Interobserver Reliability Procedures in Applied Aphasia Research: A Review with Suggestions for Change

Kevin P. Kearns, Ph.D.
University of Kansas Medical Center, Kansas City, Kansas

In his seminal review of the aphasia treatment literature Darley (1972) called for renewed scientific rigor in applied investigations of aphasia. He stated that treatment research must be "more than an interested clinician's artistic and intuitive reporting of change observed in his patients. Reliable quantitative data must be gathered with rigorous objectivity" (p.11). Despite this admonition, aphasiologists have underestimated the complexity involved in obtaining reliable measurement and serious long-range consequences may result. Specifically, compilation of a large number of studies whose apparent statistical or clinical significance is artifactually related to inadequate reliability procedures may lead to the adoption of weak or ineffective therapeutic techniques. Therefore, the purpose of this paper will be to relate applied behavioral methods of reliability assessment to the aphasia treatment literature and to suggest directions for improving reliability assessment in aphasia treatment research.

The term "reliability" refers to the consistency and replicability of measurement (Thorndike and Hagen, 1977) and it is the primary means by which we infer the accuracy of observer-recorded data. According to Hawkins and Farby (1979) there are two general approaches to the reliability issue. The traditional or group study approach is familiar to most aphasiologists. This method involves the administration of standardized tests to evaluate therapeutic effectiveness. When this method is employed, reliability is evaluated prior to the initiation of an investigation. That is, standardization data are generated to show that a test is internally consistent, has high test-retest reliability, and will result in comparable scores when given by qualified examiners. When reasonably high correlation coefficients are reported in standardization data, it is tacitly assumed that test results are independent of the examiner and representative of the patient's actual performance.

The second approach to assessing reliability, the behavior analytic approach, differs markedly from the traditional approach. This method is generally employed in single case experimental studies and it is equally applicable to other treatment investigations in which nonstandardized tests or probes are employed. Standardized tests are not available to measure behaviors such as the productive use of the verbal auxiliary and copula (Kearns and Salmon, 1981) or the confrontation naming ability of anomic individuals (Thompson and Kearns, 1981). As a result, investigators must develop tests to assess treatment effectiveness for the specific behavior under study. Since these measures are not standardized, the experimenter and an independent observer record the occurrence of the target behaviors while the investigation is being conducted and compare their records to determine the level of interobserver agreement (reliability). High percentages of agreement between observers implies that the experimenter accurately recorded the behavior of interest and that improvement in subject responding represents a true treatment effect.

Reliability Procedures Employed in Aphasia Treatment Studies

Let us now shift focus and examine the status of reliability procedures within the applied aphasia literature. A review of the aphasia treatment literature was undertaken in an attempt to evaluate the types of reliability procedures being employed and to determine if aphasiologists have begun to incorporate suggestions from the applied behavioral literature concerning reliability assessment. This survey was not intended to be exhaustive. However, both large and small N studies, planned and retrospective were included (see Appendix). A total of twenty-four aphasia treatment studies were surveyed and the following results were obtained. First, the largest category of studies examined was the "no reliability" category. Surprisingly, fourteen of twenty-four studies (approximately 58%) were included in this group. It is noteworthy that the majority of these studies (9/14) were published between 1970 and 1980 and recent and highly acclaimed investigations, such as those by Basso, Capitani and Vignolo (1975, 1979) were included.

A second finding in this survey was also somewhat surprising. Only six of the twenty-four studies reviewed (25%) used standardized testing as the primary dependent measure. It was anticipated that a larger proportion of aphasia treatment studies would have conformed to the traditional model. Finally, only a small minority of the studies (4) reported reliability data in the absence of standardized testing.

