability) confound random subject error with differences within and between scorers. Coefficients that use scores from subdivisions or alternative forms of the instruments (split-half and alternate-forms reliability) confound random subject error with differences between the subdivisions or forms. Finally, coefficients that use scores from the same instrument administered on two occasions (test–retest reliability) confound measurement errors with real changes in subject behavior that occur between the two administrations. The methods described cannot, then, separately estimate variance in test scores attributable to scorers, subtests (or forms), or occasions, nor can they consider these sources of error simultaneously. A more inclusive, multivariate theory is needed.

Generalizability Theory

Cronbach and his associates (Cronbach et al., 1972) have developed a theory that they call the theory of *generalizability*. (For a brief introduction to the theory, see J. P. Campbell, 1976, pp. 185–222.) Instead of assuming, as does classical test theory, that individual differences constitute the only lawful source of variation in test scores, generalizability theory assumes that there may be a number of sources of variation. These sources of variation other than individual differences are called *facets*. Different scorers, alternate test forms, or administration on different occasions are examples of facets that might be studied. A particular combination of facets makes up the *universe* to which test scores may be generalized.

A generalizability study (G study) is more reminiscent of a factorial study in experimental psychology than of a reliability study. In a G study, the researcher must collect data by systematically sampling conditions from each facet in the universe. For instance, two scorers might each score two alternate forms of a test given on different days to a group of subjects. Using an analysis of variance, it is then possible to independently estimate the contributions of each of the facets—scorers, forms, occasions, as well as subjects—to the overall variation in the set of test scores. Besides looking at the conventional F statistic to establish whether each facet makes a significant contribution to the scores, it is possible to compute what Cronbach calls *variance components*. These variance components reflect the size rather than the statistical significance of the contribution of each facet to the observed scores.

In examining the quality of observational data, though, not only are the absolute sizes of the variance components of interest but the relative sizes of the components are also important. The relative sizes, therefore, are the focus of this discussion.

Recall that a reliability coefficient reflects the partitioning of variance into true and error components and that the coefficient is the ratio of true score variance to obtained score variance. It represents, in other words, the proportion of the total variance that is accounted for by individual differences. In the same way, generalizability coefficient reflects the partitioning of variance into components that correspond to the facets sampled in the G study. The coefficient itself (an intraclass correlation) combines these components in a ratio that also represents the proportion of variance attributable to individual differences for a particular universe (set of conditions).

It should not be assumed, however, that a G study generates one coefficient that is appropriate for all applications of the instrument. On the contrary, one G study can generate several coefficients, each corresponding to a different universe of conditions. This fact points to an important distinction between the psychometric theory of reliability discussed earlier and the theory of generalizability. In psychometric theory, conditions of testing (or otherwise obtaining data) are assumed to influence only measurement error, not the true score on the instrument. In generalizability theory, on the other hand, the conditions of testing are assumed to influence the score itself. What Cronbach and his colleagues have shown is that while true scores all contain a common component, they also contain additional different components depending on the design; that is, it is not just the error variance that differs among the several reliability coefficients. This relationship can be illustrated by returning to the earlier example of interscorer reliability—two psychologists counting self-destructive statements from tape recordings. In generalizability terms, this is a one-facet study, that is, it samples observations from one facet (in this case, scorers) in addition to observations of different subjects. Analyzed as a G study, one can estimate variance components for subjects, for scorers, and for the interaction of subjects and scorers (which in this case is confounded with the residual error). The generalizability coefficient from this G study would reflect the dependability of a score for a subject generalized over scorers. In other words, it would indicate the proportion of variance accounted for by individual differences in subjects, above and beyond any effects accounted for by differences between scorers.

Suppose that a second facet—occasions—were added to this study, so that each scorer would count self-destructive statements for each tape two times. This study combines aspects of the interscorer and the intrascorer reliability studies. The analysis of this two-facet G study would yield variance components for subjects, for scorers, and for occasions (which in this case would be interpreted as intrascorer change). Further, variance components could be computed for each of the possible interactions of these facets: Subjects $\times$ Scorers, Subjects $\times$ Occasions, Scorers $\times$ Occasions, and Subjects $\times$ Scorers $\times$ Occasions (which in this case is confounded with residual error). The generalizability coefficient from this study would reflect the proportion of variance accounted for by individual differences in subjects apart from the effects between and within scorers.

