JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1977, 10, 133-139 NUMBER 1 (SPRING) 1977

OBSERVER AGREEMENT, CREDIBILITY, AND JUDGMENT: SOME CONSIDERATIONS IN PRESENTING OBSERVER AGREEMENT DATA

THOMAS R. KRATOCHWILL[1] AND RALPH J. WETZEL

THE UNIVERSITY OF ARIZONA

Graphical and statistical indices employed to represent observer agreement in interval recording are described as "judgmental aids", stimuli to which the researcher and scientific community must respond when viewing observer agreement data. The advantages and limitations of plotting calibrating observer agreement data and reporting conventional statistical aids are discussed in the context of their utility for researchers and research consumers of applied behavior analysis. It is argued that plotting calibrating observer data is a useful supplement to statistical aids for researchers but is of only limited utility for research consumers. Alternatives to conventional per cent agreement statistics for research consumers include reporting special agreement estimates (*e.g.*, per cent occurrence agreement and nonoccurrence agreement) and correlational statistics (*e.g.*, Kappa and Phi).

DESCRIPTORS: observational data, methodology, observer bias, observer training, reliability, validity, experimenter calculations

Applied behavior analysis emphasizes collecting human behavioral data by human observers in naturalistic settings. To establish credibility of observational data, one or more calibrating observers usually verify the primary observer's data. To aid the scientific community in its judgment of the acceptability of findings, an index of observer agreement is presented along with the findings themselves. This index simplifies and relates statistically a complex series of data points from two or more observers. The derived statistic, whether per cent or correlational, functions as a "judgmental aid" (Michael, 1974). If it falls within conventional limits, the scientific community regards the primary observer's data as "basic" and "appropriate" for subsequent analysis. In this context, the index establishes the tolerance extended to measurement error.

Recently, the extent to which certain forms of statistical observer agreement aids protect against misrepresentation of error has been questioned (Hartmann, 1977; Hawkins and Dotson, 1975; Johnson and Bolstad, 1973; O'Leary and Kent, 1973). Hawkins and Dotson (1975) demonstrated that when the formula $\frac{\text{agreements}}{\text{agreements} + \text{disagreements}} \times 100$ is used to calculate an agreement coefficient from interval recording data, the resulting scores: (a) may be highly insensitive to the adequacy of response definitions, (b) may misrepresent observer competence, and (c) cannot always be relied on to assess the "believability" of the experimental effect. Of the major problems with this conventional method, the implication that it cannot always provide a safeguard against misrepresentation of experimental findings is perhaps the most salient. This is an important issue because applied behavioral researchers have conventionally provided the scientific community with data from one observer only and have relied on these data to represent experimental effects.

These and other considerations have prompted new recommendations for applied behavioral researchers. Based on problems with the conventional per cent agreement statistic, Hawkins and Dotson (1975) recommended that calibrating

[1]The authors thank Michael J. Subkoviak and Sidney W. Bijou, for their comments on the manuscript. Reprints may be obtained from Thomas R. Kratochwill, Department of Educational Psychology, The University of Arizona, Tucson, Arizona 85721.

observer data should be plotted along with data of the primary observer. It was argued that this assists both researcher and readers of the experimental report (*i.e.,* the consumers) to interpret more adequately the credibility of statistical observer agreement indices and to detect certain threats of internal validity (*e.g.,* observer bias).

Another recommendation calls for more sophisticated examination of the properties of the statistical scores themselves and the reporting of special observer agreement statistical aids for consumers. These suggestions involved variations on conventional per cent agreements scores (*e.g.,* occurrence and nonoccurrence) (Hawkins and Dotson, 1975) and the introduction of relatively new, for applied behavior analysts, correlation-like statistics (*e.g.,* Kappa and Phi) (Hartmann, 1977). While left unstated, statistical indices are presumed to be of benefit to both researcher and consumer.

Aids for judging observer agreement must be evaluated by the ease, efficiency, and degree of accuracy with which they permit the researcher and scientific community to judge. Observer agreement can be represented statistically or graphically by plotting both the primary and calibrating observer's data. Hawkins and Dotson (1975) suggested that graphical representations would aid not only the researcher but research readers, who could then judge for themselves the adequacy of the statistical aid's representation of observer correspondence. While we agree that both graphical and statistical aids can assist in making judgments about observer agreement, we believe that both graphical and statistical aids best serve the researcher, but that statistical aids best serve the needs of the research consumer.

