Copyright ASHA 1994

Journal of Speech and Hearing Research, *Volume 37*, 264–278, April 1994

The Reliability of Observational Data: I. Theories and Methods for Speech-Language Pathology

Anne K. Cordes
University of California, Santa Barbara

Much research and clinical work in speech-language pathology depends on the validity and reliability of data gathered through the direct observation of human behavior. This paper reviews several definitions of reliability, concluding that behavior observation data are reliable if they, and the experimental conclusions drawn from them, are not affected by differences among observers or by other variations in the recording context. The theoretical bases of several methods commonly used to estimate reliability for observational data are reviewed, with examples of the use of these methods drawn from a recent volume of the *Journal of Speech and Hearing Research* (*35,* 1992). Although most recent research publications in speech-language pathology have addressed the issue of reliability for their observational data to some extent, most reliability estimates do not clearly establish that the data or the experimental conclusions were replicable or unaffected by differences among observers. Suggestions are provided for improving the usefulness of the reliability estimates published in speech-language pathology research.

KEY WORDS: reliability, interjudge agreement, observational data, behavioral data

Much research and clinical work in speech-language pathology requires human observers, or judges, to record information about the speech or language behaviors of human subjects. Such "direct observation" or "behavioral observation" methods have been referred to as the hallmark of behavioral psychology and related disciplines (see Hartmann & Wood, 1990; Suen & Ary, 1989), and it is widely accepted that these methods can provide useful and relatively objective data about the behavior of human subjects.

It is also widely acknowledged, however, that the data actually obtained from direct observations may depend as much on the behavior of the observers as on the behavior of the subjects. The effects of interobserver variability on observational data have been discussed for many years, both for behavioral psychology in general (see, e.g., Foster, Bell-Dolan, & Burge, 1988; Hartmann & Wood, 1990; Rosenthal, 1966; Wasik, 1989; Wildman & Erickson, 1977) and for speech and language research in particular (see, e.g., Ball, 1991; Bassich & Ludlow, 1986; Kearns, 1990; Kreiman, Gerratt, Kempster, Erman, & Berke, 1993; Shriberg & Lof, 1991; Young, 1984). Perceptions, judgments, and observations are affected by variables attributed to the observers, to the instrumentation or coding procedures, to the situation or conditions of observation, to the subjects being observed, and to interactions among all of these. Consequently, researchers using direct observation methods are currently expected to provide evidence that their findings are not simply the results of situational influences or observer idiosyncrasies. They are expected, in other words, to provide evidence that their data are reliable.

The question of observational data reliability is highly relevant to research in speech-language pathology. Of 92 articles, reports, and research notes published in the Speech and Language sections of the most recent complete (as of this writing)

0022-4685/94/3702-0264$01.00/0

volume of the *Journal of Speech and Hearing Research* (*35,* 1992), at least 83 contained data about speech or language development or disorders that had been collected by human judges or observers. Almost 80% of those articles (66 of 83) addressed the question of whether their experimental data were reliable. Of the 66 articles that did address reliability issues, at least 12 reported correlations between data produced by two or more observers, and at least 32 reported an interjudge percent agreement statistic. At least 23 addressed the issue of interobserver differences by using or presenting data from more than one observer, or by describing the absolute mathematical differences between measures made by different observers.[1] Reliability issues for observational data were themselves the focus of at least four articles (Cordes, Ingham, Frank, & Ingham, 1992; Martin & Haroldson, 1992; Onslow, Adams, & Ingham, 1992; Wood, Hughes, Hayes, & Wolfe, 1992), and one response to a previous article (Cannito, 1992; see also Fitch, 1990, 1992). Problems with the reliability of observational data are also among some of the most serious current methodological concerns in several sub-areas of the discipline (e.g., voice quality ratings [see Kreiman et al., 1993], transcription [see Shriberg & Lof, 1991], and stuttering [see Cooper, 1990, and Cordes & Ingham, 1994]).

This article reviews the theoretical bases of several methods currently used to estimate the reliability of observational data in speech and language research. Examples of the use of these methods are drawn almost exclusively from the *Journal of Speech and Hearing Research* (*35,* 1992). The differences and similarities among reliability estimates, correlations, and interobserver agreement indices are emphasized, and some of the difficulties in interpreting some recent reports of reliability estimates are discussed. Finally, suggestions are provided for the effective practical application of reliability estimation methods to speech and language research.

Defining Reliability and Agreement for Observational Data

Defining reliability requires differentiating between two distinctly different uses of the term. Its first and most common use is as a broad concept related to the general trustworthiness of obtained data. In this sense, synonyms for reliability include dependability, consistency, predictability, and stability (Suen & Ary, 1989, p. 99). Kearns (1990, p. 79) equated reliability and "the consistency and reproducibility of measurement"; Johnston and Pennypacker (1980, p. 191) stated that "reliability is concerned with the stability of measured values under constant conditions"; and Mitchell (1979, p. 376) defined a reliable test as one that "shows stability, consistency, and dependability of scores." These definitions all address the experimenter's need to know that data could be reproduced if the same subjects were tested again under similar circumstances (Crocker & Algina, 1986). This is one definition of reliability, then: the dependability or reproducibility of test scores or other data.

A rather different definition of reliability is provided by those who equate it with a specific psychometric property of test scores. In this sense, reliability is defined in terms of a mathematical relationship between an "observed" test score and the test-taker's "true" score on that test. True score, in this sense, is the hypothetical average score that a subject would receive across all possible forms of the same test administered on all possible occasions; it is a statistical concept for a test, not an absolute or "true" description of a state or an event (Crocker & Algina, 1986). True score is assumed to be stable for a given subject for a given test across all possible test forms, testing occasions, test administrators, test scorers, and so on, but observed scores vary because of random measurement error. The reliability coefficient is defined for a data set as the ratio of true score variance to observed score variance, or the proportion of variance in the observed scores that can be attributed to variance in the true scores (Crocker & Algina, 1986; Cronbach, 1947; Suen, 1990). In these terms, observed scores are said to be reliable if they vary with variations in true score, rather than with variations in measurement error. Thus, reliability is also defined as "the extent of absence of random error variance" (Suen & Ary, 1989, p. 118).

This second type of reliability, the more precisely defined mathematical concept, may be thought of as a subtype of the more general reliability, when that term is defined as consistency, dependability, reproducibility, or stability. In other words, one estimate of the dependability or reproducibility of a test score (i.e., of the general reliability of that score, in one use of the word) is provided by the statistical relationships among the observed score, the true score, and measurement error (i.e., by its reliability, in another use of the word).

Some confusion is introduced into definitions of reliability by statements that equate reliability (in its general sense of reproducibility) with interjudge or interobserver agreement: "Interobserver agreement, also referred to as reliability, refers to the extent to which observers agree in their scoring of behavior" (Kazdin, 1982, p. 48). It is more generally accepted, however, that there is a clear distinction between reliability and agreement (Crocker & Algina, 1986; Hartmann, 1977; Lahey, Downey, & Saal, 1983; Mitchell, 1979; Suen, 1990). Agreement is defined simply as the extent to which there is correspondence among the scores or ratings assigned by different judges or observers, or by the same observer on different occasions (Hollenbeck, 1978; Kazdin, 1982; Kearns, 1990; Mitchell, 1979; Suen, 1990). Interobserver agreement may be estimated in several ways, the most common of which is to calculate the percentage of measurement opportunities or observations recorded identically by two independent observers (see below).

High agreement is generally accepted as evidence that observers have recorded behaviors similarly and consistently. If this is the case, then the assumption is made that variations in the scores provided by one observer can reasonably be attributed to actual variation in the subject's behavior, rather than to variation or inconsistency on the part of the observers. There is a strong parallel between agreement and the precise psychometric definition of reliability: an attempt is made to separate the subject's performance from errors introduced into the recorded data while measuring that

[1] These numbers do not sum to 66 because many studies included two or more procedures for assessing the reliability of experimental data.

performance (by observer inconsistency or random measurement error).

Some authors have suggested that agreement is one component of the larger concept of reliability: "Dimensions of reliability include scorer consistency (sometimes referred to as observer agreement), temporal stability, and internal as well as situational consistency" (Hartmann & Wood, 1990, p. 121). Both Hartmann and Wood (1990) and Hollenbeck (1978) suggested that stability or dependability is only a part of reliability. Johnston and Pennypacker (1980) and Suen and Ary (1989), however, equated stability and reliability. To add to the confusion, Hollenbeck (1978) defined reliability in terms of both stability and "accuracy," which he further defined as "precision, or whether the measurement is a true representation of what is being observed" (Hollenbeck, 1978, p. 81). Similar definitions of accuracy were presented by Kearns (1990, p. 83), Kazdin (1982, p. 50), and Johnston and Pennypacker (1980, p. 190), all of whom defined accuracy in terms of whether data reflect some true or actual behavior. Accuracy has also been defined as the correspondence between an observer's scores and some established criterion (Barlow & Hersen, 1984; Kearns, 1990; Tryon, 1985a). Suen (1988, pp. 343–344) and Suen and Ary (1989, pp. 101–102) provided several examples of contradictory definitions of these terms, including suggestions from various authors that accuracy measures reliability, that reliability measures accuracy, that accuracy measures validity, and that accuracy, reliability, and validity are distinctively different.

