JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1979, 12, 523-533 NUMBER 4 (WINTER 1979)

A GRAPHICAL JUDGMENTAL AID WHICH SUMMARIZES OBTAINED AND CHANCE RELIABILITY DATA AND HELPS ASSESS THE BELIEVABILITY OF EXPERIMENTAL EFFECTS

JOHN C. BIRKIMER AND JOSEPH H. BROWN

UNIVERSITY OF LOUISVILLE

Interval by interval reliability has been criticized for "inflating" observer agreement when target behavior rates are very low or very high. Scored interval reliability and its converse, unscored interval reliability, however, vary as target behavior rates vary when observer disagreement rates are constant. These problems, along with the existence of "chance" values of each reliability which also vary as a function of response rate, may cause researchers and consumers difficulty in interpreting observer agreement measures. Because each of these reliabilities essentially compares observer disagreements to a different base, it is suggested that the disagreement rate itself be the first measure of agreement examined, and its magnitude relative to occurrence and to nonoccurrence agreements then be considered. This is easily done via a graphic presentation of the disagreement range as a bandwidth around reported rates of target behavior. Such a graphic presentation summarizes all the information collected during reliability assessments and permits visual determination of each of the three reliabilities. In addition, graphing the "chance" disagreement range around the bandwidth permits easy determination of whether or not true observer agreement has likely been demonstrated. Finally, the limits of the disagreement bandwidth help assess the believability of claimed experimental effects: those leaving no overlap between disagreement ranges are probably believable, others are not.

DESCRIPTORS: chance agreement, chance reliability, internal validity, interobserver agreement, observational data, observational technology, percentage agreement, reliability

As Kelly (1977) has indicated, research in applied behavior analysis generally produces either permanent-product data, mechanically collected data, or observational data, with the last by far the most common. When observational data are collected, a human observer watches the target individual and records instances of the behavior of interest. Sometimes simple counting of the target behavior occurs, but frequently, either interval recording, time-sampling, or trial scoring is done instead. With interval recording, the entire experimental session is divided into many brief time intervals and the observer records whether or not the behavior of interest occurs during each interval. With time-sampling, the observer either records for only some of the possible intervals during a session or records whether or not behavior is occurring at each of a prespecified subset of moments during a session (Powell, Martindale, and Kulp, 1975). With trial scoring, the observer scores "right" or "wrong" with regard to each of the subject's responses to stimulus materials.

To attempt to ensure that the data collected by the observer are similar to those that would be obtained by other competent observers, researchers arrange reliability checks. Reliability checks involve having a second observer independently record the same behavior of the same target individual through the same experimental session and then comparing the records generated by the two observers. This paper deals with several frequently used statistical procedures for summarizing the results of reliability checks. Other authors have discussed sources of bias and unreliability among observers, and procedures

Reprints of this paper are available from John C. Birkimer, Department of Psychology, University of Louisville, Louisville, Kentucky 40208.

to detect such biases and remedy them, some of this under the rubric of "generalizability theory" (Cone, 1977; Kazdin, 1977). The thrust of the current paper is, instead, toward clarifying problems which are implicit in the current commonly used procedures and toward recommending a graphical procedure which solves those problems and proves valuable as an additional tool for assessing the internal validity of behavioral research, for detecting claimed experimental effects which are not believable.

One simple reliability percentage sometimes calculated, total reliability, is simply the smaller reported frequency of occurrence of target behavior divided by the larger, the proportion then multiplied by 100. This is the only generally used agreement statistic when simple counting is the recording method, and has been correctly criticized by Hawkins and Dotson (1975) as measuring agreement on total frequencies but not on individual occurrences of target behavior. Because substantial agreement on individual occurrences is fundamental to accurate measurement, only procedures focusing on individual intervals, moments, or trials will be discussed further.

The first such procedure has been referred to as moment by moment reliability, as point by point reliability, as percentage agreement (Hartmann, 1977), and as interval by interval (I × I) reliability (Hawkins and Dotson, 1975). With this procedure the two observers' records are compared interval by interval, moment by moment, or trial by trial. The number of occasions on which the observers agree is counted, with an agreement being counted each time they agree the behavior occurred and each time they agree that it did not. The number of agreements is then divided by the total number of occasions in which they agreed and disagreed. (With this measure, the denominator equals the total number of recording occasions.) The ratio of agreements to agreements plus disagreements, when multiplied by 100, yields a percentage which is the reliability or agreement estimate. I × I reliability has been criticized by Hawkins and Dotson (1975) and Kazdin (1975), among others, with the argument that it inflates the percentage agreement estimate when response rates are low by including many cases of agreement which involve the observers agreeing the behavior has not occurred.

