JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1976, 9, 109-113 NUMBER 1 (SPRING) 1976

TECHNICAL ARTICLE
DIFFERENCES AMONG COMMON METHODS FOR CALCULATING INTEROBSERVER AGREEMENT

ALAN C. REPP,[1] DIANNE E. D. DEITZ, SHAWN M. BOLES, SAMUEL M. DEITZ,[2] AND CHRISTINA F. REPP

GEORGIA RETARDATION CENTER AND GEORGIA STATE UNIVERSITY

In most applied studies, experimenters attempt to increase the probability that data accurately reflect the subject's behavior by assessing the degree to which two observers agree that responding has occurred. While some authors report this comparison as an index of observer reliability, others report it as an index of observer or interobserver agreement.

Regardless of the term used, most authors who report agreement (instead of correlation) use variations of the same procedure. In general, each session is divided into a number of time blocks, the number of time blocks with interobserver agreement is divided by the sum of the agreements and disagreements, the quotient is multiplied by 100, and the result is reported as per cent agreement between observers. Although several variations exist within this general procedure, the most common are variations in the length of the time block and in the definition of an interval of agreement. The present experiment compared the results of computing interobserver agreement by these common methods and variations.

METHOD

General Procedure

Several methods for calculating interobserver agreement on the same data were compared. Two observers unaware of the purpose of the experiment recorded responding by five children in an early childhood program. Five response classes (each child emitted a different response) were recorded for varying numbers of sessions: R_1 (turning in seat, 13 sessions), R_2 (mouthing nonedibles, 12 sessions), R_3 (making eye contact, eight sessions), R_4 (screaming, nine sessions), and R_5 (touching objects, 21 sessions).

Although the responses, *per se,* were not of importance in this experiment, they were defined to make the data representative of field studies.[3] In the preexperimental phase, both observers discussed and wrote definitions on each response class while observing the response and then independently recorded responding. After any session in which the per cent agreement calculated by the Whole-Session method (described below) was greater than 80%, the discussions stopped, and another recording session occurred. If a session resulted in an agreement score of less than 80%, another discussion and recording session occurred. After three consecutive sessions with agreement of at least 80%, the pre-experimental phase ended and the experimental phase began.

Three rooms were used. The children were in a classroom; the observers seated approximately 3 m apart were in a second room separated from the first by a one-way mirror; the equipment (an event recorder, timer, and power supply) was in a third room acoustically isolated from the other two rooms. Sessions varied from 5 to 12 minutes, and in each session two observers recorded the responding of one child by closing one of the two silent microswitches that controlled separate pens on an event recorder. These pens were deflected from the beginning to the end of a response. A third pen on the event recorder, momentarily deflected at 5-sec intervals by an electronic timer, provided a permanent record of time against which the data could be analyzed. As the observations were made during an ongoing educational program, sessions began and ended when the children's location in the room facilitated observations. At the end of the experiment, per cent agreement was calculated by the three methods described below.

1. *Whole-session method.* In this method, the time block was equivalent to the entire session. The number of responses recorded by each observer was summed, the smaller number divided by the larger, and the quotient multiplied by 100 to yield per cent agreement.

2. *Exact-agreement method.* In this method, the event recorder pen that deflected every 5 sec provided a means by which the number of responses recorded by each observer in each interval could be compared. Responses were defined as the deflection of an observer's pen or the continuance of a deflection that had begun in a preceding interval. An agreement was defined as an interval in which both observers recorded the same number of responses; disagreement as

[1]Georgia Retardation Center 4770 North Peachtree Road Atlanta, Georgia 30341.

[2]Georgia State University.

[3]More complete definitions are available on request.

an interval in which the observers did not record the same number of responses. Interobserver agreement was determined by dividing the number of intervals of agreement by the total number of intervals in the session.

There were two variations within this method. The first was the length of the interval—5, 10, 20, and 30 sec—with the three larger being made by combining consecutive 5-sec intervals. In the second variation, calculations were based either on all the intervals in a session (All-Intervals method) or only on those intervals in which at least one observer recorded responding (Response-Intervals-Only method).