Suen (1988, 1990), Suen and Ary (1989), and Johnston and Pennypacker (1980) all present similar solutions to the problem of relating reliability and accuracy. These authors do not address accuracy in the context of reliability discussions, suggesting instead that accuracy is actually criterion-related or criterion-referenced validity. Criterion-referenced validity may be defined equally well by any of the definitions provided above for accuracy (Barlow & Hersen, 1984; Hollenbeck, 1978; Kazdin, 1982; Kearns, 1990; Johnston & Pennypacker, 1980): that is, criterion-referenced validity refers to whether data reflect some true or actual behavior, as measured by the correspondence between an observer's scores and some established criterion. Validity or accuracy, in other words, may be considered critical to data interpretation but distinct from reliability (Johnston & Pennypacker, 1980; Kazdin, 1977; Suen, 1990; Suen & Ary, 1989). The distinction between reliability and validity, or between reliability and accuracy, is also expressed in the common argument that two independent observers may score behaviors inaccurately and still show high agreement: they may both be inaccurate in the same way (e.g., Cone, 1977; Deitz, 1988; Kearns, 1990). This possibility leads to the generally accepted notion that observational data must be both reliable and valid to be meaningful or interpretable (e.g., Cone, 1977; Johnston & Pennypacker, 1980; Suen, 1988).

Finally, the distinction between definitions of reliability and methods for estimating reliability should be made explicit. Definitions of reliability were provided above: these include such formulations as "reliability is the dependability and reproducibility of obtained data" or "reliability refers to a relationship between observed scores and true scores." Complications and contradictory definitions are introduced by statements that equate reliability and a method for estimating reliability.

For observational data, these complications are compounded by the fact that the method in question is often the calculation of interobserver agreement. Wildman and Erickson (1977), for example, claimed that the reliability of observational data can be measured by the extent of observer agreement for those data. Kearns (1990, p. 79) stated, even more emphatically, that "for the purposes of single-subject treatment research, [reliability] is generally equated with the calculation of interobserver agreement scores." The latter statement is unquestionably correct: in practice, the reliability of observational data is often estimated by calculating interobserver agreement (Kelly, 1977; Mitchell, 1979). It is not true, however, that reliability and interobserver agreement are equivalent. The calculation of interobserver agreement is not the only way to measure reliability; nor is interobserver agreement simply a way to measure reliability. Most importantly, it should be recognized that the act of calculating and reporting an interobserver percent agreement figure does not in itself establish the reliability of observational data. Each of these points is addressed in greater detail below.

It is also worth noting, almost parenthetically, that the distinction between reliability and validity is not actually as clear as simplified descriptions of these two terms would make it seem. Textbook definitions of reliability refer to the replicability of measurements, and textbook definitions of validity refer to whether those measurements actually measure what they purport to measure. Validity is then traditionally divided to include, for example, content validity, criterion validity, and construct validity (e.g., Ventry & Schiavetti, 1986). More complex descriptions of validity, however, explain all validity decisions as variations on construct validity (see Cronbach, 1988, 1989; Messick, 1989; Suen, 1990), or in terms of the development of a logical argument supporting a particular interpretation of obtained data (Kane, 1992); content validity and criterion validity are viewed as either preliminary or partial aspects of construct or argument-based validity.

The distinction between construct validity and reliability has been questioned as well: Suen (1990), for example, characterized the relationship between reliability and validity as a continuum from comparisons of maximally similar (i.e., parallel) tests of the same construct to comparisons of maximally dissimilar tests of the same construct. Much earlier, Campbell and Fiske (1959) had introduced multi-trait, multi-method matrices by extending the same continuum to include similar and dissimilar tests of different constructs. These many ideas come together in Johnston and Pennypacker's (1980, p. 190) definition of the accuracy or validity of a measurement in terms of whether it provides "the best possible estimate of some dimensional quantity of a natural phenomenon"; this description arguably applies equally well to measurement reliability. Both reliability and (construct) validity, in other words, are concerned with attempts to assure that the numbers produced by some measurement process faithfully represent what they are meant to represent.

Having recognized the difficulties, however, this article will preserve the traditional distinction between reliability (or replicability) and validity. The remainder of this article discusses several methods in current use in speech-language pathology research to estimate the reliability of observational data. For the purposes of this review, reliability is defined broadly as the dependability and replicability of obtained data: data are considered to be reliable if variations to which some meaning is attached can be shown to be variations in the phenomenon being measured and not simply variations in how the data were recorded (Hartmann, 1984; see also Foster et al., 1988; Hawkins & Dotson, 1975; Johnston & Pennypacker, 1980; Kearns, 1990; Shavelson & Webb, 1991; Suen & Ary, 1989). For observational data in particular, it is the differences among observers that are the most obvious, and the most commonly discussed, source of unwanted data variation. In current practice, therefore, the question of data reliability seems to reduce to the question of agreement between observers. Most authors of speech-language pathology research reports briefly address the issue of differences among observers, but it also becomes clear that many published reliability figures do not establish that the data were, in fact, dependable, replicable, reproducible, or otherwise unaffected by the particular observer who happened to gather them. Potential solutions to this problem are also addressed.

Reliability Estimates from Classical Test Theory

Theoretical Background

Classical test theory is also known as classical reliability theory or true score theory (Suen, 1990), the parallel tests model (Suen & Ary, 1989), the classical true score model (Crocker & Algina, 1986), and the psychometric theory of reliability (Mitchell, 1979). Classical test theory posits that three elements are involved in any one test score: the subject's true score for that test (a hypothetical construct, as defined above), random measurement error, and the subject's actual obtained or observed score.

The question of reliability within classical test theory is the question of how much of the variance in observed scores may be attributed to true score variance, and how much must be attributed to random error variance. As error variance decreases, the theoretical ratio of true score variance to observed score variance approaches 1.00, and the scores are said to be increasingly reliable. This ratio, known as the reliability coefficient, cannot be calculated directly, because true scores and true score variance are hypothetical rather than actual properties of the test scores.

In practice, the reliability coefficient for a test score or observational measurement can be estimated from observed scores. The reliability coefficient for any one test is estimated by calculating the correlation, usually a Pearson product-moment correlation, between two different sets of scores from that test. The two sets of scores may be from two different forms of one test, from two different observers on the same testing occasion, from the same observer on two different testing occasions, or from dividing the test or the observations into two or more separate parts. If certain restrictive statistical assumptions are met, then the reliability of observed scores may be estimated by correlating two such sets of scores, observations, or ratings (see Crocker & Algina, 1986; Suen, 1990). One consequence of this formulation of reliability is that many estimates of reliability may be calculated, because more than two sets of scores may be obtained for each subject or for each test (Cronbach, 1947; Suen & Ary, 1988).

The most important assumptions that must be met if reliability is to be estimated within classical theory are known as the Parallel Tests Assumptions. The Parallel Tests Assumptions include the requirements that the two or more sets of scores to be correlated have equal means, equal variances, equal covariances with each other, and equal correlations with some outside criterion measure. In other words, parallel tests are statistically similar. If these assumptions are met for two sets of scores, it can be shown mathematically that the correlation between those scores equals reliability, the ratio of true score variance to observed score variance for either one of the tests.

It is equally important to note that if the Parallel Tests Assumptions are not met, then the correlation between two sets of (nonparallel) test scores does not represent the reliability of either test. In this case, the correlation represents nothing more than the correlation: it simply describes the relationship between two sets of numbers (Suen, 1988; Suen & Ary, 1989). The mathematical assumptions of parallel tests are required if the correlation is to be equivalent to the reliability coefficient. It has also been suggested that the Pearson product-moment correlation may be the highest of several correlation-based calculations when the Parallel Tests Assumptions of equal means, equal variances, or both, are violated (Ebel, 1951). Simple correlations between nonparallel sets of test scores are not estimates of reliability, in other words, and they may in fact overestimate the degree of correspondence between those scores.

Correlational methods based loosely on classical test theory have been used in observational research to estimate both interobserver reliability and intraobserver reliability. "Interobserver reliability," in this sense, usually refers to the result of correlating two sets of scores that were produced by two different observers watching and recording the same events (Suen, 1988). It may be considered a direct parallel of alternate-forms or equivalent-forms reliability for conventional tests: two judges observe at the same time and are expected to produce equivalent scores (Suen, 1990). The term "intraobserver reliability" is usually applied to the result of correlating two sets of scores produced by one observer who judged the same event on two different occasions (from a videotape, for example). This may be considered a direct parallel of test-retest reliability: the same judge observes on two different occasions, and the same score is expected on both occasions (Cone, 1977; Suen, 1990; see also Suen & Ary, 1989, for a conflicting definition of intraobserver reliability).

As common as these terms and procedures may be, some methodologists claim that they have not been thoroughly understood by behavioral researchers (Cone, 1977; Hollenbeck, 1978; Suen, 1988, 1990). First of all, as discussed above, these correlations only estimate the reliability coefficient if the assumptions of classical test theory, and especially the Parallel Tests Assumptions, are met. If these assumptions are not met, the correlation is not an estimate of reliability.

Secondly, according to classical theory, these calculations produce estimates of the reliability of a *score* or a set of scores (Crocker & Algina, 1986; Suen, 1988; 1990; Suen & Ary, 1989). They do not produce estimates of the reliability of the observers, as the terms "interobserver" and "intraobserver" might imply. This argument suggests that the reliability of every observation or set of observations must be determined, and that it is inappropriate to suggest that the "reliability of the observers" has been established.

Thirdly, interpretation of correlation coefficients is complicated by the fact that the magnitude of the correlation between data sets depends on the amount of variance in each set. Low correlations may be obtained, and low reliability inferred, if stable patterns of behavior lead to relatively homogeneous data sets (Lahey et al., 1983; Mitchell, 1979).

Finally, there is substantial functional disagreement about both the use and the meaning of the terms "interobserver reliability" and "intraobserver reliability" (Cone, 1987; Suen, 1988). Because of the imprecision with which they are used by applied behavioral researchers, these terms are equally likely to refer to a correlation between two sets of scores or (incorrectly) to a percent agreement calculation (Suen, 1988).

Reliability estimates based on the methods of classical test theory are common in observational research, although they have perhaps become less common than they once were (Hillis, 1991, 1993; Kelly, 1977; Mitchell, 1979). The decline in the use of correlations to estimate the reliability of direct observational data can be traced to two main criticisms of the method. The first, and most common, is that correlations provide no information about the reliability or replicability of any one observation (Hartmann, 1977; Hartmann & Wood, 1990; Kearns, 1990; Mitchell, 1979). The reliability coefficient is estimated by correlating sets of individual scores, but the result provides only one estimate of reliability for the set of scores as a whole.