The second such reliability calculating procedure has been referred to as occurrence reliability, as effective percentage agreement (Hartmann, 1977), and as scored interval (S-I) reliability (Hawkins and Dotson, 1975). With this procedure, again agreements are divided by the sum of agreements plus disagreements and the ratio multiplied by 100 to yield a percentage reliability or agreement figure. With S-I reliability, however, the definition of agreement is restricted to those occasions on which the observers agree the behavior did occur. The procedure has been recommended by Hawkins and Dotson (1975) and Kazdin (1975), among others, as avoiding the inflationary effects of agreements on nonoccurrences.

The third reliability procedure sometimes used in these situations has been referred to as nonoccurrence reliability, as effective percentage agreement on nonoccurrences (Hartmann, 1977) and as unscored interval (U-I) reliability (Hawkins and Dotson, 1975), and is simply S-I reliability for agreements on nonoccurrences. Hawkins and Dotson suggested that instead of using I × I reliability, researchers should determine S-I and U-I reliabilities, reporting either each reliability or the arithmetic mean of the two. Others have made similar suggestions (Baer, 1977; Hartmann, 1977).

A PROBLEM WITH S-I AND U-I RELIABILITIES

A substantial problem exists, however, with S-I and U-I reliabilities. Observers must record on each recording occasion, scoring the presence or absence of target behavior. If the rate[1] of

[1]The use of the term "rate" here and throughout is a convenience, referring to a *percentage* of intervals,

Table 1

Effect of varying rate of behavior with constant disagreement rate on I × I, S-I, and U-I reliabilities.

Behavior Rate	*Occurrence Agreements*	*Nonoccurrence Agreements*	*Disagreements*	*I × I Reliab.*	*S-I Reliab.*	*U-I Reliab.*
90%	85%	5%	10%	90%	89%	33%
50%	45%	45%	10%	90%	82%	82%
10%	5%	85%	10%	90%	33%	89%

target behavior varies and observers' disagreement rate remains constant, these two reliabilities will vary, as Table 1 shows. In the table, disagreements are taken equally from agreements and disagreements, but the following holds, however they are distributed. I × I reliability remains constant, since it is algebraically equivalent to 100% minus the constant disagreement rate. As the target behavior rate decreases, S-I reliability divides the decreasing number of occurrence agreements by the sum of those agreements plus the constant number of disagreements, causing the reliability statistic to decrease. The opposite effect is caused similarly for U-I reliability. Only if the disagreement rate decreases proportionately to the decreasing behavior rate will S-I reliability remain constant. Such a decrease could be expected only if observers had difficulty agreeing when behavior occurred but no difficulty agreeing when it did not, a distinction rarely possible to make from typical reliability data.

Fluctuations in S-I and U-I reliabilities resulting from changes in target behavior rate, with I × I reliabilities remaining constant, could be troublesome or puzzling to researchers and, with quite low or high behavior rates, lead to obtained reliability percentages below the commonly held "acceptable" 80% to 90% range, perhaps discouraging research on such behaviors.

moments, or trials, or a percentage of agreements on occurrences, on nonoccurrences, or a percentage of disagreements. Wording difficulties and potential confusion with the reliability percentages led us to use "rate" this way.

"CHANCE" RELIABILITIES: A PROBLEM WITH I × I, S-I, AND U-I RELIABILITIES

Hopkins and Hermann (1977), among others, have indicated that for each of these three reliability calculation procedures there are "chance" values obtainable when no true interobserver agreement exists. With observers reporting some rate of target behavior, but reporting randomly, the probabilities of agreements on occurrences, agreements on nonoccurrences, and disagreements are easily obtained (Hopkins and Hermann, 1977) and determine the "chance" values of I × I, S-I, and U-I reliabilities.