Suppose further that this study were extended to three facets by having each psychologist use two different scoring methods for each tape recording he or she listened to (perhaps a count of self-destructive statements and a global rating of self-destructiveness). Observational methods would then be a facet in the universe of generalization.

Clearly each facet that is added to the study makes the information available from the analysis more complete. But there is a significant cost for the extra information provided by each facet: The number of observations required of each subject is multiplied by the number of conditions sampled in the facet. In the present example, the one-facet study would have two scores, the two-facet study would have four scores, and the three-

facet study would have eight scores for each subject.

Described below are some three-facet G studies that parallel the intraobserver–interobserver, split-half, and test–retest reliability studies that were discussed earlier. All of these G studies use the same three facets (observers, observational methods, and occasions) sampled for all subjects. The studies differ in their definitions of *occasions* of measurement and thus in their interpretations of the resulting coefficients. The relationships among the reliability studies and the proposed G studies are summarized in Table 1.

Table 1
Correspondence Between Reliability Studies and Generalizability Studies

Measurement occasions	Reliability study	Generalizability study
Separate scorings of the same behavior or instrument	Intrascorer or interscorer	Duplicate
Scores on two subdivisions of an instrument or behavior sample, or two very similar instruments or behavior samples	Split half or alternate forms	Session
Scores from separate administrations of the same instrument or separate samples of behavior	Test–retest	Developmental

Duplicate Generalizability

One study with this basic design might use audiotaped or videotaped recordings of behavior, which would be scored on more than one occasion by more than one observer, using more than one form of an observational instrument. In this G study (which I call *duplicate generalizability*), the occasions of observation actually consist of exactly the same behavior by the subjects. It is, then, an extension of the traditional intrascorer or interscorer reliability study. A reliability study uses two scores for each subject (usually from two different scorers) and confounds measurement error with differences within and between scorers. A duplicate G study, on the other hand, has many scores for each subject and separately estimates the contribution of differences within and among observers. Although *occasions* is a facet in this G study, variance attributable to occasions cannot be interpreted as within-subject change, since the same behavior occurs on each occasion of observation. In this study, occasions variance should be interpreted as a measure of within-observer stability. A duplicate G study would be appropriate for demonstrating the dependability of an instrument used in a study in which the stability of the behavior over time was not an issue.

Session Generalizability

A second G study with the basic design might be called *session generalizability*. It would use as measurement occasions two subdivisions of some behavioral sequence (i.e., first and second halves or odd and even minutes) and would be an extension of the traditional split-half reliability study. In a split-half reliability study, recall that errors of measurement are confounded with differences between the two halves of the test or observation. In a session G study, differences between subdivisions of the behavioral sequence are estimated by the variance component for occasions. A session G study would be used to estimate the dependability of scores reflecting traits and behaviors expected to be stable during the course of the behavioral sequence being observed, although perhaps no longer than that.

Developmental Generalizability

A third G study with the same basic design might be called *developmental generalizability*. It would use as measurement occasions two or more administrations of the same instrument, perhaps at different ages or developmental stages. It is, then, an extension of the

traditional test–retest reliability study, which confounds measurement error with true changes in behavior that have occurred between the two administrations of a test. In this developmental G study, these changes in behavior over time would be estimated by the occasions facet of the design. The developmental G study is best suited to measure of traits or characteristics believed to be relatively enduring.