The primary benefits of plotting data occur before publication of the scientific report. If plotting both sets of data reveals to researchers that only one observer recorded an experimental effect, the data should not be published, acceptable statistical representations of observer agreement notwithstanding. Clearly, the situation requires experimenters to redesign the observation technology. Plotting the calibrating observer's data makes the judgment of the acceptability of the primary observer's data individual to researchers and is in keeping with the reliance on graphical aids so pervasive in applied behavioral research.

There are several limitations in providing graphical aids (*i.e.,* both sets of data) for observer agreement judgment for research consumers. First, it will take considerable expense on the part of researchers to report such data. Second, there is already growing concern that exclusive use of visual analysis of graphical displays may lead to misrepresentation of experimental effects (*cf.* Jones, Weinrott, and Vaught, Note 1). Plotting data from two observers would further complicate the already complex conventional process of evaluating experimental effects through visual analysis.

While applied behavioral researchers have preferred to communicate data via graphical displays, and have avoided the use of statistical tests to evaluate experimental effects, they have relied almost exclusively on a statistical index to communicate observer agreement. In one respect, this has been an advantage for the scientific community. Reporting a statistical score provides a more objective judgment of the credibility of experimental data. Such objective criteria can be standardized and consumers can react to them with perfect reliability. For the researcher, statistical measures of observer agreement are also objective, in that anyone can calculate them and obtain the same result.

However, because most statistical indices of observer agreement are abbreviations, they have achieved their simplifying effects with some loss of information. The combined effects of this abbreviating process and the possible lack of understanding of the properties of some conventional scores have threatened the data base of applied behavior analysis. This paper suggests that the choice of both graphical and statistical aids must be dictated by the nature of the experiment, rather than depending on convention. First, we describe the conditions under which graphical aids provide an additional source of information

that facilitates making data respectable and eventually, publicly specifiable. Second, we review some statistical aids that can increase the credibility of observational data.

Graphical Judgmental Aids

The benefits of plotting data extend beyond those of the per cent agreement index to offer researchers some additional judgmental aids over all statistical scores. Plotting the calibrating observers' data allows the researcher to detect critical behaviors that might otherwise go unnoticed, to maintain contact with the data, to evaluate absolute differences between the two observers, to detect certain threats to internal validity, and requires relatively little sophistication in statistics to evaluate any observer differences.

Detection of behavior. Clearly, one major purpose of having two or more observers is to ensure detection of behavior. We have encountered situations in which the calibrating observer detected extremely important behaviors missed by the primary observer. For example, autistic children frequently exhibit physically self-destructive behaviors. Depending on the type of recording technology used and the method of calculating agreement, a relatively high observer agreement statistic can be obtained after a "successful" program has been established and an "applied criterion" of significance (*cf.* Risley, 1970) is achieved (in this case, zero occurrence of the target behavior). A calibrating observer can detect target behaviors when the primary observer did not. Plotting both sets of data may be important, not only to detect behavior and establish its covariation with other behaviors or stimuli in the environment, but also to note temporal relationships to nonoccurrences over the duration of the treatment program. Such data displays would also allow determination of why certain behaviors are maintained despite the absence of data from the primary observer. This additional judgmental aid for behavioral occurrence allows researchers and perhaps under some conditions, the reader of the experimental report, to view the interrelationships of behavior with other behaviors displayed graphically.

Staying in touch with the data. Data presented by Hawkins and Dotson (1975) raised the possibility that some applied behavioral researchers have not understood the properties of the statistical aids used in reporting observer agreement. Such a finding gives some credibility to the argument that too heavy reliance on "statistics" can lead researchers away from basic contact with their data (*cf.* Michael, 1974). Aside from basic tabulation of raw data, plotting a calibrating observer's data involves the least amount of transformation of the observer's record forms. As a "stimulus simplifying technique" it provides a point-to-point relationship between data sets, thereby allowing visual examination of agreement. The graphical index complements a statistical index and serves as a safeguard against misrepresentation through reliance on only statistical scores.

Evaluation of absolute differences. Plotting calibrating observer data allows researchers to evaluate absolute differences between data records, and not just differences relative to total frequency or duration. For example, the conventional per cent agreement estimate and other statistical aids allow no easy visual evaluation of differences in base rates. Absolute differences can be large and may be of significance to the researcher. Consider a situation in which during baseline, a relatively high-frequency behavior is observed. Observer 1 reports 19 scored intervals and Observer 2 reports 10. During the treatment phase, differences in recorded occurrences are three and six for the two observers, respectively. While a statistical index could show an acceptable agreement level, the researcher would be unable to observe graphically the differences of nine and three, respectively.