If observer 1 reports O_1 occurrences of target behavior (and N_1 nonoccurrences) while observer 2 reports O_2 occurrences and N_2 nonoccurrences, over T observation occasions, but observers report randomly, then the chance probabilities are calculated as:

$$p(\text{chance agreements on occurrences}) = \frac{O_1 \cdot O_2}{T^2}$$

$$p(\text{chance agreements on nonoccurrences}) = \frac{N_1 \cdot N_2}{T^2}$$

$$p(\text{chance disagreements}) = \frac{O_1 \cdot N_2 + N_1 \cdot O_2}{T^2}$$

and

$$\text{I} \times \text{I chance} = \frac{O_1 \cdot O_2 + N_1 \cdot N_2}{T^2}\,(100)$$

$$\text{S-I chance} = \frac{O_1 \cdot O_2}{O_1 \cdot O_2 + O_1 \cdot N_2 + N_1 \cdot O_2}(100)$$

$$\text{U-I chance} = \frac{N_1 \cdot N_2}{N_1 \cdot N_2 + O_1 \cdot N_2 + N_1 \cdot O_2}(100)$$

Our denominators for S-I chance and U-I chance differ from Hopkins and Hermann's (1977), since their use of T^2 for S-I and U-I is inconsistent with the definitions of these reliabilities.

These "chance" reliabilities explain why Hawkins and Dotson (1975) obtained substantial I × I reliabilities when observers were given different definitions of target behavior or when one observer's records were marked as showing 100% occurrences, results which led those authors to conclude I × I reliabilities were uninterpretable.

Hopkins and Hermann (1977) showed that calculating the chance values for each of the three reliabilities makes each obtained reliability interpretable. Each of the "chance" reliabilities, however, varies with the rate of target behavior, essentially as Hopkins and Hermann's figures indicate (with correction for our more appropriate S-I and U-I denominators). Thus, as behavior rates vary, the chance values of each reliability also vary. Hopkins and Hermann suggested researchers report I × I, S-I, and U-I reliabilities along with the chance values of each of these which reported rates of target behavior would have produced. Investigators would thus examine each reliability to see if it were reasonably close to 100% and also to see if it were above chance, then present all these to consumers in their research reports. While it is clearly necessary to avoid accepting reliability percentages as evidence of observer agreement which are, in fact, likely as a result of chance or random responding, elaborate tables featuring each reliability and its chance value for every reliability check would not likely be read by consumers. Presenting means and ranges for each would remove the reliability data from direct access by consumers and the implications of such measures for any particular target behavior data points would be obscure.

Some authors have proposed the use of various correlational measures to summarize interobserver agreement, and to avoid the problems described above (Hartmann, 1977; Kratochwill and Wetzel, 1977). As Baer (1977) suggests, such measures are rather far removed from the basic data they summarize and are thus counter to our tradition of basing conclusions on data as little modified as possible. Also, as Kratochwill and Wetzel (1977) note, such summarizing procedures invariably discard information in the process of summarizing. The procedure recommended below does not modify reliability data at all, preserves all the information contained in the data, and presents them all graphically (rather than in computational or tabular form), generally believed a superior way to present data whenever possible. Finally, the correlational procedures do not address the believability of claimed experimental effects as directly as the procedure we recommend.

A PROPOSED SOLUTION

We believe the problems discussed earlier can be easily solved by a simple graphical presentation of disagreement rates, on the basic graph(s) of target behavior, with substantial gain in implications for assessing the believability of claimed experimental effects. The understanding of that graphical presentation is facilitated, though, by a rearrangement of the usual formulas for I × I, S-I, and U-I reliabilities. If the calculating formula for each is subtracted from 100%, then terms are algebraically collected over the denominator of the original ratio and simplified, the nature of each measure is clearer.

$$100\% - \text{I} \times \text{I} = 100 \times \left[\frac{\text{Disagreements}}{\begin{array}{c}\text{Agreements on Occurrences} + \\ \text{Agreements on Nonoccurrences} + \text{Disagreements}\end{array}}\right]$$

$$100\% - \text{S-I} = 100 \times \left[\frac{\text{Disagreements}}{\text{Agreements on Occurrences} + \text{Disagreements}} \right]$$

$$100\% - \text{U-I} = 100 \times \left[\frac{\text{Disagreements}}{\text{Agreements on Nonoccurrences} + \text{Disagreements}} \right]$$

Each measure of interobserver agreement is the inverse of disagreements divided by a denominator, but the denominators and thus the interpretations of the reliability measures differ. I × I reliability basically compares disagreements to the total number of observation occasions; S-I reliability compares disagreements to the sum of disagreements and agreements on occurrences; and U-I reliability compares disagreements to the sum of disagreements and agreements on nonoccurrences. Each focuses on disagreements, but each compares those to a different base and thus gives different information.