This property of classical reliability theory is not a limitation within the confines of classical analyses of test scores: many researchers are indeed interested in the reliability of groups of scores from a given test, or in the reliability of the scores obtained from a set (composite) of test items, rather than in the reliability of responses to particular test items (Crocker & Algina, 1986). In addition, classical test theory provides methods for estimating the standard error of measurement and for creating confidence intervals around individual scores; the use of confidence intervals to interpret individual scores will be familiar to users of standardized assessment instruments. Behavioral researchers, in contrast, are often concerned about observers' judgments of individual behaviors or individual experimental trials, information that cannot be provided by a correlation.

The second criticism leveled at classical reliability methods by behavioral researchers is that the correlation may be quite high even in the absence of agreement between two sets of scores (Hartmann, 1977; Kearns, 1990; Mitchell, 1979; Tryon, 1985b). A typical example of this criticism demonstrates that the two sets of scores 1, 2, 3, 4, 5, and 101, 102, 103, 104, 105 correlate perfectly, or produce an estimate of the reliability coefficient equal to 1.00, even though it is obvious that the scores are very different. The emphasis in observational research appears to have been on assuring that the absolute scores produced by two different observers are relatively similar, and the correlation provides no information about absolute scores.

Again, this potential limitation of classical reliability theory as applied to observational data generally does not present a problem within classical test theory. The latter is concerned with the relative abilities of different subjects based on their total test scores (Crocker & Algina, 1986; Suen, 1990). It is also worth noting that this criticism of classical methods is based on a questionable application of classical reliability methods. Specifically, correlating the sets of scores 1, 2, 3, 4, 5, and 101, 102, 103, 104, 105 cannot produce an estimate of the reliability coefficient, because these scores do not satisfy the Parallel Tests Assumption of equal means. According to classical theory, correlating these scores produces only a correlation, not an estimate of reliability. Some variants of classical theory are based on random sampling of "exchangeable" tests and test items; within these variant theories, the correlation of scores from any two exchangeable tests may represent the test's reliability. In either case, however, the fact remains that the correlation of 1.00 fails to describe adequately the relationship between these two sets of scores, and for this reason alone correlational methods are problematic for observational research.

In summary, classical test theory provides one method for estimating the reliability of behavioral observation data. If the Parallel Tests Assumptions are met, then the Pearson product-moment correlation between two sets of test scores provides an estimate of the reliability coefficient, or the proportion of observed score variance that is true score variance. Strategies for obtaining parallel tests in direct observation research have been described (e.g., Suen, 1990; Suen & Ary, 1989), but the consensus among those concerned with methodology for behavioral observation appears to be that correlations do not provide adequate information about reliability.

Correlations as Reliability Estimates in Speech-Language Research

Correlations remain relatively common as reliability estimates in speech-language pathology research, despite the methodological concerns discussed above. At least 12 studies published in the *Journal of Speech and Hearing Research, 35,* included correlations as estimates of the reliability of observational data.[2] Bishop and Adams (1992), for example, reported correlations of between .910 and .963 for the scores provided by two judges, each of whom had independently coded children's responses to a receptive language task as either correct, partially correct, or incorrect.

[2] Many studies are cited in the following sections as examples of the better or poorer use of reliability-estimation methods in speech-language pathology research. That these studies are so mentioned should not be construed as criticism of their methods or findings in any area other than their reliability-estimation methods.

Kreiman, Gerratt, Precoda, and Berke (1992) correlated the scores provided by all possible pairs of a group of judges rating the similarity of voice samples, and Zwirner and Barnes (1992) correlated scores provided on two occasions by the same judge.

These three studies based their correlations on two measures of all available experimental data, but it is probably more common for investigators to correlate repeated measurements of some fraction of their data. This practice introduces the possibility that the reported correlations may be based on a potentially unrepresentative data sample, or may not reflect the relationship that would have existed between judges for all observations. Walker, Archibald, Cherniak, and Fish (1992), for example, correlated articulation rate scores from the primary judge and a second judge for only 2 of their 40 subjects; Onslow, Hayes, Hutchins, and Newman (1992) remeasured stuttering and speech rate for 12 of 36 speech samples; Sussman, Hoemeke, and McCaffrey (1992) remeasured data from 1 of 16 subjects; and Hall and Yairi (1992) correlated scores provided by two judges for 10% of their data.

Some reports of correlations as reliability estimates are difficult to interpret for other reasons: it may not be clear exactly what was correlated, or how the data that were correlated relate to the actual experimental data, or what the experimenters intended the correlations to represent. Postma and Kolk (1992), for example, reported correlations of .97 and .94 for total number of speech errors and total number of disfluencies, respectively, as measured by two independent judges. Their experimental data were reported in terms of specific error types, not in terms of total errors, yet no reliability estimates were provided at the level of specific types. Zwirner and Barnes (1992), similarly, provided correlations for part of their perceptual analyses, but not for the acoustic analyses that formed the center of their study or for their overall severity and intelligibility judgments. Zwirner and Barnes also appear to have estimated only self-correlations for their two judges, with no comparison between judges, although this may be an incorrect interpretation of their statement that "test-retest procedures (Pearson product-moment correlation) for both listeners indicated high reliability (0.85 and 0.91)" (Zwirner & Barnes, 1992, p. 763).

There are also at least two examples within the recent speech-language pathology literature of high correlations masking substantial differences between judges. Sussman et al. (1992, p. 772) reported that correlations of scores from two judges "exceeded .95." They also reported, however, that the mean difference between the two judges for these acoustic measures was 97.2 Hz, a difference that could quite conceivably have affected, if not created, some of their significant group differences (on the order of 100 to 300 Hz) (p. 772; see also their Table 1, p. 773). Onslow, Hayes, et al. (1992) reported differences between two judges of up to 13 percentage points in percent syllables stuttered measurements and up to 71 syllables per minute in speech rate measurements. The correlations between these two judges were high enough (.97 and .92 for the two measurements), however, that Onslow, Hayes et al. declared the data sufficiently reliable for their purposes.

Finally, Ansel and Kent's (1992) use of correlation-based reliability estimates illustrates some of the difficulties with many applications of these methods. In the first of their two studies, listener intelligibility scores were derived based on the number of phonetic contrasts within CVC syllables that were correctly transcribed by a group of listeners ($n = 8$). Intelligibility scores for repeated transcriptions of the same syllables were "consistent," as evidenced by an "intrajudge agreement coefficient" (which they had previously defined as a Pearson product-moment correlation) of .89 (Ansel & Kent, 1992, p. 300). Intelligibility scores also "varied significantly," however, between the two transcriptions for one particular type of stimulus, as evidenced by an "intrajudge agreement coefficient of .67" (p. 300). It is not clear whether the reference is to true statistical significance (an interpretation that might be unwarranted; see Suen, 1990) or to the fact that this was the lowest of their correlations, nor is it clear why correlations were not reported individually for the other stimulus types. Listeners were also randomly divided into two groups, and a "correlation was calculated to determine the relationship between the intelligibility judgments of the two groups of listeners . . . in an attempt to ascertain the consistency of judgments across listeners" (p. 300). The result of this calculation was an "intrajudge reliability coefficient between the two sets of judges" (pp. 300–301) of .85. Clearly, this statistic was not an "intrajudge reliability coefficient" at all, but simply a correlation comparing the mean scores of the two groups (i.e., an inter-group comparison).

In addition, the .85 figure was interpreted as suggesting that "72% of the variance was accounted for" (p. 301) when, if it were a reliability coefficient, it would have indicated that 85% of the variance in observed scores could be attributed to variance in true scores (see Suen, 1990). This figure was also interpreted as indicating "substantial agreement between listeners" (p. 301). A more precise interpretation of these figures, however, provides no information about agreement between individual listeners; the correlation between the groups' mean scores shows only that those means rank-ordered the intelligibility of the CVC tokens similarly. Intelligibility scores from all listeners were averaged to create a mean intelligibility score for data analyses; perhaps agreement among the listeners might have been most clearly shown in this case by simply reporting the range of intelligibility scores as well as the mean for each speaker (see below).

Correlations of the data provided by two judges (or by one judge on two occasions; Zwirner & Barnes, 1992), in summary, are as problematic in recent speech-language pathology research as in other observational contexts. It is generally unclear whether researchers intended the correlations to represent reliability coefficients, in the particular sense described within classical test theory, or whether the correlations were intended merely as comparisons between two data sets. In either case, both major criticisms of correlations as reliability estimates (discussed above in the context of general behavioral observation methods) obviously continue to be problematic: correlations provide no information about the reliability or replicability of any one observation, and correlations may be quite high even in the absence of agreement between the two sets of scores. Correlations have been used in recent speech-language pathology research as reliability estimates for some fraction of the experimental data, as well as for data that are only tangentially related to the main experimental data. They have been interpreted as establishing some absolute (as opposed to relative) relationship or agreement between judges, even though correlations compare only the relative rank-ordering of data sets. Some of this research, in other words, suffers from the same problem that troubles observational research in other areas: correlations alone cannot establish that the experimental data were replicable or dependable, or were not substantially affected by the particular observers who happened to be involved.

Reliability Estimates from Interjudge Agreement

Theoretical Background

Perceived limitations in correlational methods have led to the widespread use of interjudge agreement calculations for estimating the reliability of observational data. Percent agreement methods do not address reliability in the strict psychometric sense of true score variance and error variance, and they do not fully address reliability in the general sense of the dependability or reproducibility of any one numerical estimate of behavior. They do address one part of general reliability: that part of the inconsistency of measurements that can be attributed to differences among observers.