Use of Generalizability Coefficients

The comments made about the reliability studies discussed earlier can be repeated for these G studies. First, observational studies very rarely report data in G-study terms. The method does appear occasionally in dissertations (see, e.g., Leler, 1971; Mitchell, 1977), and it surfaces now and then in educational psychology research (Medley & Mitzel, 1963; McGaw, Wardrop, & Bunda, 1972). But to return again to developmental psychology as an example, there were no studies published in 1976 in either *Child Development* or *Developmental Psychology* that reported generalizability coefficients. Second, regardless of the sizes of the variance components for the facets, it is necessary to have a relatively large variance component for subjects to obtain a large generalizability coefficient. All other things being equal, a sample of subjects with greater variability on the trait being measured will yield a higher generalizability coefficient than will a sample of subjects with lesser variability on the trait.

The three G-study designs outlined here (duplicate, session, and developmental) all sample three important facets that may influence scores in observational studies: observers, observational methods, and occasions of observation. They differ in the nature of the occasions that are sampled, and these occasions tell us something about the nature of the universe to which scores can be generalized.

One way to contrast these universes is to imagine that the three types of studies were conducted so that exactly the same subjects, observers, and observational methods were used for all three. When only the nature of the occasions facet is different, one can hypothesize certain relationships among the sizes of the coefficients derived from the three studies. When duplicate G studies are conducted, it is expected that variance due to occasions will be the smallest and hence that the generalizability coefficient will be the largest. Further, when developmental G studies are conducted, it is expected that variance due to occasions will be the greatest and that the generalizability coefficients will be the smallest. Finally, when session G studies are conducted it is expected that the variance due to occasions and the resulting generalizability coefficients will be intermediate.

Other Uses for Generalizability Theory

Although measures of interobserver agreement and reliability have important uses in observational research, it should be clear from the earlier discussion that it is the generalizability coefficient that is potentially the most useful source of information about the quality of such data. Generalizability theory, however, has many applications of interest for the developmental psychologist other than the computation of a coefficient. Some of these applications are discussed below.

Multitrait–Multimethod Matrix

D. T. Campbell and Fiske (1959) introduced the notion of determining the validity of psychological measurement instruments by using a *multitrait–multimethod matrix.* As the name suggests, this matrix consists of scores for an individual on several traits, each trait assessed by two or more different methods. Such a design is clearly an instance of a G study in which traits and methods are the two facets. The data analysis proposed by Campbell and Fiske uses a matrix of correlations, but other authors (i.e., Kavanagh, MacKinney, & Wolins, 1971) have used analyses of variance with the multitrait–multimethod design. In this form, such a matrix closely resembles a G study.

The multitrait–multimethod matrix was used by Wicker (1975) to examine the reliability of observational records generated

from transcriptions of conversations. In this study observers were treated as "methods," and behavior samples were treated as "traits." By applying Campbell and Fiske's criteria to the correlational matrix, Wicker concluded that his data showed both convergent and discriminant validity.

Attribution of Variance

Traditionally, psychological studies have sought either to demonstrate mean differences between groups of subjects or to show consistent individual differences among subjects. Another kind of study, far less common, tries to systematically apportion the variance in a set of research data among several independent variables. This approach has been popular in efforts to resolve the issue of whether individual differences (personality) or situational differences (environment) are the most important determinants of human behavior. In such studies, data are gathered on a group of individuals in several situations. The relative importance of individual differences and of situational differences is estimated by the use of the statistic known as omega-square. As Golding (1975) has ably demonstrated, this experimental design is more profitably viewed as a G study; generalizability coefficients answer questions about the relative importance of different facets (here, individuals and situations) more suitably than does omega-square.

The logic involved in this kind of study is not limited to the person–situation controversy, of course. A similar question might be asked about a study in which several raters rate the behavior of a number of subjects. As Norman and Goldberg (1966) pointed out, these data can be interpreted as reflecting the behavior of the ratees (subjects) or the behavior of the raters. Once again, this study appears to be a straightforward, one-facet generalizability study in which raters is the facet. Although Norman and Goldberg did not use analysis of variance to analyze their data either, it is clear that the conceptualization is similar to that presented earlier: There are several sources of meaningful variation in a set of data, and only a multifacet study can illuminate the relative contributions of different facets.

Observer Generalizability

The use of generalizability coefficients to estimate the contributions of facets other than individual differences to a set of test scores has a particular application to observational studies. Specifically, it allows a researcher to look at the proportion of variance in scores that is attributable to the consistent behavior of the observers.