Investigations in this survey were selected without preestablished criteria for inclusion, and these findings may not, therefore, be representative of the approach that every aphasiologist might take to the reliability issue. The large number of studies in the "no reliability" category suggests, however, that aphasiologists have failed to appreciate the importance and complexity of reliable measurement. As Neale and Liebert (1973) (p.86) remind us "...the first and most important aspect of collecting useful information is reliable measurement." The usefulness of our treatment data will remain suspect unless we begin routinely to incorporate procedures which evaluate the reliability of our data collection procedures.

Methodological Issues in Reliability Assessment

Although routine utilization of traditional or interobserver reliability procedures will strengthen the results and conclusions obtained from aphasia treatment studies, the collection of reliability data will not ensure the objectivity of measurement procedures. Even well standardized aphasia tests do not, for example, control sources of experimenter bias which may confound measurement. Bias due to observer expectancy and drift, and witnessing of consequation are difficult to control and even more difficult to detect. O'Leary, Kent and Kanovitch (1975) demonstrated that observer expectancy, particularly in combination with contingent feedback, may influence an experimenter's scoring of patient behaviors. Apparently, observers who anticipate change in subject behaviors are likely to reflect that bias in their scoring. Similarly, observer drift may also confound treatment data (Kent, Kanowitz, O'Leary and Cheiken, 1977). Examiners within the same environment tend to develop and share idiosyncratic scoring habits. For example, co-workers may be reliable with one another in their scoring of the Porch Index of Communicative Ability (PICA) (Porch, 1971) but they may not agree with an objective standard. As these examples demonstrate, test standardization does not guarantee the collection of accurate data. Appropriate control procedures, such as those utilized in the recent V.A. cooperative study (Wertz, et al., 1978), are needed to avoid the confounding effects of experimenter bias when standardized tests are used as the primary measure of treatment effectiveness.

Confounding factors are of equal concern when the behavior analytic approach to reliability assessment is employed (Kazdin, 1977). However, the remainder of this discussion will be concerned with methodological aspects of calculating reliability in behavior analytic studies. High levels of interobserver reliability, like their correlational counterparts in standardized testing, do not ensure the accuracy of observer recorded data. As Hawkins and Dotson (1975) have demonstrated, the point-to-point agreement method may not be sensitive to factors such as observer incompetence, experimental bias or inadequate response definitions. The basic limitation of the point-to-point, percentage agreement method of determining reliability relates to the fact that it is integrally related to rate of responding. (Bijou et al., 1968; Hawkins and Dotson, 1975; Hopkins and Herman, 1977). During periods of low or high rates of correct responding, judges are virtually assured of attaining a high percentage of interobserver agreement on the basis of chance alone. That is, "agreements" on errors inflates reliability coefficients during periods of low rates of correct responding and "agreements" on correct responses inflates the level of reliability obtained during sessions in which a high rate of correct responding is evident.

A brief example may demonstrate these points. Since 80% interobserver reliability is the minimum acceptable level (Kazdin, 1977b) we will assume that an 85% level of agreement was calculated for a hypothetical aphasia study using the standard formula:

$$\text{Reliability} = \frac{\text{Total \# of Agreements}}{\text{Total \# Agreements + Disagreements}} \times 100$$

More important for our present purposes, the level of agreement which would be expected on the basis of chance could be calculated using Hopkins and Herman's (1977) formula:

$$\text{Chance Reliability} = \frac{(O_1 \times O_2) + (N_1 \times N_2) \times 100}{(T)^2}$$

The O_1 and O_2 in the formula designate the number of occurrences (correct responses) recorded by observers 1 and 2 respectively and N_1 and N_2 refer to the number of nonoccurrences or error responses recorded by the observers. The T represents the total number of trials which were observed. In order to calculate the level of chance agreement which would be expected, it is necessary to know the subject's rate of responding. Therefore, further suppose that the 85% reliability level was calculated for a one hundred trial session in which Observer $_1$ recorded 95 correct responses and 5 incorrect and Observer $_2$ recorded 90 correct and 10 incorrect responses. We are now in a position to calculate the overall chance level of agreement as follows:

$$\text{Chance Overall Agreement} = \frac{(95 \times 90) + (5 \times 10) \times 100}{(100)^2} = 86\%$$

As the results of our calculation reveal, the level of chance agreement actually exceeds the 85% agreement level reported from our hypothetical example. Since the level of agreement expected on the basis of chance alone actually exceeds the reported level, data from this study would be considered unreliable and any results and conclusions based on the data would be of questionable validity. As this example demonstrates, reported levels of interobserver reliability must, at a minimum, exceed the level which would be expected on the basis of chance alone.