Detection of threats to internal validity. It has not been customary for applied behavioral researchers to establish observer agreement checks on every observation session. This undoubtedly reflects availability of additional observers, cost involved in their training, and the complexity of

behavior being observed. At times, researchers may obtain useful information by plotting calibrating observer data to determine the relationship between check and noncheck sessions. If the researcher found that the primary observer consistently reported more (or less) of a target behavior on check sessions relative to noncheck sessions, the hypothesis of bias could be entertained (Hawkins and Dotson, 1975).

There has been increased attention to observer bias (Hersen and Barlow, 1976; Johnson and Lotitz, 1974; McNamara and MacDonough, 1972; Reid, 1970; Scott, Burton, and Yarrow, 1967; Kass and O'Leary, Note 2) and related issues of observer feedback and drift (Hanley, 1970; Kent, O'Leary, Diament, and Dietz, 1974; O'Leary, Kent, and Kanowitz, 1975; Patterson, 1969). However, bias may occur regardless of the level of observer agreement and may similarly reflect that both observers were biased (O'Leary *et al.,* 1975). Furthermore, the response definition of the two observers could "drift" (*cf.* Hanley, 1970; O'Leary and Kent, 1973), causing both to maintain high levels of agreement but leaving a deterioration in original response definitions over the course of the study. When *both* observers are biased and drift, plotting both data sets will not serve a useful detection function.

Some authors have suggested that a third observer could be employed to detect observer drift (Bijou, Peterson, Harris, Allen, and Johnson, 1969; Hanley, 1970). Occasional checks with two regular observers would assist in determining whether either primary or calibrating observers had drifted. O'Leary and Kent (1973) noted that a third *group* of observers could be trained after several weeks of data gathering by regular observers. If comparison with regular observers demonstrated no systematic differences over the course of the experiment, it could be concluded that drift and/or bias did not occur. Plotting the third observer's (or third group of observers') data could facilitate detection of this possible threat to interval validity. A possible difficulty with the procedure is that even a third or fourth observer could be biased. There are no clear guidelines for how many such checks against bias should be employed. However, once the researcher determines the source of error, the observers could be retrained or the observational system could be redesigned.

Sophistication in statistics. A graphical representation of observer agreement data demands little sophistication in statistics. Visual examination of a graphical display of two observers' data will generally allow easy discrimination of disagreement. Researchers working with paraprofessionals, clients involved in therapy, or even children, will find that such individuals easily comprehend fully the meaning of observer agreement. Furthermore, researchers could easily establish guidelines for the amount of deviation between two observers that will be tolerated. Such tolerance could easily be "measured" rather than calculated. In addition, individuals working in applied settings could easily construct agreement graphs.

Statistical Judgmental Aids

Statistical indices have promoted ease of judgment. Yet, important judgmental information is lost in the process of abbreviation. Also, the type of information lost varies as a function of the abbreviating process (*i.e.,* the method of calculation). The major problem with the conventional per cent agreement statistic for interval recording is that it does not adequately take into account chance agreements between two observers. A thorough discussion of this problem can be found in several sources (*e.g.,* Gelfand and Hartmann, 1975; Hawkins and Dotson, 1975; Hartmann, 1977; Johnson and Bolstad, 1973). Because of this factor, some special statistical agreement indices have been proposed. These scores include per cent occurrence agreement and nonoccurrence agreement, an average of these two scores, which we will call "average agreement", and special "correlation-like" coefficients.

Special percentage scores. To circumvent problems associated with conventional per cent

agreement indices, some writers have proposed that per cent occurrence agreement and nonoccurrence agreement scores could be reported (*e.g.,* Bijou *et al.,* 1969; Hawkins and Dotson, 1975; Hartmann, 1977).[2] Both occurrence agreement and nonoccurrence agreement are derived, in part, from the conventional per cent agreement formula. These scores differ in the information they use from the scoring intervals. In occurrence agreement, only intervals in which both observers recorded the presence of a behavior are scored as an agreement. Disagreements are scored in the same manner as those in the conventional method. Nonoccurrence agreement scores reflect the situation in which both observers agree on the nonoccurrence of some behavior. Disagreements are scored when one observer records the presence of a behavior and the other records its absence.