On the surface, the function of the generalizability coefficient sounds much like the function of the interobserver agreement percentage, but in fact it is not. Recall that the agreement percentage did not take into account the extent of overall variability in a set of data, whereas a generalizability coefficient does. It should be noted that the G study necessary to compute this coefficient is exactly the same study as that used to compute the more conventional coefficient based on the behavior of the subjects. Nothing has changed except one's point of reference.

Mitchell (1977) computed both subject and observer generalizability coefficients in a study in which 67 observers made repeated observations of 10 mother–infant pairs during the first year of life. She found that the coefficients reflecting the variability accounted for by observers were, in this study at least, greater than the coefficients reflecting subject differences. Although this result is specific to this particular set of data, the study is an example of the usefulness of observer generalizability coefficients.

Single-Subject Studies

Studies of individual subjects have had few ways of reporting reliability in the traditional sense. However, it is possible to conduct G studies using only a single subject. For example, several observers, several occasions, or several methods of observation might be used with a single subject. In this case, the generalizability coefficient would reflect the generalizability of a score recorded by a single observer under the circumstances sampled in the

study; that is, it would be an observer generalizability coefficient.

Such single-subject studies may be appropriate even when many subjects are part of a research project. It is commonly assumed that the behavior of all subjects is recorded with equal accuracy, that is, measurement errors are approximately equal for all subjects. If this assumption is true, then the data for all subjects are presumably equally "good." On the other hand, if this assumption is incorrect, so that subject behavior is recorded with variable accuracy, then data for different subjects may have different meanings.

The possibility of systematic differences in measurement error among subjects has been explored for traditional psychological tests (Ghiselli, 1963). Berdie (1969), for example, found that intraindividual variability (i.e., measurement error) was a stable trait for some pencil-and-paper performance tests. Ghiselli (1960) examined the prediction of "predictability." He was able to predict the errors of measurement for two groups of subjects on a reaction time test. Reliability for the high group was .97, compared with .82 for the low group.

A similar result was found in a quite different study by Gorsuch, Henighan, and Barnard (1972). Their interest was the internal consistency of a children's pencil-and-paper test of locus of control. They found that the reliability of the scale differed significantly according to the reading ability of the children. The errors of measurement were quite small for the good readers, but were large for the poor ones.

Observational studies, however, rarely have enough subjects to permit analysis of this kind. An alternative is to make use of observer generalizability coefficients. Suppose that the data collected for each subject were considered to be a mini-G-study. If the basic G-study design outlined earlier were used, each mini-G-study would have two facets (methods and occasions) that would be sampled for each observer. From each of these mini-studies it would be possible to compute an observer generalizability coefficient. Using this design, Mitchell (1977) found that although there were differences in observer agreement for different subjects, subjects did not differ in their observer generalizability coefficients.

Summary

Three different coefficients that purport to reflect the quality of data gathered in observational studies have been discussed. The first and most commonly used of these was observer agreement. Coefficients of observer agreement are a source of important information about the quality of observational data: the objectivity of different observers using the same method to record the same behavior. Determination of interobserver agreement is a necessary part of the development and use of observational measures. Interobserver agreement is not, however, sufficient by itself.

The second coefficient discussed was the reliability coefficient, obtained by fitting observational data into the pattern used by developers of standardized psychological tests. There are really many different reliability coefficients, each defined by the way the scores are obtained for its computation. Reliability coefficients provide useful information about the stability and consistency of individual differences among subjects, but confound measurement error with other sources of variability.

The third coefficient was the generalizability coefficient, as defined by Cronbach et al.'s (1972) multivariate theory. In one sense, a generalizability coefficient supersedes a reliability coefficient, because it too provides information about the stability and consistency of individual differences among subjects. Its superiority to the reliability coefficient lies in its ability to account for variance from sources other than individual differences and measurement error. Besides giving information that can be reported as the generalizability coefficient, a G study also permits innovative ways of looking at results from observational studies. These ways include variations on the multitrait–multimethod matrix, the attribution of variance to independent variables, observer generalizability, and studies of single subjects. It is therefore especially unfortunate that G-study designs are not used more frequently in observational research.