We believe much potential confusion for researchers and consumers can be avoided by using the disagreement rate itself as the primary measure of reliability, then examining its magnitude relative to occurrence agreements and/or to nonoccurrence agreements if desired. Thus observer error in general would be assessed, as percentage of disagreements (or as I × I reliability, if preferred), then observer error relative to agreements on occurrences and/or agreements on nonoccurrences would be examined as the ratio of disagreements to each (or as S-I and/or U-I reliability, if preferred) to see if any special implications exist. (We are not proposing a simple return to the use of I × I reliability, but present a graphical summary below we believe superior to any procedures so far discussed.)

In every case in which I × I and either S-I or U-I reliabilities are reasonable but one of the latter two is low, understanding is facilitated if researchers attend to the disagreement percentage itself as a measure of observer error and then to its magnitude relative to agreements on occurrences and/or nonoccurrences, while remembering that a constant disagreement rate produces lower S-I reliability at low response rates and lower U-I reliability at higher response rates.

GRAPHIC REPRESENTATION OF DISAGREEMENT RANGE

A simple graphic representation of disagreements aids greatly in making the recommended comparisons, conveys and summarizes all the information obtained from reliability checking, and is much simpler to present and understand than tables of obtained and chance S-I, U-I, and I × I reliabilities would be. In addition, the procedure has clear implications with respect to the believability of claimed experimental effects, as explained below.

For a given reliability check the percentage of intervals, moments, or trials that produced disagreements is found. This disagreement percentage is then graphed as a band-width or confidence interval around the primary observer's data for that day; but it is centered around the mean or the median (they are the same) of the two observers' reported rates of target behavior for that day. (Centering the disagreement range around the mean produces the properties to be described here; other ranges or other placements of the disagreement range would not.) The mean itself is not plotted, but both observers' data points are. The range is thus centered halfway between the two observers' reported rates of target behavior. The disagreement range so graphed effectively represents all the information collected by the reliability checking procedure, as explained below.[2]

The left part of Figure 1 demonstrates the use of the disagreement range for this purpose. The figure represents the results of a reliability check performed on Day 6 of an imaginary ex-

[2]Pioneering suggestions along these lines were made by Hawkins and Dotson (1975) and by Morris, Rosen, and Clinton (Note 1).

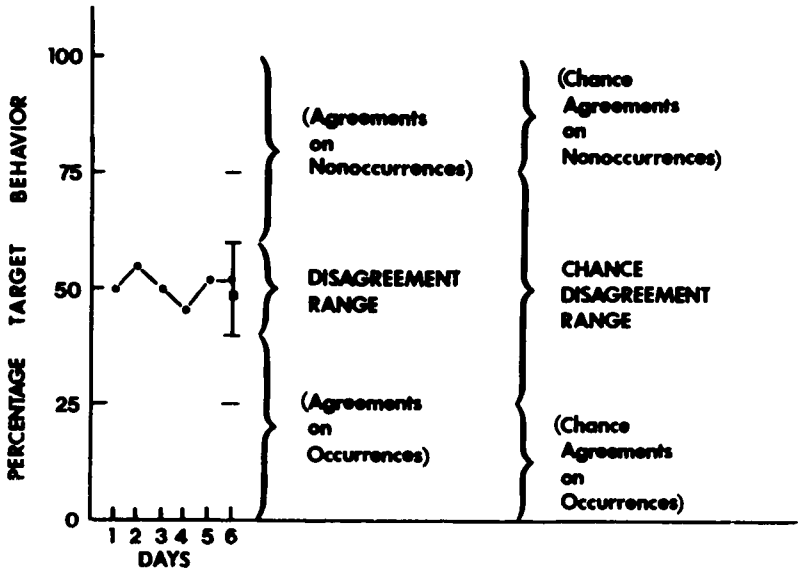

Fig. 1. Use of disagreement range to partition all observation occasions into those producing agreements on occurrences, disagreements, and agreements on nonoccurrences, and use of chance disagreement range to show chance rates of agreements on occurrences, disagreements, and agreements on nonoccurrences.