The primary basis for using these scores is that they address the issue of chance agreements that can occur at varying rates of occurrence of target behaviors (Johnson and Bolstad, 1973). Both per cent occurrence and nonoccurrence agreement can provide a great deal of information if used under conditions dictated by the nature of the data (Bijou *et al.,* 1969). Occurrence agreement should be used in a program where there is a very low rate of behavioral occurrence. Since nonoccurrence agreement can inflate the conventional per cent agreement estimate, it should not be reported under such conditions. If behavior is occurring at a very high rate, occurrence agreement can produce "high" conventional per cent agreement estimates and should not be included in calculations. In such cases, per cent nonoccurrence agreement should be calculated. With an understanding of chance agreement (see also Hartmann, 1977) and methods of calculating occurrence and nonoccurrence, these scores will provide important safeguards in reporting agreement data from interval recording. Their advantages appear to be the ease of calculation and the ease with which researchers can conceptualize information used in the scores.

Currently, there are some potential limitations in using occurrence and nonoccurrence agreement scores. Of greatest concern are the conditions under which these scores should be reported and the related issue of what level of agreement index is acceptable. We concur with Hartmann's (1977) observation that it may not be an easy task to decide under which conditions these scores should be calculated and reported when differential rates of behavior occur during an experiment. With a great deal of variability in the data, this problem is compounded.

Another problem with these scores is that the amount of information that will need to be presented in experimental reports will greatly add to the complexity of the judgmental process. For example, while Hawkins and Dotson (1975) suggested that per cent occurrence and nonoccurrence could be presented in combination, they were uncertain whether the combination should involve simply reporting the scores and/or presenting a mean of the two scores. This "average per cent agreement" score reputedly reduces the problem of dealing with variable behavior frequencies, but is not completely free of this problem under all conditions.[3] Some authors have

[2]Hartmann's (1977) effective percentage agreement statistics for occurrence and nonoccurrence are equivalent to our agreement occurrence and agreement nonoccurrence, respectively, and Hawkins and Dotson's S-I and U-I, respectively.

[3]One alternative to the "average agreement" procedure would be to establish some "rules of thumb" for acceptable occurrence and nonoccurrence agreement scores. For example, if the two observers reported more than 80% occurrence, nonoccurrence agreement could be computed and reported. Where the total number of recording intervals is 100 (N=100), a chance agreement for nonoccurrence would be 4% (probability of the first observer reporting a nonoccurrence times the probability of the second observer reporting a nonoccurrence). A nonoccurrence agreement of 75% would indicate considerable nonrandom agreement between observers. Similarly, occurrence agreement could be computed and reported when the behavior rate is less than 20%. In this case, a chance agreement for occurrence would be 4%. At an intermediate rate of behavior (*i.e.,* when behavior is occurring 50% of the time), nonoccurrence would not add much information, since chance levels are rela-

suggested that base-rate chance agreements should be computed and that differences between obtained agreement and chance agreement should be reported (Hersen and Barlow, 1976; Johnson and Bolstad, 1973). In general, if researchers decide to employ per cent occurrence and nonoccurrence agreement scores, research consumers will face a more complex judgmental process when reading scientific reports.

Special correlational statistics. Kappa and Phi statistics have been proposed as options for researchers using interval recording observational methods (Gelfand and Hartmann, 1975; Hartmann, 1977).[4] These scores, also designed to deal with chance agreements, employ all the data from the scoring intervals. Hartmann (1977) has already described the many advantages and relatively few disadvantages of these scores. The primary advantage that we perceive in the use of these scores is that only one score will have to be reported. This relative ease of judgment, coupled with the fact that these scores can be reported over varying levels of behavior change during an experiment, make them attractive.

The chief disadvantage of these scores is that their "novel" feature could cause investigators to employ them to the exclusion of simpler statistical aids that could adequately represent observer agreement. Another disadvantage is that they do require relatively greater sophistication in statistics. It remains to be seen whether they will be accepted in applied behavioral research with the current concerns on more formal reliance on statistics in general (Michael, 1974).

SUMMARY AND RECOMMENDATIONS

Applied behavioral researchers have conventionally relied on statistical judgmental aids to present observer agreement data.[5] Problems in conventional per cent agreement statistics suggest that new methods of reporting observer agreement data be used. Plotting the calibrating observer's data is useful to the researcher to ensure credibility of observational data, but requires a degree of subjectivity that could be generally unacceptable to research consumers. The reporting of special agreement estimates (per cent occurrence agreement and nonoccurrence agreement) and correlational statistics (Kappa and Phi) can greatly improve statistical judgmental aids. Correlational statistics decrease subjectivity and the degree of inference in judging agreement scores over the per cent statistics. Correlational sttaistics should be given greater attention in applied behavioral research.