Recommendations

Researchers doing observational studies are obliged to show that their measuring instruments are reliable—that they have small errors of measurement and that the scores of individuals show stability, consistency, and dependability for the trait, characteristic, or behavior being studied. The reason for this obligation is practical: If the measure is not reliable, it cannot be expected to show lawful relationships with other variables being studied. It is well-known that the reliability of a standardized test sets the limits of its validity (Nunnally, 1967). Similarly, the predictive usefulness of observational measures is limited by the stability and consistency of the scores obtained from the observational instruments.

Observer agreement coefficients alone, regardless of their mathematical sophistication, are inadequate to demonstrate this stability and consistency. What alternative or additional information ought to be reported in observational studies, and how should it be collected?

First, and most basically, the coefficients computed should be based on the same scores that are used in the substantive analysis of the study. If a composite score (such as a total of several categories or time units) is to be used for analysis, it is this composite—and not its component individual categories or time units—that should be examined for agreement, reliability, or generalizability. Of course, during the training of observers it is extremely helpful to compare the records of different observers on a trial-by-trial (or time unit by time unit) basis, but such a comparison does not suffice for reporting in a published research article. It is possible, albeit rare, to have acceptable agreement on a time unit basis and yet unacceptable levels of agreement on a total score. It is also possible, and much more common, for observers to be in only moderate agreement for small time units, but to show good agreement for a total score. In this case, an analysis of the trial-by-trial agreement would underestimate the agreement for the measures actually employed in the study.

Similarly, if several different scores are to be analyzed, coefficients should be computed for each of them. Good agreement or reliability on the frequency of particular behaviors, for example, does not insure good agreement or reliability on their duration. The data that are to be analyzed are the data that should be scrutinized for their stability and consistency.

Second, the coefficients should be computed from data that are part of the actual study being reported. The studies of observer agreement cited earlier (i.e., Reid, 1970; Romanczyk et al., 1973; Taplin & Reid, 1973) show clearly that the quality of data collected during a study may not be the same as the quality collected during reliability assessment or training. This difference can also be expected for data collected during a pilot study or during a previous study that used the same instrument.

This means that the researcher must plan to collect data that can be used to compute coefficients of agreement, reliability, or generalizability at the same time the rest of the data are gathered. Although this does entail some additional data collection, the addition need not generally be enormous (see, e.g., Rowley, 1976).

Third, interobserver agreement should always be obtained and reported. In most observational studies, observer disagreement is an important source of error, and it should be carefully and systematically monitored. This monitoring needs to be regular and unobtrusive for the most accurate results. Researchers also need to be alert to the possibility that interobserver agreement may differ for different subjects, therefore observations from many—preferably all—subjects should be used when determining the level of agreement.

Although the observer agreement percentage is the most widely used and most easily computed index of agreement, it may be desirable in some cases to substitute other indices, such as that suggested by Lawlis and Lu (1972). Whatever the exact form of the coefficient, however, both the researcher and the reader should remember that it reflects only one source of error and that it reports this error in absolute rather than in relative terms.

Fourth, a reliability or generalizability

coefficient that uses two or more measurement occasions should be presented for any score that is used to predict other behavior. The researcher is obliged to demonstrate that the individual differences among subjects are stable over different occasions as well as over different observers. This stability can be reported as a split-half, alternate-forms, or test–retest reliability coefficient, or as a session or developmental generalizability coefficient. The single exception to this rule is a study that focuses on some behavior not expected to show stability over time (e.g., first response to a new stimulus). In this case only, an interscorer reliability coefficient or a duplicate generalizability coefficient is appropriate.