Although the above example demonstrates the inflationary effects of high rates of responding, it is perhaps worth reiterating that similar results would have been obtained at extremely low rates of responding. In fact, the only rate of responding which does not result in inflated chance agreement levels is when correct and incorrect responses are produced at approximately the same rate (Hawkins and Dotsun, 1975). As Hopkins and Herman (1977) point out, therefore, percentage agreement reliability coefficients are uninterpretable unless the calculable level of chance agreement is known.

Numerous other methods of handling the inflationary effects of rate of responding and chance agreement have been proposed in the literature. These include the occurrence, nonoccurrence, weighted average, and graphic methods (Bijou, Peterson and Auldt, 1968; Hawkins and Dotson, 1975; Hopkins and Herman, 1977; Harris and Lahey, 1978; Birkimer and Brown, 1979 a,b). One other approach, the statistical method, has generated considerable discussion and controversy. It will, therefore, be briefly considered.

The question of how much better than chance does a reliability coefficient need to be has been considered in the applied literature. However, many behavior analytic researchers have eschewed the statistical approach which is implied in this question (Michael, 1974; Baer, 1977). It has been argued that statistical analyses are likely to result in reporting statistically significant but clinically weak variables. In addition, statistical methods may also distract researchers from fine grain data analysis and mask data variability. Despite these objections, however, statistical procedures have been proposed to control for rate of responding and the level of chance agreement. Birkimer and Brown (1979), for example, report what they label an "Easier Way" to resolve the chance agreement problem. Based on probability theory, they have generated significance tables which ultimately led to their "50-10-10 (90) rule." They note that for periods of 50 or more trials, in which the observer disagreement rate is less than or equal to 10% and the rate of correct responding is between 10 and 90%, the level of agreement is not likely due to chance. Essentially, this method proposes a statistical criterion ($p \leq .01$) for rejecting the hypothesis that observer agreement is due to chance.

Again, statistical solutions to the reliability issue represent a considerable departure from the more familiar interobserver agreement methods. Computation of chance and interobserver agreement levels may be preferable to statistical approaches since, as Hawkins and Farby (1979) note, "Statistical sophistication often becomes confused with scientific merit." (548).

In summary, aphasiologists have not given sufficient attention to the reliability of data collection procedures employed in treatment studies. A review of the treatment literature revealed numerous investigations in which neither the traditional nor the behavior analytic approach to reliability assessment procedures were employed. It is imperative, therefore, that some form of reliability assessment be incorporated into future investigations in clinical aphasiology. In addition, control procedures must be adopted to rule out sources of experimenter bias and the deleterious effects of rate of subject responding. Unless such procedures are routinely incorporated into our applied efforts, treatment research in aphasia will be based largely on subjective clinical assessment. And, as Aaron Smith (1972, p.275) has commented, "If clinical assessments and endurance were the primary criteria for usefulness and validity of prevailing dogmas in the history of science, we might still be using leeches to cure patients with stroke..."

APPENDIX

A SURVEY OF RELIABILITY PROCEDURES IN APPLIED APHASIA RESEARCH: STUDIES REVIEWED

Basso, A., Faglioni, P. and Vignolo, L.A. Etude controlee de la reeducation du lange dans l'aphasie: Comparaison entre aphasiques traites et non-traites. Rev. Neurology, 131, 607-614 (1975).