In contrast to the correlational methods based on classical test theory, percentage agreement calculations have been described as essentially atheoretical descriptions of data (Suen & Ary, 1989). Percentage agreement indices describe the extent of correspondence between ratings, judgments, or observations made by two independent observers or by one observer on two different occasions. This correspondence is expressed as one summary statistic that estimates the percentage (or, less often, the proportion) of trials, observations, or other recording opportunities in which behavior was scored identically. There are several types of percentage agreement indices, three of which will be briefly reviewed.

The simplest of the percentage agreement indices is usually labeled "total percentage agreement" (Kearns, 1990); it has also been called the "smaller/larger index" (Suen & Ary, 1989). This estimate of agreement is calculated for two sets of scores by simply dividing the smaller total number of behaviors observed or correct responses identified by the larger number; the resulting proportion is usually multiplied by 100 to produce a percentage. This calculation assumes that the observer with the higher total identified all the behaviors that were identified by the observer with the lower total. Agreement is defined as the overlap in the total number of behaviors identified by the two observers or on the two occasions, and disagreement is defined as the discrepancy between the smaller total number and the larger.

This method of estimating agreement provides very little information about the correspondence between two sets of scores, and it is generally not recommended for use with observational data (Kearns, 1990; Suen & Ary, 1989). The assumption that the smaller total includes only behaviors that are included in the larger total is untenable, and no information is provided about agreement for particular observations. For these reasons, total percent agreement is currently used in observational research primarily for cases where the data structure prevents the use of another, more desirable, statistic.

The most common improvement to the total percentage agreement index produces a statistic that is variously known as "point-by-point," "event-by-event," "item-by-item," "occasion-by-occasion," "interval-by-interval," or "exact" percent agreement (Kelly, 1977; Suen & Ary, 1989). Point-by-point agreement is estimated for two observers (or for one observer for two occasions) by directly comparing judgments for each recording opportunity. An opportunity may result in an agreement, if both observers scored that trial or that interval in the same way, or a disagreement, if they scored it differently. The agreement index is calculated by dividing the number of agreements by the total number of agreements plus disagreements and multiplying by 100. Information about interobserver agreement is available for each individual recording opportunity (although that information is collapsed to report a percentage for the test or group of observations as a whole), so this calculation addresses one of the main criticisms of smaller/larger agreement estimates. One limitation of exact percent agreement is that recording opportunities must be defined for agreements and disagreements to be defined. For the case where recorded data consist solely of frequency tallies or ongoing descriptions of behavior, this method of calculating agreement is inappropriate.

The point-by-point agreement method is among the most common in behavioral research (Kelly, 1977; Mitchell, 1979; Suen, 1988). Point-by-point agreement is readily calculated, and a simple interpretation of the information it provides is easily made. Several variations are also prevalent, including methods for calculating agreement for correct responses or occurrences of some behavior only (the percent agreement for occurrences) and methods for calculating agreement for the absence of some behavior (the percent agreement for nonoccurrences) (Harris & Lahey, 1978; Hartmann, 1977; Hopkins & Hermann, 1977). In general, percent agreement for occurrences defines agreements in terms of the desired or target behavior, ignoring trials where the observers agreed on the absence of the behavior. Percent agreement for nonoccurrences uses the opposite procedure. These variations are used to control for the artificial inflation of the percent agreement index that may result when a large proportion of ratings or trials are essentially irrelevant to the judgments being made by the observers.

Concern over artificial inflation of the agreement index also leads to a third method for estimating agreement. It is generally recognized that some agreements between observers will occur by chance. Chance agreement may result when behavior is especially frequent or infrequent relative to the recording opportunities, as discussed above. It may also result when the judgment process is based on a small number of fixed categories or decisions, such as decisions regarding correct or incorrect responses. Arguably, the best estimate of interjudge agreement in these cases should indicate the level of agreement that exists above the level expected by chance alone (Cohen, 1960; Fleiss, 1975; Hirji & Rosove, 1990). Cohen (1960) introduced a statistic, known as "kappa," that essentially corrects obtained percent agreement for expected chance agreement. Kappa is defined as the ratio of the difference between obtained percent agreement and expected chance agreement to the difference between perfect agreement (i.e., 1.00) and expected chance agreement. Thus, kappa indicates the extent to which obtained interjudge agreement exceeds chance agreement.

There exists a family of statistics similar to kappa (Fleiss, 1975; Zwick, 1988) that differ primarily in their mathematical definition of chance agreement (Zwick, 1988). Kappa and the other kappa-like statistics are widely recommended by behavioral methodologists as the best estimators of interobserver agreement (Hartmann, 1984; Suen, 1988; Suen & Ary, 1989). Perreault and Leigh (1989, p. 137) claimed that kappa is "the most widely used measure of interjudge agreement across the behavioral science literature," but other reviews suggest that kappa is not nearly as widely used as the simpler point-by-point agreement methods (Hillis, 1991; Mitchell, 1979; Suen, 1988).

Several potential mathematical weaknesses of kappa have been identified (e.g., Uebersax, 1988; Zwick, 1988). One relatively simplistic problem results from the fact that kappa is defined as a ratio of differences: the difference between obtained agreement and chance agreement is divided by the difference between perfect agreement and chance agreement. Kappa is to be interpreted as indicating the "proportion of the distance between chance and perfect agreement at which [obtained agreement] is located" (Hillis, 1991, p. 10). This interpretation of kappa assumes that any obtained agreement that could possibly have been obtained by chance was, in fact, obtained by chance. A similar flaw in the smaller/larger index is seen as fatal to the interpretation of that statistic: it assumes that any overlap in total counts that could have produced agreement did, in fact, produce agreement. This assumption is widely held to be unreasonable, but the fact that kappa is essentially a variation on a smaller/larger index for obtained and chance agreement does not appear to have been cause for concern.

Interjudge Agreement in Speech-Language Pathology

Point-by-point agreement appears to be the currently most frequently used agreement statistic in observational research (Mitchell, 1979; Suen, 1988). Interjudge agreement estimates were also widely used in the *Journal of Speech and Hearing Research, 35,* but several common problems complicate interpretation of these figures.

First, it is often difficult to determine if agreement calculations were smaller/larger indices or more exacting point-to-point comparisons. Ryan (1992) clearly used smaller/larger indices to estimate agreement between two judges; many other researchers clearly used point-to-point comparisons (e.g., Gierut, 1992; Imai & Michi, 1992; Nippold, Schwarz, & Undlin, 1992; Otomo & Stoel-Gammon, 1992; Rice, Buhr, & Oetting, 1992; Violette & Swisher, 1992; Wetherby & Rodriguez, 1992). Many reports, however, do not include enough information to allow conclusive determination of how agreement figures were calculated or what they might mean. Evans and Craig (1992, p. 347), for example, calculated some unspecified form of "percentage agreement between experimenter and observer for transcription and scoring reliability." Purcell and Liles (1992, p. 358) reported that "intraexaminer reliability" and "interexaminer reliability" were calculated by re-scoring a portion of the data and "computing the percentage of agreement for the major scoring categories," and Abkarian, Jones, and West (1992, p. 583) reported only that two judges independently rated each subject response and "between-judge agreement levels exceeded 90%." Yairi and Ambrose (1992), finally, assessed point-by-point agreement using an agreement index introduced by Sander (1961) as a total counts agreement measure (smaller/larger); it is difficult to determine whether these were point-by-point or total counts analyses.

A related difficulty with interpreting some reports of percent agreement calculations is that it is not clear what defined one opportunity for an agreement, or how different judgments were compared. Tompkins, Boada, and McGarry (1992), for instance, calculated agreement between two judges who had classified subjects' responses as correct or incorrect; it seems a reasonable assumption that these comparisons were item-by-item. They also reported that agreement was calculated for five subcategories that were used to classify error types for incorrect responses; it is less clear what was compared or how these category-by-category agreement figures were computed.

A report by Light, Dattilo, English, Gutierrez, and Hartz (1992) contains another example of the same problem. Agreement methods used during training phases were explicitly described as point-by-point comparisons, but interobserver agreement assessments for the experimental data were described only as dividing number of agreements by total number of agreements plus disagreements. This description most probably refers to point-to-point comparisons (in this case, probably utterance-by-utterance), but it is not at all clear how those points were defined or how comparisons were derived from the study's descriptions of on-going communicative interactions.

Transcription data represent a special set of difficulties for reliability assessment, several of them related to the issue of defining the items to be compared (see Shriberg & Lof, 1991). McGregor and Schwartz's (1992, p. 597) report is typical: for 20% of a 500-utterance speech sample, "intertranscriber reliability was 85% overall and 92% when differences in voicing characteristics were dismissed." Perhaps the most reasonable assumption is that "intertranscriber reliability" refers to an intertranscriber agreement calculation; comparisons were probably made phoneme-by-phoneme, with gross differences in transcription (such as the omission of words) either ignored or considered one disagreement, although this is not at all obvious. Otomo and Stoel-Gammon (1992) allowed transcribers to choose two broad phonetic symbols to represent the vowels produced by 20- to 30-month-old children. Token-by-token interjudge agreement was reported for all pairs of the three judges, but the authors did not specify whether two or four of the phonetic symbols chosen by any pair of judges had to agree for an agreement to be registered for that pair. It is equally difficult to determine what constituted an agreement when Paul and Jennings (1992, p. 102) calculated "agreement on the percentage of consonants correct" (probably consonant-by-consonant agreement for correct and incorrect productions rather than agreement on the final PCC value) and "reliability for the consonant inventories" (probably consonant-by-consonant agreement for the presence or absence of phonemes).

Another problematic issue in some recent reports of interobserver agreement is that agreement comparisons have been carefully described for data that were not the focus of the experiment's main data analyses. Wetherby and Rodriguez (1992), for example, provided interjudge agreement data for the coded intentionality of toddlers' communicative acts, but they did not address interobserver agreement for the total numbers of communicative acts produced by each child or for the intentionality of acts coded in different (structured or unstructured) play situations. Gierut (1992) and Morrison and Shriberg (1992) assessed interobserver agreement for transcription, but not for their determinations of phonological process descriptions. Some studies have assessed interobserver agreement for their subject-selection and pre-test data but not for the experimenter's decisions about children's picture-pointing or other test responses (e.g., Fuller & Lloyd, 1992; Rice et al., 1992).