It is impossible to overemphasize the importance of using two or more measurement occasions to compute coefficients of reliability or generalizability. The purpose of reporting such a coefficient is to demonstrate that the data being analyzed reflect stable, consistent, and dependable individual differences among subjects. If, however, a single measurement occasion has been used, then the coefficient can demonstrate only the competence and consistency of the observers. Since many, if not most, observational studies obtain repeated measures as part of the experimental design, it is seldom necessary to collect additional data. What is necessary is to analyze and report these observational measures in terms of their stability over time.

Fifth, a generalizability study is usually preferable to the computation of a reliability coefficient. First, a G study provides more useful information about sources of variability in a set of data than does a reliability coefficient. Leler (1971), for instance, used the variance components from a G-study analysis midway through her research to help refine the observation instrument and to retrain observers on some items. Second, the G-study design makes other kinds of analysis possible for most observational studies. These include, particularly, observer generalizability coefficients and coefficients for individual subjects.

Finally, the design of the generalizability study should correspond to the overall design of the research. The G study for most applications does not need to be complex. Two facets—observers and occasions—are usually sufficient. There are some studies, however, that require three-facet designs.

Studies that measure behavior before and after some intervention (experimental treatment) should include a third facet in the G study—before versus after the intervention. This is especially important if the intervention is likely to reduce the variability of the observed behavior. For example, suppose a study were undertaken to reduce the amount of aggressive behavior exhibited by school children on the playground. At the start of the study, the children would be quite variable in their playground aggression. If the intervention were successful, however, the variability after intervention would be quite low (all kids showing low aggression). A measure that would be quite reliable (in the classical sense) in differentiating among the children before the intervention might be inadequate afterwards. The G-study design allows the researcher to evaluate this possibility.

In the same way, studies that compare two groups of subjects should include group membership as a third facet in their G study; that is, it is necessary to demonstrate that the scores for both groups show approximately equal stability and consistency over different occasions and different observers.

Conclusions

These recommendations have an important empirical implication: Studies that follow them will report coefficients that are lower, perhaps substantially lower, than the coefficients reported by studies that do not follow them. The procedures suggested here are stringently conservative, and the coefficients that they yield should be considered lower limits of the true dependability of the observational data that are collected. Reviewers and readers, who are used to seeing reports of observer agreement in the .80s and .90s, will have to change their expectations for reliability and generalizability coefficients, which will often be in the .50s and .60s. In fact, these new coefficients are not low; rather the old ones were inappropriately high. Observer agreement percentages, interobserver reliabil-

ity, and reliability or agreement determined during pretests or previous studies are all spuriously high estimates of the quality of the data that are collected. Although we may have to revise our standards downward concerning the size of reported coefficients, we will be revising our standards upward concerning the ways in which the data for the coefficients are gathered and analyzed.

There is a methodological implication of these recommendations as well. Most observational studies currently being published are pretty straightforward in experimental design. Their sophistication is usually based on the nature of the observations themselves: the complexity of the behavioral record, the length of time included in the record, or the specific nature of the situation or setting in which the data are gathered. Future studies that follow the recommendations given here will be far more complex in design than is now typically the case.

One final question concerns whether it is really worth the additional time and money necessary to determine and report on the quality of the data in the ways suggested here. If including a G-study design in an observational study has no benefits except the computation of a generalizability coefficient, the answer is probably no. In fact, though, including a G study usually provides a great deal of substantive information to the researcher. Are there differences in the behavior reported by different observers? Do subjects act differently on different occasions? Is the variance of the data different before and after an experimental treatment? Are all groups in this study represented by data of equal quality? All these important questions can be addressed by including a G study in the overall research plan.

Even more importantly, though, the use of generalizability designs focuses the attention of the researcher (and the reader) on both the individual differences among subjects and on the influence of other (usually environmental) factors on behavior. As Cronbach (1957) eloquently pointed out, psychologists have historically tended to focus on one or the other of these two aspects: individual differences (using correlational methods) or group differences (using experimental methods). Fittingly, it is now Cronbach's theory of generalizability that makes it possible to combine these two viewpoints. And observational research is particularly well suited to the task of looking at individual differences in behavior in the context of systematic environment variation.

References

Berdie, R. F. Consistency and generalizability of intraindividual variability. *Journal of Applied Psychology,* 1969, *53,* 35–41.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, *56,* 81–105.