Basso, A., Capitani, E. and Vignolo, L.A. Influence of rehabilitation on language skills in aphasic patients: A controlled study. Archives of Neurology, 36, 190-196 (1979).

Butfield, E. and Zangwill, O. Re-education in Aphasia: A review of 70 cases. J. Neurology, Neurosurgery, Psychiatry, 9, 75-79 (1946).

Deal, J.L. and Deal, L.A. Efficacy of aphasia language rehabilitation: Preliminary results. In R.H. Brookshire (Ed.), Clinical Aphasiology Conference Proceedings, 1978. Minneapolis, MN: BRK Publishers (1978).

Godfrey, C.M. and Douglas, E. The recovery process in aphasia. Canadian Medical Association J., 80, 618-624 (1959).

Goodkin, R., Diller, L. and Shah, N. Training spouses to improve the function of speech of aphasic patients. In B. Lahey (Ed.), The Modification of Language Behavior. Springfield: C.C. Thomas Publishers (1973).

Hagen, C. Communication abilities in hemiplegia: Effect of speech therapy. Archives of Physical Medicine and Rehabilitation, 54, 454-463 (1973).

Holland, A. The usefulness of treatment for aphasia: A seredipitous study. In R.H. Brookshire (Ed.), Clinical Aphasiology Conference Proceedings, 1980. Minneapolis, MN: BRK Publishers (1980).

Holland, A. and Levy, C. Syntactic generalization in aphasia as a function of re-learning an active-declarative sentence. Acta Symbolica, 2-2, 34-41 (1971).

Holland, A. and Sonderman, T.C. Effects of a program based on the Token Test for teaching comprehension skills to aphasics. J. Speech and Hearing Research, 17, 589-598 (1974).

Kearns, K.P. and Salmon, S.J. An experimental analysis of auxiliary and copula verb generalization in aphasia. In submission (1981).

Marks, M., Taylor, M. and Rusk, H.A. Rehabilitation of the aphasic patient: A summary of three years experience in a rehabilitation setting. Archives Physical Medicine and Rehabilitation, 38, 219-226 (1957).

Naeser, M.A. A structured approach to teaching aphasics basic sentence types. British J. of Disordered Communication, 10, 70-76 (1975).

Sarno, M.T., Sands, E. and Shankweiler, D. Long-term assessment of language function in aphasia due to stroke. Archives Physical Medicine and Rehabilitation, 50, 202-206 (1969).

Sarno, M.T., Silverman, M. and Sands, E. Speech therapy and language recovery in severe aphasia. J. Speech and Hearing Research, 13, 607-623 (1970).

Seron, X., Deloshe, G., Bastard, U., Chasscon, G., and Hermand, N. Word finding difficulties and learning transfer in aphasic patients. Cortex, 15, 149-155 (1979).

Shewan, C.M. Facilitating sentence formulation: A case study. J. of Communication Disorders, 9, 197-199 (1976).

Smith, M.D. Operant conditioning of syntax in aphasia. Neuropsychologia, 12, 403-405 (1973).

Thompson, C.K. and Kearns, K.P. An experimental analysis of acquisition, generalization and maintenance of naming behavior in a patient with anomic aphasia. In R.H. Brookshire (Ed.), Clinical Aphasiology Conference Proceedings. Minneapolis, MN: BRK Publishers (1981).

Vignolo, L.A. Evaluation of aphasia and language rehabilitation: A retrospective exploratory study. Cortex, 1, 344-367 (1964).

Wiegel-Crump, C. and Koenigsknecht, R.A. Tapping the lexical store of the adult aphasic: An analysis of improvement made in word retrieval skills. Cortex, 9, 411-418 (1973).

Wepman, J.M. Recovery From Aphasia. New York: Ronald Press (1951).