Recent publications in speech-language research also include evidence of at least two variations on consensus procedures. These procedures are combined with percent agreement calculations in attempts to establish that the data used for experimental analyses were not affected by the idiosyncrasies of individual judges. Many experimenters used consensus procedures to resolve any disagreements identified from agreement comparisons, in order to develop data sets that reflected the judgments of more than one observer (e.g., Abkarian et al., 1992; Denny & Smith, 1992; Gutierrez-Clellen & Iglesias, 1992; Nippold et al., 1992; Thal & Tobias, 1992; Tompkins et al., 1992; Violette & Swisher, 1992). An essentially opposite route was followed for some smaller number of studies (e.g., Duchan, Meth, & Waltzman, 1992; Perlman, Grayhack, & Booth, 1992; Prins & Hubbard, 1992): the original decisions in these studies were made by pairs or groups of judges working together, and the reliability of their decisions was assessed by requiring judges to re-rate the data independently.

Consensus procedures such as these appear to have much to recommend them. At the very least, they may lead to an increased awareness of the extent and location of disagreements between judges, or perhaps even to explanations for or solutions to those disagreements. Consensus judgment procedures also have potential problems of their own, however, and these should be recognized. Shriberg, Kwiatkowski, and Hoffman (1984), for example, presented a method for consensus transcription that seems to have formed the basis of many current transcription consensus methods. Their initial report included the finding that segment-by-segment agreement was only 68% when all diacritics were considered, or 76% when disagreements on "nonerror" diacritics were ignored, between transcriptions produced by the same highly-skilled two-person judge team on two different occasions.

The same problem is apparent in Morrison and Shriberg's (1992) recent study of articulation accuracy in different contexts: transcription agreement between judge-teams began as low as 61.3% for narrow transcription, and agreement between transcriptions produced by the same judge pair at two different times began as low as 65.5%.[3] Shriberg et al. (1984) also noted that when consensus procedures are not thoroughly described, the possibility exists that differences reportedly solved by consensus may have been solved simply by including the judgment of the higher-ranked or otherwise more forceful judge (a conceivable occurrence when students serve as secondary observers).

Consensus transcriptions, in summary, suffer from the same problem that afflicts most other recent reports of interjudge agreement calculations in speech-language pathology research: interjudge agreement figures, as they have been presented, tend not to establish that two or more equivalently trained or experienced judges agreed on the decisions that formed the basis of the main experimental data. Chance-corrected agreement figures such as Cohen's (1960) kappa, the most recommended percent agreement measures, remain rare in the speech-language pathology research literature (but see Adamson, Remski, Deffebach, & Sevcik, 1992; Cordes et al., 1992; Gutierrez-Clellen & Iglesias, 1992; Weston & Shriberg, 1992). Most percent agreement figures in recent speech-language research have been the result of relatively unspecified agreement calculations, based on some fraction of the experimental data; many of these reports have not established that the data were reliable, dependable, replicable, or unaffected by observer idiosyncrasies.

Reliability Estimates: Data from Multiple Observers

The argument resurfaces periodically within the behavior analysis literature that the correlational and percent agreement measures discussed in the preceding two sections are inadequate for establishing the reliability of observational data and should be replaced by methods that more directly reflect the judgments of multiple observers. Hawkins and Dotson (1975), for example, demonstrated that percent agreement estimates may be quite high even in the absence of agreement between observers for the experiment's overall data trends. They suggested the calculation of both occurrence and nonoccurrence percent agreement, as well as the presentation of raw data from two independent observers. Kratochwill and Wetzel (1977) commented positively on the practice of graphing data from two observers, although they saw little need for publishing such data: "If plotting both sets of data reveals . . . that only one observer recorded an experimental effect, the data should not be published, acceptable statistical representations of observer agreement notwithstanding" (p. 134). Birkimer and Brown (1979) made the slightly different suggestion that the range of observers'

[3] Morrison and Shriberg's (1992) study also includes an excellent discussion of how transcriber differences might affect research in child phonology, as do the reports of Pye, Wilcox, and Siren (1988) and Shriberg and Lof (1991).

disagreements should be plotted on graphic displays of experimental data. This suggestion was endorsed by Hawkins and Fabry (1979, p. 550), who noted that traditional reliability measures have been "of little use" in evaluating the "believability of an experimental effect." With some reservations, the suggestion was also endorsed by Kratochwill (1979).

Other arguments against statistical agreement figures are less convincing (e.g., Hawkins & Fabry's [1979] reasoning that undue emphasis on statistical analyses makes for difficult, uninteresting reading), and there are several cogent arguments in favor of mathematical and statistical agreement indices for observational data (see Hartmann, 1977; Hopkins & Hermann, 1977; Kratochwill & Wetzel, 1977). In general, though, the suggestion that observer agreement might best be assessed through direct comparisons of data from multiple observers is a compelling one; if one goal of reliability estimation for observational data is to establish that experimental results were not affected by judges' idiosyncrasies, then certainly the demonstration that results do not vary across two observers is a reasonable starting point.

There are examples within recent speech-language pathology research of several variations on presenting data from multiple observers. The first involves simply reporting a mean absolute difference between the measures made by one observer and those made by another. Ratner (1992), for example, reported mean differences between the experimenter and a second judge for 20% of her study's subjects for several dependent variables, including speech rate, mean length of utterance, and percentage of disfluent words. Mean differences between judges were also presented in several studies involving acoustic measures (Onslow, van Doorn, & Newman, 1992; Prins & Hubbard, 1992; Sussman et al., 1992; Swanson, Leonard, & Gandour, 1992). Reports of mean differences are subject to many of the interpretation problems that affect other agreement measures; specifically, they may or may not establish that the experimental effects would have been obtained had the second judge served as the primary data source. In Ratner's (1992) study, for example, it seems that the differences between judges were not problematic: interjudge differences were substantially smaller than the differences between groups that were attributed to experimental manipulations. The mean difference reported by Sussman et al. (1992), however, was large enough to have affected some of that study's significant group differences (as discussed above).

Two studies by Onslow and his colleagues (Onslow, Hayes, et al., 1992; Onslow, van Doorn, & Newman, 1992) presented mean differences between observers and took the additional step of repeating their experimental analyses with data provided by the second judge. Despite substantial differences between judges in one of these studies (Onslow, Hayes et al., 1992), the experimental findings were essentially preserved. Gow and Ingham (1992) included scores from two observers for each dependent variable (stuttering frequency, speech rate, and speech naturalness) in their graphic data presentations; some differences between judges were noticeable, but most differences between experimental phases were replicated by the second judge.

Finally, another group of studies has attempted to reduce the effects of differences among observers by developing measurement systems that rely on more than one judge. Such systems exploit the fact that increasing the number of item scores used to create a total score increases the reliability of that total score; that is, the reliability of a subject's total test score may be increased by increasing the number of test questions or the number of observers or observation occasions that contribute to that score (see Crocker & Algina, 1986; Shavelson & Webb, 1991; Suen, 1990). For observational data, this principle suggests that it is more likely that mean or total scores calculated for each subject from, for example, 10 observers could be reproduced by any other 10 observers than that a score based on only one observer could be reproduced by any one other observer (see Shavelson & Webb, 1991).[4]

Dagenais and Critz-Crosby (1992), in one example of depending on multiple judges, presented their listener judgment data by describing target vowel judgments made by a panel of five listeners who each judged five tokens from each of 20 subjects. Otomo and Stoel-Gammon (1992), as mentioned above, derived their data from two judgments (broad transcription symbols) from each of three judges. Imai and Michi (1992) used mean data from six ratings, made by three judges on two occasions, as their primary experimental data. Bishop and Adams (1992), similarly, used mean scores across two judges, and Bedrosian, Hoag, Calculator, and Molineux (1992) used summed scores from 24 judges.

Using data from multiple judges does not address the issue of the level of agreement that exists among those judges. In fact, it is clear in at least some of these studies that agreement was actually quite poor, even though experimental conclusions were preserved (e.g., Onslow, Hayes, et al., 1992). The advantage of using multiple judges is not that disagreements among judges are eliminated, but that the potential for those differences to create spurious experimental conclusions is minimized.

Generalizability Theory and Investigations of Reliability

The three methods described above (correlations based on classical test theory, interjudge percent agreement calculations, and the presentation or analysis of data from more than one observer) constitute the three most common methods for addressing the reliability of data obtained in speech-language research. They have also been reported to be the most common in other behavioral and observational research (Kelly, 1977; Mitchell, 1979). They are not, however, the most comprehensive methods for assessing the reliability of observational data; interestingly enough, they are not even the most widely recommended. Arguably, those distinctions fall to generalizability theory (Cronbach, Gleser, Nanda, &

[4] This principle was developed within classical test theory, for the restricted case of parallel composites of test items, as the Spearman-Brown prophecy formula. It was extended to composites of (nonparallel) test items by Cronbach's (1951) alpha, and further extended within generalizability theory's decision studies to other types of data (see Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991).

Rajaratnam, 1972), a theoretical framework for estimating data reliability that has been described as the most comprehensive of the available alternatives. Generalizability theory was developed for traditional psychometric test-question data, but it has been recommended almost since its introduction as ideally suited to behavior observation data (e.g., Cone, 1977; Kazdin, 1977).