periment. On that day a reliability check was run which resulted in a disagreement percentage of 20%. Observer 1, the primary observer in this study, reported a rate of target behavior for that day of 52%, plotted as a solid circle. Observer 2, the reliability checker, reported a rate of 48%, plotted as a solid square. The mean or median of their reported rates is 50% and the disagreement range was centered around this number. This produced a disagreement range running from 40% to 60%, shown as a vertical line with horizontal limits.

The disagreement range is interpreted as follows. Both observers agreed that the behavior *occurred* on the 40% of observation occasions lying below the disagreement range. In addition, both observers agreed that the behavior *did not occur* on the 40% of observation occasions lying above the disagreement range. They disagreed on the 20% of observation occasions contained within the range.

The 20% disagreement range itself is rather substantial. Comparing it to the "agreement on occurrences" height indicates disagreements are half as great as such agreements, and comparing it to the "agreement on nonoccurrences" height also shows disagreements to be half as frequent as those agreements. With lower target behavior rates and the same disagreement rate, the disagreement range would be greater relative to agreements on occurrences and lesser relative to agreements on nonoccurrences. The opposite would be true if behavior rates were greater. Thus the graphical display of the disagreement range facilitates examination of the disagreement range, comparison of it to each of the two sorts of agreements, and identification of any special implications those comparisons might produce.

For those preferring to interpret these results in terms of the three reliability percentages, I × I reliability is the total height of the graph (100%) minus the disagreement range (20%), so equals 80%. S-I reliability is the "agreement on occurrences" height (40%) relative to that height (40%) plus the disagreement height (20%), so equals 67%. U-I reliability is the "agreement on nonoccurrences" height (40%) relative to that height (40%) plus the disagreement height (20%) so also, in this example, equals 67%.

The disagreement range thus used presents

visually all the information which would be summarized by I × I, S-I, and U-I reliability percentages, and shows clearly the relative rates of disagreement, agreements on occurrences, and agreements on nonoccurrences.

The disagreement range as used here has additional useful features. It always includes observers' reported rates of target behavior; if their reported rates were more unequal, the disagreements so produced would increase the range to include those reported rates. It always includes all disagreements, regardless of how they occurred. Graphically, the precise nature of disagreements is shown: for each observer the distance from the lower limit to the reported behavior rate is the percentage of occasions one observer reported behavior occurring and the other did not.

These points are illustrated in Table 2. In the first line of that table, while observer 1 reported a behavior rate of 52% and observer 2 a rate of 48%, the two observers disagreed on 20% of occasions, thus agreed on occurrences 40% of the time and on nonoccurrences the same. Thus observer 1 reported behavior occurring 12% of occasions on which observer 2 reported no behavior, and observer 2 reported behavior on 8% of occasions when observer 1 did not. In the second line observer 1 reports 60% occurrences, observer 2 reports 40%, and disagreements equaled 20%. Thus, on 20% of the observations observer 1 said behavior occurred while observer 2 reported it did not, and no converse cases occurred. In line three, the observers' rates differ by more than 20%, so the disagreement range must be 30% or more, with the agreements of each sort 35% or less.

GRAPHIC REPRESENTATION OF "CHANCE" DISAGREEMENT RANGE

A simple graphic summary of "chance" reliability information is also easily achieved. The "chance" disagreement percentage for two observers reporting frequencies of occurrence of target behavior O_1 and O_2, respectively, and frequencies of nonoccurrence of N_1 and N_2, across T observation occasions is given by:

$$\text{Disagreements (chance)} = 100 \times \frac{O_1N_2 + N_1O_2}{T^2}$$

If this chance disagreement percentage for each reliability check is then centered around the two observers' mean or median reported rate of target behavior for that day, the researcher and consumer can quickly see whether or not the obtained disagreement ranges are substantially smaller than "chance," smaller than random responding by observers would produce, and thus provide evidence of true observer reliability. Figure 1 illustrates the use of the chance disagreement range for this purpose. The chance disagreement range is shown by horizontal lines as limits at 75% and 25% above Day 6, explained to the right in that figure. The height below the chance disagreement range is the chance agreement on occurrences rate and the height above the chance disagreement range is the chance agreement on nonoccurrences rate. Using the imaginary data from Figure 1, the chance disagreement range calculates to equal 50% (rounded), is centered around the two observers' mean behavior rate, so ranges from 25% up to 75% in Figure 1, and leaves the