Wertz, T., Collins, M., Weiss, D., Brookshire, R.H., Friden, T., Kurtzke, J.F., and Pierce, J. Veterans Administration cooperative study on aphasia: Preliminary report on a comparison of individual and group treatment. A paper presented at the annual meeting of the American Association for the Advancement of Science. Washington, D.C. (1978).

West, J.A. Auditory comprehension in aphasic adults: Improvement through training. Archives of Physical Medicine and Rehabilitation, 54, 78-86 (1973).

REFERENCES

Basso, A., Faglioni, P. and Vignolo, L.A. Etude controlee de la reeducation du lange dans l'aphasie: Comparaison entre aphasiques traites et non-traites. *Rev. Neurology*, 131, 607-614, 1974.

Basso, A., Capitani, E. and Vignolo, L.A. Influence of rehabilitation on language skills in aphasic patients: A controlled study. *Archive Neurology*, 36, 190-196, 1979.

Baer, D.M. Reviewer's comment: just because it's reliable doesn't mean that you can use it. *Journal of Applied Behavior Analysis*, 10, 117-120, 1977.

Bkjou, S.W., Peterson, R.F., and Ault, M.H. A method to integrate descriptive and experimental field studies at the level of data and experimental concepts. *Journal of Applied Behavior Analysis*, 1, 175-191, 1968.

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

Birkimer, J.C. and Brown, J.H. Back to basics: Percentage agreement measures are adequate, but there are easier ways. *Journal of Applied Behavior Analysis*, 12, 535-544, 1979b.

Darley, F.L. The efficacy of language rehabilitation in aphasia. *Journal of Speech and Hearing Disorders*, 37, 3-21, 1972.

Harris, F.C. and Lahey, B.B. A method for combining occurrence and non-occurrence interobserver agreement scores. *Journal of Applied Behavior Analysis*, 11, 523-527, 1978.

Hawkins, R.P. and Dotson, V.A. Reliability scores that delude: An Alice in Wonderland trip through the misleading characteristics of inter-observer agreement scores in interval recording. In E. Ramp and G. Semb (Eds.), *Behavior Analysis Areas of Research and Application*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1975.

Hawkins, R.P. and Farby, B.D. Applied behavioral analysis and interobserver reliability: A commentary on two articles by Birkimer and Brown. *Journal of Applied Behavior Analysis*, 12, 545-552, 1979.

Hopkins, B.L. and Herman, R.J. Evaluating interobserver reliability of interval data. *Journal of Applied Behavior Analysis*, 10, 121-126, 1977.

Kazdin, A.E. Artifact, bias and complexity of assessment: the ABC;s of reliability. *Journal of Applied Behavior Analysis*, 10, 141-150, 1977a.

Kazdin, A.E. Methodology of applied behavior analysis. In T.A. Brigham and A.E. Catania (Eds.), *Social and Instructional Process: Foundations and Applications*. New York: Irvington/Naiburg. John Wiley and Sons, 1977b.

Kearns, K.P. and Salmon, S.J. An experimental analysis of auxiliary and copula verb generalization in aphasia. In submission, 1981.

Kent, R.N., Kanowitz, J., O'Leary, K.D., and Cheiken, M. Observer reliability as a function of circumstances of assessment. *Journal of Applied Behavior Analysis*, 10, 317-324, 1977.

Michael, J. Statistical inference for individual organism research. Mixed blessing or curse? *Journal of Applied Behavior Analysis*, 7, 647-653, 1974.

Neale, J.M. and Liebert, R.M. *Science and Behavior*. *An Introduction to Methods of Research*. Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1973.

O'Leary, K.D., Kent, R.N. and Kanowitz, J. Shaping data collection congruent with experimental hypotheses. Journal of Applied Behavior Analysis, 8, 43-51, 1975.

Porch, B.E. The Porch Index of Communicative Ability. Palo Alto, California, 1971.

Smith, A. Replies to two comments on "Objective indices of severity of chronic aphasia..." Journal of Speech and Hearing Disorders, 37, 274-278, 1972.