Generalizability theory was developed to assess simultaneously the effects on obtained data of two or more sources of error. It was developed from the methods of classical test theory, and it extends those methods by explicitly allowing that measurement error may have more than one identifiable source (Cronbach, Rajaratnam, & Gleser, 1963; Cronbach et al., 1972; Shavelson & Webb, 1991). Cone (1977) identified six error sources that usually concern behavioral researchers: scorer (or observer), item (or individual behavior), time (or occasion), setting, method (of testing or recording), and the "dimension" of the behavior (its construct validity). The goal of reliability investigations within generalizability theory is to estimate the extent to which scores obtained under certain combinations of these conditions (e.g., with certain observers on certain occasions) are representative of the scores that would be obtained under all other conditions. Under this formulation, unreliable observations are those that depend on some particular combination of observers or settings, and reliable observations are those that may be generalized to other conditions because they are relatively unaffected by the particular conditions under which they were obtained (Cronbach et al., 1972).

The methods of generalizability theory assess the contributions of measurement conditions to obtained data through the (computer-assisted) calculation of variance components. These methods remain undeniably rare in behavioral research in general and in speech-language research in particular (but see Demorest & Bernstein, 1992). An explanation for this state of affairs would be purely speculative: perhaps these methods have been perceived as too statistical (cf. Baer, 1977) for behavioral observation, or perhaps they are simply too recently developed as compared with other available methods.

It is also worth noting, however, that the conceptual bases of generalizability theory are common among behavioral researchers. With some tolerance for differences in terminology, some of the defining features of generalizability theory may actually be seen to be defining features of experimental and applied behavior analysis. The conceptual bases of a generalizability framework, for example, may be found in lists of conditions that affect observations and observational data (e.g., Hartmann & Wood, 1990; Wildman & Erickson, 1977). They may also be seen in the behavior analyst's concerns with operationally defining behavior, sampling behavior in several functional environments, and controlling extraneous variables (e.g., Cone, 1988; Foster et al., 1988); in the terminology of generalizability theory, these are attempts to sample observations of behavior from well-defined conditions of well-defined facets.

The conceptual underpinnings of generalizability theory, and even simple examples of ANOVA-based methods, may also be seen in four recent reliability investigations in speech-language pathology. All four attempted to identify the multiple measurement conditions that influence reliability or observer agreement for measurements of speech disorders. Wood et al. (1992), for example, used intraclass correlations to separate the error variance associated with repeated trials, the error variance associated with repeated testing sessions, and the true variance associated with differences among subjects, for labial closure force measurements in several groups of subjects. Onslow, Adams, and Ingham (1992) used several mathematical and statistical comparisons, including ANOVA-based intraclass correlations, to assess the influences of observer experience, speaker, and range of assigned ratings on the reliability of speech naturalness ratings. Martin and Haroldson (1992) also investigated the influences of several conditions on the reliability of speech naturalness ratings: audiovisual and audio-only speech sample presentations, stuttered and nonstuttered speech, mild to severe stuttering, stuttering frequency, and speech rate. Finally, Cordes et al. (1992) assessed the effects of observer experience, repeated rating occasions, and analysis intervals ranging from 0.5 sec to 7.0 sec on interjudge and intrajudge agreement for the occurrence of stuttering events.

Final Considerations

One issue that has not been addressed in this discussion of reliability estimates involves the specification of an adequate or satisfactory level of numerical agreement. Recommended minimum levels of point-by-point percent agreement figures, for example, vary from 70% to 90% (Foster et al., 1988; Hartmann, 1984; Kelly, 1977), with 80% identified as a "traditional" lower limit for acceptable agreement (Kazdin, 1982). Because behavior frequency affects levels of chance agreement and of obtained agreement, however, it is difficult to establish an absolute criterion for percent agreement levels; equal agreement percentages may not actually reflect equivalent levels of agreement (Suen & Ary, 1989). More importantly, very high numeric agreement estimates can be obtained for individual trials or experimental sessions even if the two observers disagree almost entirely on data trends across the experiment as a whole (Hawkins & Dotson, 1975).

Calculating and reporting both agreement for occurrences and agreement for nonoccurrences has been recommended as a way to control for the inflation of percent agreement that may occur when behavior is of either relatively low or relatively high frequency (Foster et al., 1988; Harris & Lahey, 1978; Hartmann, 1977; Hawkins & Dotson, 1975; Hillis, 1991, 1993; Hopkins & Hermann, 1977). This recommendation does not solve the problem of establishing acceptable limits for occurrence agreement and nonoccurrence agreement. Other authors recommend the use of kappa or another kappa-like statistic instead of simpler percent agreement methods, and the problem of defining an adequate level of percent agreement or occurrence/nonoccurrence agreement is thus transformed into the problem of defining an adequate level of kappa. The results are similar: it is recommended that kappa be anywhere from .50 to .90 to represent satisfactory agreement (Crocker & Algina, 1986; Hillis, 1991; Suen & Ary, 1989).

One measurement method under investigation for use in stuttering research combines the "traditional" 80% minimum level with a separation between occurrence and nonoccurrence agreement to describe judgment opportunities (speech intervals) in terms of whether 80% or more of a group of judges agreed (Cordes et al., 1992; Ingham, Cordes, & Finn, 1993; Ingham, Cordes, & Gow, 1993). This method has proved useful in preliminary attempts to develop a reliable measurement method for stuttering behaviors (see Cordes & Ingham, 1994).

Another possible solution to the problem of defining a satisfactory level of agreement was provided by Hartmann (1984) and discussed above in the context of defining reliability for this review. Hartmann suggested that agreement is just one piece of evidence in the determination of the overall adequacy of data, and that no single numeric limit for agreement estimates should be set. The important question is whether the data provide a "powerful means of detecting experimentally produced or naturally occurring response covariation" (Hartmann, 1984, p. 128). This is essentially the same point made earlier by Hawkins and Dotson (1975) in their discussion of the relationships among agreement figures and the "believability" of experimental effects. It is important to establish for a given data set that data variation interpreted as meaningful cannot be attributed to variation in how those data were recorded or who recorded them (Foster et al., 1988; Hawkins & Dotson, 1975), in other words, but it is not necessarily important for a given data set to report that interobserver agreement met some predetermined 80% or 90% standard.

This is, in summary, the essence of reliability: If the obtained numeric representations of some natural phenomenon are reliable, then variations in those data will represent variations in the phenomenon being measured, rather than variations in the conditions of its measurement (cf. Hartmann, 1984; Johnston & Pennypacker, 1980; Shavelson & Webb, 1991). Because variability may be introduced from so many sources, it is difficult, if not impossible, to establish the replicability or dependability of direct observation data by simply correlating two sets of scores or calculating interobserver agreement (Mitchell, 1979; Suen, 1990). Generalizability theory provides one framework for investigating the effects of multiple error sources and determining how those effects might be mitigated in subsequent experiments, but generalizability theory has not yet been widely adopted for speech-language research.

The many reports of interobserver agreement estimates described in recent speech-language publications do suggest, however, that researchers in speech-language pathology are concerned about establishing the reliability of their observational data. No single reliability estimation procedure is necessarily better than any other, but they are different: correlations compare the relative rank-ordering of scores in different groups, whereas interjudge agreement indices compare the absolute scores assigned by different observers. It seems that some speech-language research might benefit from explicit evaluation of the conclusions to be drawn from the experimental data and whether the provided reliability estimates support those conclusions. In addition, some of the issues discussed in this review suggest several other ways in which the informative value of reliability estimates in speech-language research might be increased.

The first point seems rather obvious: reliability or agreement assessments must be based on directly relevant data. Assessing agreement for subject-selection or stimulus-development data, for example, is important to establishing the reliability of subject-selection or stimulus-development decisions. Assessing agreement for experimental data is important to establishing the reliability of any conclusions drawn from that experiment. The reliability of subject-selection data, however, or of the decisions involved in stimulus development, does not imply the reliability of experimental data.

A second, related, point is that agreement or reliability assessment is intended to establish the reliability of the entire data set. Ideally, therefore, reliability assessments will use the entire data set, until or unless it can be established that a given measurement procedure consistently produces reliable data. If time constraints or personnel limitations prevent the completion of thorough reliability analyses, researchers should at least base their analyses on representative samples of data, from all experimental conditions. How large a sample must be to produce representative estimates of the reliability of an entire data set seems to be an empirical question; in the absence of an empirical answer, one reasonable recommendation might be to estimate reliability from at least half of the experimental data. Previous recommendations also provide other guidelines (see Hartmann & Wood, 1990; Hollenbeck, 1978; Kearns, 1990): reliability analyses should be based on multiple behavior samples from each experimental group or condition, and they should be calculated during and throughout the experiment, rather than once at the end, if the information gained can be used to improve reliability for later conditions.

Third, meaningful interpretation of reliability estimates requires that experimenters describe explicitly what constituted a recording opportunity and what constituted an agreement. Reports that "observer agreement was 90%" (or, worse, that "reliability was 90%") do little to communicate to a research consumer the extent of the differences or similarities among observers, unless those reports are accompanied by clear descriptions of the methodological steps that were completed to obtain those figures. If trials or recording opportunities are clearly defined, the report that reliability estimates were item-by-item interobserver agreement calculations should be explicit enough. However, if data are derived from an ongoing behavior stream through transcription or other analysis of spontaneous speech or communicative interactions, or if individual trials are otherwise undefined, then reliability analyses should acknowledge the many decisions involved: Point-to-point agreement for transcriptions, for example, depends not only on phoneme-by-phoneme comparisons, but on transcribers' decisions in glossing words and in separating utterances.

Finally, reliability should be assessed at the same measurement or judgment level at which experimental effects are assessed (cf. Hollenbeck, 1978; Kearns, 1990). If observers divided speech behaviors into several types of correct responses and several types of incorrect responses, for example, then reliability estimates should be completed for those judgments, rather than simply for the distinction between correct and incorrect responses. Reliability estimates should also reflect the type of conclusions that are being drawn from a given study. Correlational estimates, for example, may be appropriate in the context of relative decisions, or decisions about the ranking or relative standing of subjects or groups (Shavelson & Webb, 1991; see Onslow, Hayes, et al., 1992). Absolute decisions, however, or decisions about absolute levels of subjects' knowledge or behavior, depend on establishing the replicability of individual scores, not just the replicability of the relative or ranked relationships among those scores (see Walker et al., 1992).