To understand better the conditions under which per cent and correlational statistics promote accurate decision-making regarding the credibility of observational data, a series of simulated data series needs to be evaluated. Such a series would hopefully vary all the parameters that indicate their limitations under conditions faced by applied behavioral researchers. The current development in more refined statistical aids should be perceived as advancing the science of applied behavioral research.

tively low (*i.e.,* 25% in this case). Under these conditions, total agreement could be more easily interpreted.

[4]The reader is referred to Hartmann (1977) in this series for the method of calculation of Kappa and Phi and for a description of their statistical properties.

[5]Observer "reliability" adds an element of confusion to other terms used in conventional psychometric theory (*cf.* Johnson and Bolstad, 1973). Furthermore, it is not "reliability" in the traditional measurement sense (*cf.* Herbert and Attridge, 1975). Given these considerations, the recent work of Hawkins and Dotson (1975), and our suggestions regarding situations where plotting data from calibrating observers may be useful, we would argue that the term "observer agreement" more adequately handles the representation of observer agreement data, whether graphical or statistical.

REFERENCE NOTES

1. Jones, R. R., Weinrott, M., and Vaught, R. S. *Visual versus statistical inference in operant research.* Paper presented as a part of a symposium on the use of statistics in N=1 research at the annual meeting of the American Psychological Association, Chicago, Illinois, September 1975.
2. Kass, R. E. and O'Leary, K. D. *The effects of observer bias in field-experimental settings.* Paper

presented in a symposium, Behavior Analysis in Education, University of Kansas, Lawrence, April 1970.

REFERENCES

Bijou, S. W., Peterson, R. F., Harris, F. R., Allen, K. E., and Johnson, M. S. Methodology for experimental studies of young children in natural settings. *Psychological Record,* 1969, **19,** 177-210.

Gelfand, D. M. and Hartmann, D. P. *Child behavior analysis and therapy.* New York: Pergamon Press, 1975.

Hanley, E. M. Review of research involving applied behavior analysis in the classroom. *Review of Educational Research,* 1970, **40,** 597-625.

Hartmann, D. P. Notes on methodology: on choosing an interobserver reliability estimate. *Journal of Applied Behavior Analysis,* 1977, **10,** 103-116.

Hawkins, R. P. and Dotson, V. A. Reliability scores that delude: an Alice in Wonderland trip through the misleading characteristics of interobserver 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, 1975. Pp. 359-376.

Herbert, J. and Attridge, C. A guide for developers and users of observational systems and manuals. *American Educational Research Journal,* 1975, **12,** 1-20.

Hersen, M. and Barlow, D. H. *Single case experimental designs. Strategies for studying behavior change in the individual.* New York: Pergamon Press, 1976.

Johnson, S. M. and Bolstad, O. D. Methodological issues in naturalistic observation: some problems and solutions for field research. In L. A. Hamerlynck, L. C. Handy, and E. J. Mash (Eds.), *Behavior change: methodology, concepts, and its practice.* Champaign, Illinois: Research Press, 1973. Pp. 7-67.

Johnson, S. M. and Lobitz, G. K. Parental manipulaion of child behavior in home observations: a methodological concern. *Journal of Applied Behavior Analysis,* 1974, **7,** 23-31.

Kent, R. N., O'Leary, K. D., Diament, C., and Dietz, A. Expectation biases in observational evaluation of therapeutic change. *Journal of Consulting and Clinical Psychology,* 1974, **42,** 774-780.

McNamara, J. R. and MacDonough, T. S. Some methodological considerations in the design and implementation of behavior therapy research. *Behavior Therapy,* 1972, **3,** 361-378.

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

O'Leary, K. D. and Kent, R. N. Behavior modification for social action: research tactics and problems. In L. A. Hamerlynck, L. C. Handy, and E. J. Mash (Eds.), *Behavior change: methodology, concepts, and practice.* Champaign, Illinois: Research Press, 1973. Pp. 69-96.

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

Patterson, G. R. A community mental health program for children. In L. A. Hamerlynck, P. O. Davidson, and L. E. Acker (Eds.), *Behavior modification and ideal mental health services.* Calgary, Canada: University of Calgary Press, 1969. Pp. 130-179.

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

Risley, T. R. Behavior modification: an experimental-therapeutic endeavor. In L. A. Hamerlynck, P. O. Davison, and L. E. Acker (Eds.), *Behavior modification and ideal mental health services.* Calgary, Canada: University of Calgary Press, 1970. Pp. 103-127.

Scott, P. M., Burton, R. V., and Yarrow, M. R. Social reinforcement under natural conditions. *Child Development,* 1967, **38,** 53-63.

Received 19 June 1975.
(Final acceptance 15 May 1976.)