3. *Category method.* Calculations were similar to the previous method, the difference being that an interval was defined as one of agreement either if both observers failed to record any responding or if both observers recorded at least one response. An interval of disagreement was defined as one in which one observer recorded responding and the other did not. The same two variations made within the Exact-Agreement method were made within the Category method.

Data analysis.[4] The per cent of interobserver agreement was used for each method. Analyses were conducted for the various calculation methods on each of the five response classes. The main comparisons were: (1) Whole Session *versus* Exact Agreement *versus* Category, (2) All Intervals *versus* Response Intervals Only, and (3) 5-sec *versus* 10-sec *versus* 20-sec *versus* 30-sec time intervals.

RESULTS

The rates of responding (1.4, 6.6, 3.7, 1.7, and 3.2 rpm for responses classes one through five, respectively) are representative of experiments dealing with moderate response rates.

Whole Session versus *Exact Agreement* versus *Category*

In the first analysis, data were combined in two ways: (1) across the All-Intervals and Response-Intervals-Only methods, and (2) across the four interval sizes so that the Exact-Agreement method and the Category method could be compared with the Whole-Session method for each of the five response classes. Means, which were calculated for each response class (Figure 1), indicated that: (1) the Category and Whole-Session methods produced similar results, with the former producing higher scores for three of the five response classes (mean difference = 2%), (2) the Category method produced higher percentages (mean difference = 14%) than the Exact-Agreement method for all five response classes, and (3) the Whole-Session method produced higher percentages than the Exact-Agreement method for four of the five response classes (mean difference = 12%).

All Intervals versus *Response Intervals Only*

In the second analysis, data were combined across the Exact Agreement, Category, and Interval Size factors so that the All-Intervals and Response-Intervals-Only methods could be compared. Means calculated for each method and for each response are plotted in Figure 1 (lower portion). For each response class, the Response-Interval-Only method consistently produced lower agreement percentages than the All-Intervals method. The mean difference across the five response classes was 7%.

Interval Size

Means for interobserver agreement calculated by the Exact-Agreement and Category methods were calculated for the four interval sizes (Figure 2) and indicate that: (1) as the interval size increased, the difference between the agreement percentages derived from the two methods increased; (2) as the interval size increased, the agreement percentages calculated by the Exact-Agreement method decreased; and (3) as the interval size increased, the agreement percentages calculated by the Category method increased.

Combined Methods

With the data from the five response classes at the smallest and at the largest intervals averaged, the results indicated that: (1) the Category—All-Intervals method produced the highest scores at both the 5-sec (mean = 94%) and the 30-sec intervals (mean = 94%), (2) the Exact-Agreement-Response-Intervals-Only method produced the lowest scores (means = 77% and 64%), (3) the Category-Response-Intervals Only method produced increasing scores with increasing interval size (means = 85% and 92%), and (4) the Exact-Agreement-All-Intervals method produced decreasing scores with increasing interval size (means = 91% and 71%).

DISCUSSION

Exact Agreement versus *Category*

For all five response classes, the data indicated considerable differences between these two methods of calculating interobserver agreement, with the Category method producing the highest scores. This difference occurred because the Category method is the limiting value of the Exact-Agreement method. That is, any interval that is scored as one of agreement in the Exact-Agreement method must be scored as one of agreement in the Category method, but not all intervals scored as agreement in the Category method are scored as agreement in the Exact-Agreement method (*e.g.*, if one observer recorded four responses and another observer recorded three).

[4]Appropriate statistical analyses were conducted on all comparisons and are available in an expanded version of this paper.