Table 2

Illustration of Additional Useful Features of Disagreement Range Used as Recommended

Reported Behavior Rate		*Disagreement Range*	*Agreements on Occurrences*	*Agreements on Nonoccurrences*
Observer 1	*Observer 2*			
52%	48%	20%	40%	40%
60%	40%	20%	40%	40%
65%	35%	30% or more	35% or less	35% or less

25% chance agreement on occurrences height below it and the 25% chance agreement on nonoccurrences height above it. If, as in Figure 1, the obtained disagreement range is smaller than the chance disagreement range, then agreements on occurrences and agreements on nonoccurrences are necessarily greater than chance, and disagreements are fewer.

Graphical representation of the chance disagreement range provides all the information contained in the three chance reliabilities Hopkins and Hermann (1977) recommended calculating. The chance level of I × I reliability is the chance disagreement range subtracted from the total height of the graph (100%; all observation occasions). Chance S-I reliability is the height below the chance disagreement range relative to that height plus the chance disagreement range itself. Similarly, chance U-I reliability is the height above the chance disagreement range relative to that height plus the chance disagreement range itself. Using the Figure 1 data and calculations, chance I × I reliability is 100% minus the 50% chance disagreement range, so equals 50%. Chance S-I reliability is the 25% chance agreement on occurrences height divided by that 25% plus the chance disagreement range of 50%, or 25%/25% + 50%, so equals 33%. Chance U-I reliability is the 25% chance agreement on nonoccurrences height divided by that 25% plus the chance disagreement range of 50%, so also equals 33%.

Each of the three chance reliabilities is a function of the chance disagreement range and the heights above and below it, and each of the actually obtained reliabilities is a function of the true disagreement range and the heights above and below it. Consequently, any time the true disagreement range is substantially smaller than the chance disagreement range then the obtained reliabilities will be substantially larger than their corresponding chance reliabilities. Thus, rather than examining tables of numerous reliability measures, researchers and consumers can attend to the graphical comparison of obtained and chance disagreement ranges to determine that random observer responding does not account for obtained observer agreement.

IMPLICATIONS FOR BELIEVABILITY OF EXPERIMENTAL EFFECTS

The disagreement range, when plotted as recommended here, leaves below the range all reports of target behavior on which both observers agree and leaves above the range all reports of the nonoccurrence of target behavior on which they agree. Within the range are only occasions when one observer reported the behavior and the other disagreed. Since both observers agree the behavior occurred at a rate equal to the lower limit of the disagreement range and agree it did not occur on the percentage of observation occasions above the range, then they agree only that the rate of target behavior may be anywhere within the range. (Of course, it is possible the true rate is actually outside the range and biases affecting the observers are leading them to err systematically. Assuming all efforts have been made to prevent such biases, however, our interobserver agreement only permits us to say our observers agree that the rate is probably within the disagreement range.)

Graphical presentation of the disagreement range as exemplified by Figure 1 can serve as the kind of graphical judgmental aid called for by Hawkins and Dotson (1975) and by Kratochwill and Wetzel (1977), a judgmental aid useful to consumers as well as researchers in determining whether or not, given the levels of observer agreement in a study, claimed experimental effects are believable. For experimental effects to be believable they must be substantial enough to produce no overlap between disagreement ranges resulting from reliability checks taken before and after the claimed experimental effect occurred. If overlap exists between these disagreement ranges, the apparent change in target behavior is not great enough, given the degree of observer disagreement, to

permit certainty that an effect of treatment has been shown.