The reliability of observational data may always be problematic, simply because there will always be differences among human observers. Observer training programs, complete operational definitions of target behaviors and of behavior rating categories, and careful attention to other methodological details should continue to be recommended as ways to reduce observer differences (see, e.g., Hartmann & Wood, 1990; Kearns, 1990). It will also continue to be important to estimate data reliability, but the mere calculation of these estimates should not be expected to eliminate differences among observers. The point that seems most important to current speech-language research is that assessing the reliability of observational data is more complex than reporting that some vaguely described observer agreement statistic fell at some certain numeric level. Direct behavior observation methods can indeed provide important and relevant data about human speech and language behaviors (see Wertz & Rosenbek, 1992), but researchers should be careful to establish that their conclusions were not influenced by the error introduced into their observational data by their all-too-human judges.

Acknowledgments

The influence of psychometrics courses taught by Richard J. Shavelson at the Graduate School of Education, University of California, Santa Barbara, is gratefully acknowledged, but errors of fact, application, or interpretation are entirely my own. My thanks to Marilyn Demorest and Martin Young for their reviews of this manuscript. Preparation of this manuscript was supported by funds from grant #DC00060, awarded to Roger J. Ingham by the National Institutes of Health.

References

Abkarian, G. G., Jones, A., & West, G. (1992). Young children's idiom comprehension: Trying to get the picture. *Journal of Speech and Hearing Research, 35,* 580–587.

Adamson, L. B., Remski, M. A., Deffebach, K., & Sevcik, R. A. (1992). Symbol vocabulary and the focus of conversations: Augmenting language development for youth with mental retardation. *Journal of Speech and Hearing Research, 35,* 1333–1343.

Ansel, B. M., & Kent, R. D. (1992). Acoustic-phonetic contrasts and intelligibility in the dysarthria associated with mixed cerebral palsy. *Journal of Speech and Hearing Research, 35,* 296–308.

Baer, D. M. (1977). Reviewer's comment: Just because it's reliable doesn't mean you can use it. *Journal of Applied Behavior Analysis, 10,* 117–119.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1987). Some still-current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 20,* 313–327.

Ball, M. J. (1991). Recent developments in the transcription of nonnormal speech. *Journal of Communication Disorders, 24,* 59–78.

Barlow, D. H., & Hersen, M. (1984). *Single case experimental designs: Strategies for studying behavior change* (2nd ed.). New York: Pergamon.

Bassich, C. J., & Ludlow, C. L. (1986). The use of perceptual methods by new clinicians for assessing voice quality. *Journal of Speech and Hearing Disorders, 51,* 125–133.

Bedrosian, J. L., Hoag, L. A., Calculator, S. N., & Molineux, B. (1992). Variables influencing perceptions of the communicative competence of an adult augmentative and alternative communication system user. *Journal of Speech and Hearing Research, 35,* 1105–1113.

Birkimer, J. C., & Brown, J. H. (1979). A graphical judgmental aid which summarizes obtained and chance reliability data and helps assess the believability of experimental effects. *Journal of Applied Behavior Analysis, 12,* 523–534.

Bishop, D. V. M., & Adams, C. (1992). Comprehension problems in children with specific language impairment: Literal and inferential meaning. *Journal of Speech and Hearing Research, 35,* 119–129.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 402–414.

Cannito, M. P. (1992). A questionable consistency: Response to Fitch (1990) [Letter to the editor]. *Journal of Speech and Hearing Research, 35,* 1268–1269.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavior Therapy, 8,* 411–426.

Cone, J. D. (1987). Behavioral assessment: Some things old, some things new, some things borrowed? *Behavioral Assessment, 9,* 1–4.

Cone, J. D. (1988). Psychometric considerations and the multiple models of behavioral assessment. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (3rd ed.) (pp. 42–66). New York: Pergamon.

Cooper, J. A. (Ed.). (1990). *Research needs in stuttering: Roadblocks and future directions. ASHA Reports, 18.*

Cordes, A. K., & Ingham, R. J. (1994). The reliability of observational data: II. Issues in the identification and measurement of stuttering events. *Journal of Speech and Hearing Research, 37,* 279–294.

Cordes, A. K., Ingham, R. J., Frank, P., & Ingham, J. C. (1992). Time-interval analysis of interjudge and intrajudge agreement for stuttering event judgments. *Journal of Speech and Hearing Research, 35,* 483–494.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth, TX: Holt, Rinehart and Winston.

Cronbach, L. J. (1947). Test "reliability": Its meaning and determination. *Psychometrika, 12,* 1–16.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–171). Urbana, IL: University of Illinois Press.

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

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *The British Journal of Statistical Psychology, 16(2),* 137–163.

Dagenais, P. A., & Critz-Crosby, P. (1992). Comparing tongue positioning by normal-hearing and hearing-impaired children during vowel production. *Journal of Speech and Hearing Research, 35,* 35–44.

Deitz, S. M. (1988). Another's view of observer agreement and observer accuracy. *Journal of Applied Behavior Analysis, 21,* 113.

Demorest, M. E., & Bernstein, L. E. (1992). Sources of variability in speechreading sentences: A generalizability analysis. *Journal of Speech and Hearing Research, 35,* 876–891.

Denny, M., & Smith, A. (1992). Gradations in a pattern of neuromuscular activity associated with stuttering. *Journal of Speech and Hearing Research, 35,* 1216–1229.

Duchan, J., Meth, M., & Waltzman, D. (1992). *Then* as an indicator of deictic discontinuity in adults' oral description of a film. *Journal of Speech and Hearing Research, 35,* 1367–1375.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16,* 407–424.

Evans, J. L., & Craig, H. K. (1992). Language sample collection and analysis: Interview compared to freeplay assessment. *Journal of Speech and Hearing Research, 35,* 343–353.

Fitch, J. L. (1990). Consistency of fundamental frequency and perturbation in repeated phonations of sustained vowels, reading, and connected speech. *Journal of Speech and Hearing Disorders, 55,* 360–363.

Fitch, J. L. (1992). Response to Cannito [Letter to the editor]. *Journal of Speech and Hearing Research, 35,* 1269.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics, 31,* 651–659.

Foster, S. L., Bell-Dolan, D. J., & Burge, D. A. (1988). Behavioral observation. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (3rd ed.) (pp. 119–160). New York: Pergamon.

Fuller, D. R., & Lloyd, L. L. (1992). Effects of configuration on the paired-associate learning of Blissymbols by preschool children with normal cognitive abilities. *Journal of Speech and Hearing Research, 35,* 1376–1383.

Gierut, J. A. (1992). The conditions and course of clinically induced phonological change. *Journal of Speech and Hearing Research, 35,* 1049–1063.

Gow, M. L., & Ingham, R. J. (1992). Modifying electroglottograph-identified intervals of phonation: The effect on stuttering. *Journal of Speech and Hearing Research, 35,* 495–511.

Gutierrez-Clellen, V. F., & Iglesias, A. (1992). Causal coherence in the oral narratives of Spanish-speaking children. *Journal of Speech and Hearing Research, 35,* 363–372.

Hall, K. D., & Yairi, E. (1992). Fundamental frequency, jitter, and shimmer in preschoolers who stutter. *Journal of Speech and Hearing Research, 35,* 1002–1008.

Harris, F. C., & Lahey, B. B. (1978). A method for combining occurrence and nonoccurrence interobserver agreement scores. *Journal of Applied Behavior Analysis, 11,* 523–527.

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis, 10,* 103–116.

Hartmann, D. P. (1984). Assessment strategies. In D. H. Barlow & M. Hersen, *Single case experimental designs: Strategies for studying behavior change* (2nd ed.) (pp. 107–139). New York: Pergamon.

Hartmann, D. P., & Wood, D. D. (1990). Observational methods. In A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), *International handbook of behavior modification and therapy* (2nd ed.) (pp. 107–138). New York: Plenum.

Hawkins, R. P., & Dotson, V. A. (1975). Reliability scores that delude: An Alice in Wonderland trip through the misleading characteristics of interobserver agreement scores in interval recording. In E. Ramp & G. Semb (Eds.), *Behavior analysis: Areas of research and application* (pp. 359–376). Englewood Cliffs, NJ: Prentice-Hall.

Hawkins, R. P., & Fabry, B. D. (1979). Applied behavior analysis and interobserver reliability: A commentary on two articles by Birkimer and Brown. *Journal of Applied Behavior Analysis, 12,* 545–552.

Hillis, J. W. (November, 1991). The perceptual identification of speech characteristics. Microcomputer instructional lab and paper presented at the Annual Convention of the American Speech-Language-Hearing Association, Atlanta, GA.

Hillis, J. W. (1993). Ongoing assessment in the management of stuttering: A clinical perspective. *American Journal of Speech-Language Pathology: A Journal of Clinical Practice, 2,* 24–37.

Hirji, K. F., & Rosove, M. H. (1990). A note on interrater agreement. *Statistics in Medicine, 9,* 835–839.

Hollenbeck, A. R. (1978). Problems of reliability in observational research. In G. P. Sackett (Ed.), *Observing behavior, Volume II, Data collection and analysis methods* (pp. 79–98). Baltimore, MD: University Park Press.

Hopkins, B. L., & Hermann, J. A. (1977). Evaluating interobserver reliability of interval data. *Journal of Applied Behavior Analysis, 10,* 121–126.

Imai, S., & Michi, K. (1992). Articulatory function after resection of the tongue and floor of the mouth: Palotometric and perceptual evaluation. *Journal of Speech and Hearing Research, 35,* 68–78.

Ingham, R. J., Cordes, A. K., & Finn, P. (1993). Time-interval measurement of stuttering: Systematic replication of Ingham, Cordes, and Gow. *Journal of Speech and Hearing Research, 36,* 1168–1176.