Thompson, C.K. and Kearns, K.P. An experimental analysis of acquisition, generalization and maintenance of naming behavior in a patient with anomic aphasia. In R.H. Brookshire (Ed.), Clinical Aphasiology Conference Proceedings. Minneapolis, MN: BRK Publishers, 1981.

Thorndike, R.L. and Hagen, E.P. Measurement and Evaluation in Psychology and Education (4th ed.). New York: John Wiley and Sons, 1977.

Wertz, T., Collins, M., Weiss, D., Brookshire, R.H., Friden, T., Kurtzke, J.F., and Pierce, J. Veterans Administration cooperative study on aphasia: Preliminary report on a comparison of individual and group treatment. A paper presented at the annual meeting of the American Association for the Advancement of Science. Washington, D.C., 1978.

DISCUSSION

Q: There would be less chance of agreement with a scoring system which has, say, five codes than one which only has two.

A: That's a good point.

Q: I think that this is an important issue for aphasia research (and any other research for that matter). It would be interesting to do a study that demonstrates what would happen if you did get reliability data versus if you did not and how different the results and conclusions might be.

A: I think the danger is that if you don't obtain reliability data you could get significant results, and, then, as studies accumulate we begin to feel very comfortable about our conclusions and applications. However, an accumulation of a lot of weak studies result in the impression that we have strong data even though we really don't have data that we can trust. Secondly, as Audrey Holland pointed out last year, as we move away from the "big" efficacy question of "Does aphasia therapy work?" and move towards more specific treatment questions, we'll be moving away from the use of standardized tests to look at these specific issues. As we do that, I think we'll need to be more aware of the reliability of our measurement tools.

Q: What do you have to say to the clinician who is working in a fairly isolated setting who wants to do clinical research and who would like to do it well, and would like to get reliability?

A: I'd suggest the use of videotaping when it is available and audiotaping when it is not. One can usually, for example, bring in a naive observer. Although we often use trained observers, perhaps we don't always need to use other speech pathologist as observers.

Q: Could you give us some guidelines for the use of single case designs in aphasia research?

A: With regard to single case designs and the reliability issue, there are two simple points I've tried to make. First of all, make sure you obtain reliability for observer recorded data. Second, the use of standardized tests may not be sufficient to insure the collection of accurate data because of the types of bias which I have mentioned. Similarly, in single subject studies we should first of all, calculate and report the level of chance agreement. In addition, I would suggest we all become more familiar with the procedures I have mentioned; for example, the weighted averaging approach. What this method does is proportion out, on the basis of rate of responding, those types of responses which are most likely to contribute to high levels of chance agreement.

Q: Sometimes there are practical problems in gathering ideal reliability data. I wonder how satisfied you would be if clinicians at least insured that their scoring system was well defined, and made explicit in print so that we all knew exactly what the scoring system was. Do you think that this would be a good compromise for some of the practical problems we might run into?

A: My personal opinion is no. I think that it is a good start but, as the review by Hawkins and Dotson (1975) demonstrates, even when we appear to have pretty good definitions or good reliability, there can be a low level of accuracy. What it comes down to is that we don't have an objective standard as a basis of comparison. Making our definitions operational is a good start, but without some form of reliability I wouldn't be comfortable with this as a solution.

Q: Are there people who have looked at agreement on ordinal versus interval scales?

A: Observations have been recorded during specified intervals within an observation period, but I'm not familiar with studies which have specifically compared interval versus ordinal level data.

Q: The issue of simple versus complex scoring systems becomes important when we begin to assess the effects of treatment as it relates to the outside world. We're going to have people go out to the outside world and make judgments about our treatment and we'll probably have to use non-professionals. We will have to develop scoring systems which are reasonably straightforward. I think that the literature on social validation shows that when we go into the patient's world the rating system becomes crucial.

A: We're going to have to get into the area of social validation and view it as another form of reliability.