The use of the disagreement range to examine the believability of experimental effects is illustrated in Figure 2. Panel A of Figure 2 shows imaginary data from interval recording through a baseline phase, a treatment phase, a reversal phase, and a second treatment phase. Reliability checks are performed on Days 5, 10, 15, and 20, with each reliability check yielding a disagreement percentage of 10%. (The reliability checker's reported percentage of target behavior is shown for each reliability checking day.) The primary observer's data indicate a decreased per-

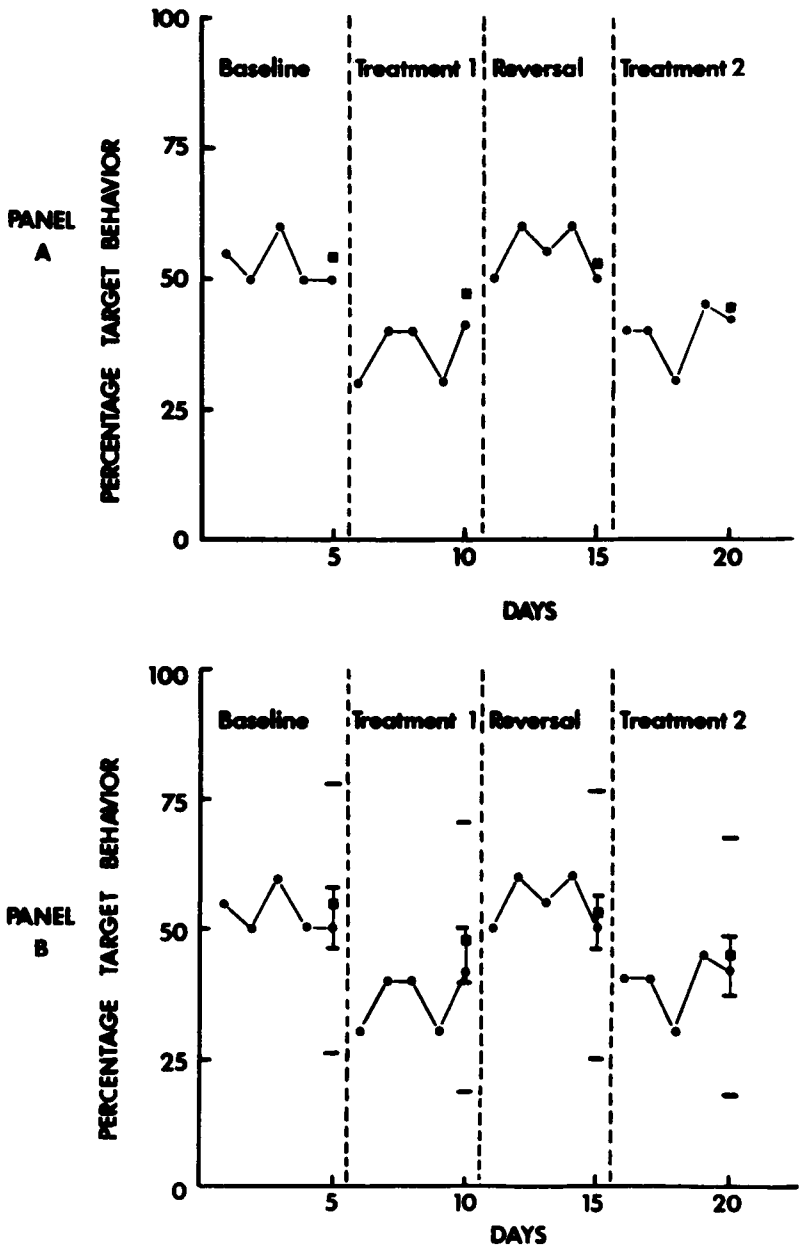

Fig. 2. Imaginary data showing primary observer's data over days and phases along with the second observer's data on reliability checking days (Panel A) and the same imaginary data with disagreement and chance disagreement ranges added (Panel B). Panel B assumes disagreement rates of 10% from each reliability check.

centage of target behavior during Treatment 1 relative to baseline, a return to baseline levels during the reversal, and a reduction back to Treatment 1 levels during Treatment 2. The reliability checker's data are consistent, showing a decrease to Treatment 1, an increase during the reversal, and a reduction back to Treatment 1 level during Treatment 2. Ninety percent I × I reliability is substantial enough to be considered generally acceptable and the agreement as to the effects of treatments and reversal shown by the reliability checker's data is the "cofunctional" reliability discussed by Goldiamond (1969), Hawkins and Dotson (1975), Kelly (1977), and by Kratochwill and Wetzel (1977). Thus the data *appear* to demonstrate a believable experimental effect.