Ingham, R. J., Cordes, A. K., & Gow, M. L. (1993). Time-interval measurement of stuttering: Modifying interjudge agreement. *Journal of Speech and Hearing Research, 36,* 305–323.

Johnston, J. M., & Pennypacker, H. S. (1980). *Strategies and tactics of human behavioral research.* Hillsdale, NJ: Lawrence Erlbaum.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527–535.

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis, 10,* 141–150.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings.* New York: Oxford University Press.

Kearns, K. J. (1990). Reliability of procedures and measures. In L. B. Olswang, C. K. Thompson, S. F. Warren, & N. J. Minghetti (Eds.), *Treatment efficacy research in communication disorders* (pp. 79–90). St. Louis, MO: American Speech-Language-Hearing Foundation.

Kelly, M. B. (1977). A review of the observational data-collection and reliability procedures reported in *The Journal of Applied Behavior Analysis. Journal of Applied Behavior Analysis, 10,* 97–101.

Kratochwill, T. R. (1979). Just because it's reliable doesn't mean it's believable: A commentary on two articles by Birkimer and Brown. *Journal of Applied Behavior Analysis, 12,* 553–558.

Kratochwill, T. R., & Wetzel, R. J. (1977). Observer agreement, credibility, and judgment: Some considerations in presenting observer agreement data. *Journal of Applied Behavior Analysis, 10,* 133–140.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36,* 21–40.

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research, 35,* 512–520.

Lahey, M. A., Downey, R. G., & Saal, F. E. (1983). Intraclass correlations: There's more than meets the eye. *Psychological Bulletin, 93,* 586–595.

Light, J., Dattilo, J., English, J., Gutierrez, L., & Hartz, J. (1992). Instructing facilitators to support the communication of people who use augmentative communication systems. *Journal of Speech and Hearing Research, 35,* 865–875.

Martin, R. R., & Haroldson, S. K. (1992). Stuttering and speech naturalness: Audio and audiovisual judgments. *Journal of Speech and Hearing Research, 35,* 521–528.

McGregor, K. K., & Schwartz, R. G. (1992). Converging evidence for underlying phonological representations in a child who misarticulates. *Journal of Speech and Hearing Research, 35,* 596–603.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). New York: MacMillan.

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin, 86,* 376–390.

Morrison, J. A., & Shriberg, L. D. (1992). Articulation testing versus conversational speech sampling. *Journal of Speech and Hearing Research, 35,* 259–273.

Nippold, M. A., Schwarz, I. E., & Undlin, R. A. (1992). Use and understanding of adverbial conjuncts: A developmental study of adolescents and young adults. *Journal of Speech and Hearing Research, 35,* 108–118.

Onslow, M., Adams, R., & Ingham, R. (1992). Reliability of speech naturalness ratings of stuttered speech during treatment. *Journal* of Speech and Hearing Research, 35, 994–1001.

Onslow, M., Hayes, B., Hutchins, L., & Newman, D. (1992). Speech naturalness and prolonged-speech treatments for stuttering: Further variables and data. *Journal of Speech and Hearing Research, 35,* 274–282.

Onslow, M., van Doorn, J., & Newman, D. (1992). Variability of acoustic segment durations after prolonged-speech treatment for stuttering. *Journal of Speech and Hearing Research, 35,* 529–536.

Otomo, K., & Stoel-Gammon, C. (1992). The acquisition of unrounded vowels in English. *Journal of Speech and Hearing Research, 35,* 604–616.

Paul, R., & Jennings, P. (1992). Phonological behavior in toddlers with slow expressive language development. *Journal of Speech and Hearing Research, 35,* 99–107.

Perlman, A. L., Grayhack, J. P., & Booth, B. M. (1992). The relationship of vallecular residue to oral involvement, reduced hyoid elevation, and epiglottic function. *Journal of Speech and Hearing Research, 35,* 734–741.

Perreault, W. D., & Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research, 26,* 135–148.

Postma, A., & Kolk, H. (1992). The effects of noise masking and required accuracy on speech errors, disfluencies, and self-repairs. *Journal of Speech and Hearing Research, 35,* 537–544.

Prins, D., & Hubbard, C. P. (1992). Constancy of interstress intervals in the fluent speech of people who stutter during adaptation trials. *Journal of Speech and Hearing Research, 35,* 799–804.

Purcell, S. L., & Liles, B. Z. (1992). Cohesion repairs in the narratives of normal-language and language-disordered school-age children. *Journal of Speech and Hearing Research, 35,* 354–362.

Pye, C., Wilcox, K. A., & Siren, K. A. (1988). Refining transcriptions: The significance of transcriber "errors." *Journal of Child Language, 15,* 17–37.

Ratner, N. B. (1992). Measurable outcomes of instructions to modify normal parent-child verbal interactions: Implications for indirect stuttering therapy. *Journal of Speech and Hearing Research, 35,* 14–20.

Rice, M. L., Buhr, J., & Oetting, J. B. (1992). Specific-language-impaired children's quick incidental learning of words: The effect of a pause. *Journal of Speech and Hearing Research, 35,* 1040–1048.

Rosenthal, R. (1966). *Experimenter effects in behavior research.* New York: Appleton-Century-Crofts.

Ryan, B. P. (1992). Articulation, language, rate, and fluency characteristics of stuttering and nonstuttering preschool children. *Journal of Speech and Hearing Research, 35,* 333–342.

Sander, E. K. (1961). Reliability of the Iowa Speech Disfluency Test. *Journal of Speech and Hearing Research, Monograph Supplement 7,* 21–30.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Beverly Hills, CA: Sage.

Shriberg, L. D., Kwiatkowski, J., & Hoffman, K. (1984). A procedure for phonetic transcription by consensus (Research note). *Journal of Speech and Hearing Research, 27,* 456–465.

Shriberg, L. D., & Lof, G. L. (1991). Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics & Phonetics, 5,* 225–279.

Suen, H. K. (1988). Agreement, reliability, accuracy, and validity: Toward a clarification. *Behavioral Assessment, 10,* 343–366.

Suen, H. K. (1990). *Principles of test theories.* Hillsdale, NJ: Lawrence Erlbaum.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data.* Hillsdale, NJ: Lawrence Erlbaum.

Sussman, H. M., Hoemeke, K. A., & McAffrey, H. A. (1992). Locus equations as an index of coarticulation for place of articulation distinctions in children. *Journal of Speech and Hearing Research, 35,* 769–781.

Swanson, L. A., Leonard, L. B., & Gandour, J. (1992). Vowel duration in mothers' speech to young children. *Journal of Speech and Hearing Research, 35,* 617–625.

Thal, D. J., & Tobias, S. (1992). Communicative gestures in children with delayed onset of oral expressive vocabulary. *Journal of Speech and Hearing Research, 35,* 1281–1289.

Tompkins, C. A., Boada, R., & McGarry, K. The access and processing of familiar idioms by brain-damaged and normally aging adults. *Journal of Speech and Hearing Research, 35,* 626–637.

Tryon, W. W. (1985a). Introduction and overview. In W. W. Tryon (Ed.), *Behavioral assessment in behavioral medicine* (pp. 1–18). New York: Springer.

Tryon, W. W. (1985b). Measurement of human activity. In W. W. Tryon (Ed.), *Behavioral assessment in behavioral medicine* (pp. 200–256). New York: Springer.

Uebersax, J. S. (1988). Validity inferences from interobserver agreement. *Psychological Bulletin, 104,* 405–416.

Ventry, I. M., & Schiavetti, N. (1986). *Evaluating research in speech pathology and audiology.* New York: MacMillan.

Violette, J., & Swisher, L. (1992). Echolalic responses by a child with autism to four experimental conditions of sociolinguistic input. *Journal of Speech and Hearing Research, 35,* 139–147.

Walker, J. F., Archibald, L. M. D., Cherniak, S. R., & Fish, V. G. (1992). Articulation rate in 3- and 5-year-old children. *Journal of Speech and Hearing Research, 35,* 4–13.

Wasik, B. H. (1989). The systematic observation of children: Rediscovery and advances. *Behavioral Assessment, 11,* 201–217.

Wertz, R. T., & Rosenbek, J. C. (1992). Where the ear fits: A perceptual evaluation of motor speech disorders. *Seminars in Speech and Language, 13,* 39–54.

Weston, A. D., & Shriberg, L. D. (1992). Contextual and linguistic correlates of intelligibility in children with developmental phonological disorders. *Journal of Speech and Hearing Research, 35,* 1316–1332.

Wetherby, A. M., & Rodriguez, G. P. (1992). Measuring communicative intentions in normally developing children during structured and unstructured situations. *Journal of Speech and Hearing Research, 35,* 130–138.

Wildman, B. G., & Erickson, M. T. (1977). Methodological problems in behavioral observation. In J. D. Cone & R. P. Hawkins (Eds.), *Behavioral assessment: New directions in clinical psychology* (pp. 255–273). New York: Brunner/Mazel.

Wood, L. M., Hughes, J., Hayes, K. C., & Wolfe, D. L. (1992). Reliability of labial closure force measurement in normal subjects and patients with CNS disorders. *Journal of Speech and Hearing Research, 35,* 252–258.

Yairi, E., & Ambrose, N. (1992). A longitudinal study of stuttering in children: A preliminary report. *Journal of Speech and Hearing Research, 35,* 755–760.

Young, M. A. (1984). Identification of stuttering and stutterers. In R. F. Curlee & W. H. Perkins (Eds.), *Nature and treatment of stuttering: New directions* (pp. 13–30). San Diego, CA: College-Hill.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103,* 374–378.

Zwirner, P., & Barnes, G. J. (1992). Vocal tract steadiness: A measure of phonatory and upper airway motor control during phonation in dysarthria. *Journal of Speech and Hearing Research, 35,* 761–768.

Received March 20, 1993
Accepted October 29, 1993

Contact author: Anne K. Cordes, Department of Speech and Hearing Sciences, University of California, Santa Barbara, Santa Barbara, CA 93106-7050.