Campbell, J. P. Psychometric theory. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology.* Chicago: Rand McNally, 1976.

Cronbach, L. J. The two disciplines of scientific psychology. *American Psychologist,* 1957, *12,* 671–684.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley, 1972.

Ghiselli, E. E. The prediction of predictability. *Educational and Psychological Measurement,* 1960, *20,* 3–8.

Ghiselli, E. E. Moderating effects and differential reliability and validity. *Journal of Applied Psychology,* 1963, *47,* 81–86.

Golding, S. L. Flies in the ointment: Methodological problems in the analysis of the percentage of variance due to persons and situations. *Psychological Bulletin,* 1975, *82,* 278–281.

Gorsuch, R. L., Henighan, R. P., & Barnard, C. Locus of control: An example of dangers in using children's scales with children. *Child Development,* 1972, *43,* 579–590.

Johnson, S. M., & Bolstad, O. D. Methodological issues in naturalistic observations: Some problems and solutions for field research. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior change: Methodology, concepts and practice.* Champaign, Ill.: Research Press, 1973.

Kavanagh, M. J., MacKinney, A. C., & Wolins, L. Issues in managerial performance: Multitrait-multimethod analyses of ratings. *Psychological Bulletin,* 1971, *75,* 34–39.

Lawlis, G. F., & Lu, E. Judgment of counseling process: Reliability, agreement, and error. *Psychological Bulletin,* 1972, *78,* 17–20.

Leler, H. O. Mother-child interaction and language performance in young disadvantaged Negro children (Doctoral dissertation, Stanford University, 1970). *Dissertation Abstracts International,* 1971, *31,* 4971B. (University Microfilms No. 71-2793)

Lytton, H. Three approaches to the study of parent-child interaction: Ethological, interview, and experimental. *Journal of Child Psychology and Psychiatry,* 1973, *14,* 1–17.

Mash, E. J., & McElwee, J. D. Situational effects on observer accuracy: Behavioral predictability, prior experience, and complexity of coding categories. *Child Development,* 1974, *45,* 367–377.

McDowell, E. E. Comparison of time-sampling and continuous recording techniques for observing developmental changes in caretaker and infant behaviors. *Journal of Genetic Psychology,* 1973, *123,* 99–105.

McGaw, B., Wardrop, J. L., & Bunda, M. A. Classroom observational schemes: Where are the errors? *American Educational Research Journal,* 1972, *9,* 13–27.

Medley, D. M., & Mitzel, H. E. Measuring classroom behavior by systematic observation. In L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally, 1963.

Mitchell, S. K. The reliability, generalizability and interobserver agreement of data collected in observational studies (Doctoral dissertation, University of Washington, 1976). *Dissertation Abstracts International,* 1977, *37,* 3583B. (University Microfilms No. 77-611)

Norman, W. T., & Goldberg, L. R. Raters, ratees, and randomness in personality structure. *Journal of Personality and Social Psychology,* 1966, *4,* 681–691.

Nunnally, J. C. *Psychometric theory.* New York: McGraw-Hill, 1967.

Reid, J. B. Reliability assessment of observation data: A possible methodological problem. *Child Development,* 1970, *41,* 1143–1150.

Romanczyk, R. G., Kent, R. N., Diament, C., & O'Leary, K. D. Measuring the reliability of observational data: A reactive process. *Journal of Applied Behavior Analysis,* 1973, *6,* 175–184.

Rowley, G. L. The reliability of observational measures. *American Educational Research Journal,* 1976, *13,* 51–59.

Taplin, P. S., & Reid, J. B. Effects of instructional set and experimenter influence on observer reliability. *Child Development,* 1973, *44,* 547–554.

Tinsley, H. E. A., & Weiss, D. J. Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology,* 1975, *22,* 358–376.

Wicker, A. W. An application of the multitrait-multimethod logic to the reliability of observational records. *Personality and Social Psychology Bulletin,* 1975, *1,* 575–579.

Received December 21, 1977 ■