Panel B of Figure 2 presents the same data as Panel A but shows the disagreement range and the chance disagreement range around the mean of the two observers' reported percentage of target behavior for each reliability checking day. Since the obtained disagreement ranges are consistently much smaller than the chance disagreement ranges, true observer agreement appears to have been obtained. However, comparison of the disagreement ranges across each of the four phases suggests that no believable experimental effects were demonstrated. In each case the disagreement range for the two observers is substantial enough that the true rate of target behavior may not have changed at all across these four conditions. If the true rate of target behavior was near the lower end of the range on Day 5, in the upper half of the range on Day 10, in the lower half of the range on Day 15, and in the uppermost part on Day 20, then no effects of treatments or the reversal would have occurred. Thus, with these apparent rate changes and levels of observer agreement, believable experimental effects were not obtained.

If greater effects of the treatment had been demonstrated, then with these levels of observer agreement believable experimental effects could have been demonstrated. Conversely, if greater observer agreement had been achieved, believable experimental effects might also have been shown. With the magnitude of treatment effects and the degree of observer agreement in these examples, however, neither Treatment 1 nor Treatment 2 demonstrates conclusive changes in rates of target behavior from baseline and reversal levels. The use of the disagreement range has permitted us to avoid accepting as believable experimental effects which, in fact, are not. Given that the primary observer's data and cofunctional reliability had both suggested that treatment effects had occurred, this added protection for researchers and consumers against type one errors, urged by Baer (1977), and provided by graphical presentation of the disagreement range, is strongly recommended.[3]

RECOMMENDATIONS

(1) Researchers in applied behavior analysis should attend to their obtained observer disagreement percentages as their primary measure of observer agreement, then compare it to observers' rates of agreements on occurrences and rates of agreements on nonoccurrences, recalling that lower and higher target behavior rates produce lower S-I and U-I reliabilities, respectively, with constant disagreement rates.

(2) The second observer's data should be plotted, for each reliability check, on graphs of the primary observer's data, as Hawkins and Dotson (1975) proposed.

(3) Disagreement percentages should be

[3]Early readers of this manuscript have asked how our recommendations relate to the fact that *sets* of data points within experimental conditions usually form the basis for conclusions regarding change across conditions. If the primary observer's data on reliability checking days are consistent with the data when no checks are being conducted, then the disagreement range is likely representative of what would be obtained by multiple checks within conditions, and so can be viewed as establishing a band around the data points within that condition. No overlap between such bands across conditions would then be a strong argument for believability of claimed experimental effects.

presented on graphs of the primary observer's data, centered halfway between the two observers' reported rates of target behavior. This provides all the information on observer agreement collected, in an easily understood format, and permits identification of claimed experimental effects which are not believable.

(4) Researchers should also graph the limits of chance disagreement ranges for each reliability check, thus providing all available information on chance agreement and permitting easy determination as to whether or not true observer agreement has been demonstrated.

REFERENCE NOTE

1. Morris, E. K., Rosen, H. S., and Clinton, L. P. *Reliability considerations in single-subject research.* Paper presented at meeting of the Midwestern Association for Behavior Analysis, Chicago, Illinois, May 1975.

REFERENCES

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

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

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

Goldiamond, I. Stuttering and fluency as manipulatable operant response classes. In L. Krasner and L. P. Ullman (Eds), *Research in behavior modification.* New York: Holt, Rinehart & Winston, 1969.

Hartmann, P. Considerations in the choice of interobserver reliability estimates. *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), *Behavioral analysis: areas of research and application.* Englewood Cliffs, New Jersey: Prentice-Hall, 1975.

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

Kazdin, A. E. *Behavior modification in applied settings.* Homewood, Illinois: Dorsey Press, 1975.

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

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

Kratochwill, T. R. and Wetzel, R. J. Observer agreement, credibility and judgment: Some consideration in presenting observer agreement data. *Journal of Applied Behavior Analysis,* 1977, **10,** 133-139.

Powell, J., Martindale, A., and Kulp, S. An evaluation of time-sample measures of recording. *Journal of Applied Behavior Analysis,* 1975, **8,** 463-464.

Received 3 April 1978.
(Final Acceptance 22 January 1979.)