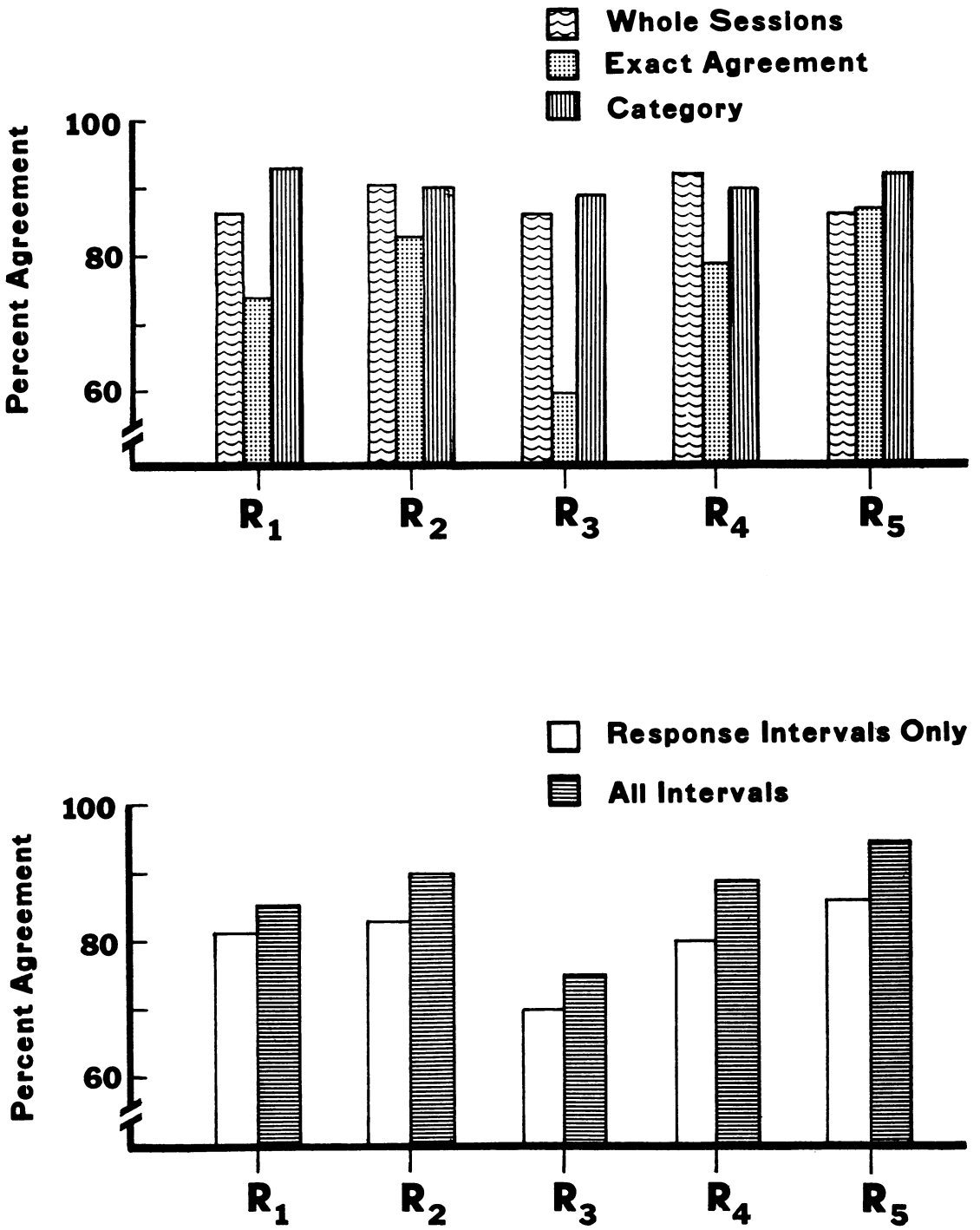

Fig. 1. The mean per cent of interobserver agreement for each method for each response class. All calculations were made on the same data.

All Intervals versus *Response Intervals Only*

For all response classes, the All-Intervals method produced higher agreement percentages than the Response-Intervals-Only method; and the formula for calculating per cent interobserver agreement is the reason for the difference. Any interval defined as one of agreement and any interval defined as one of disagreement within the Response-Intervals-Only method is defined in the same manner in the All-Intervals method. The difference between these two methods is the inclusion by the latter method of all intervals in which neither observer recorded any responding. These intervals, of course, are defined as intervals of agreement in the All-Intervals method but are excluded from calculations in the Response-Intervals-Only method.

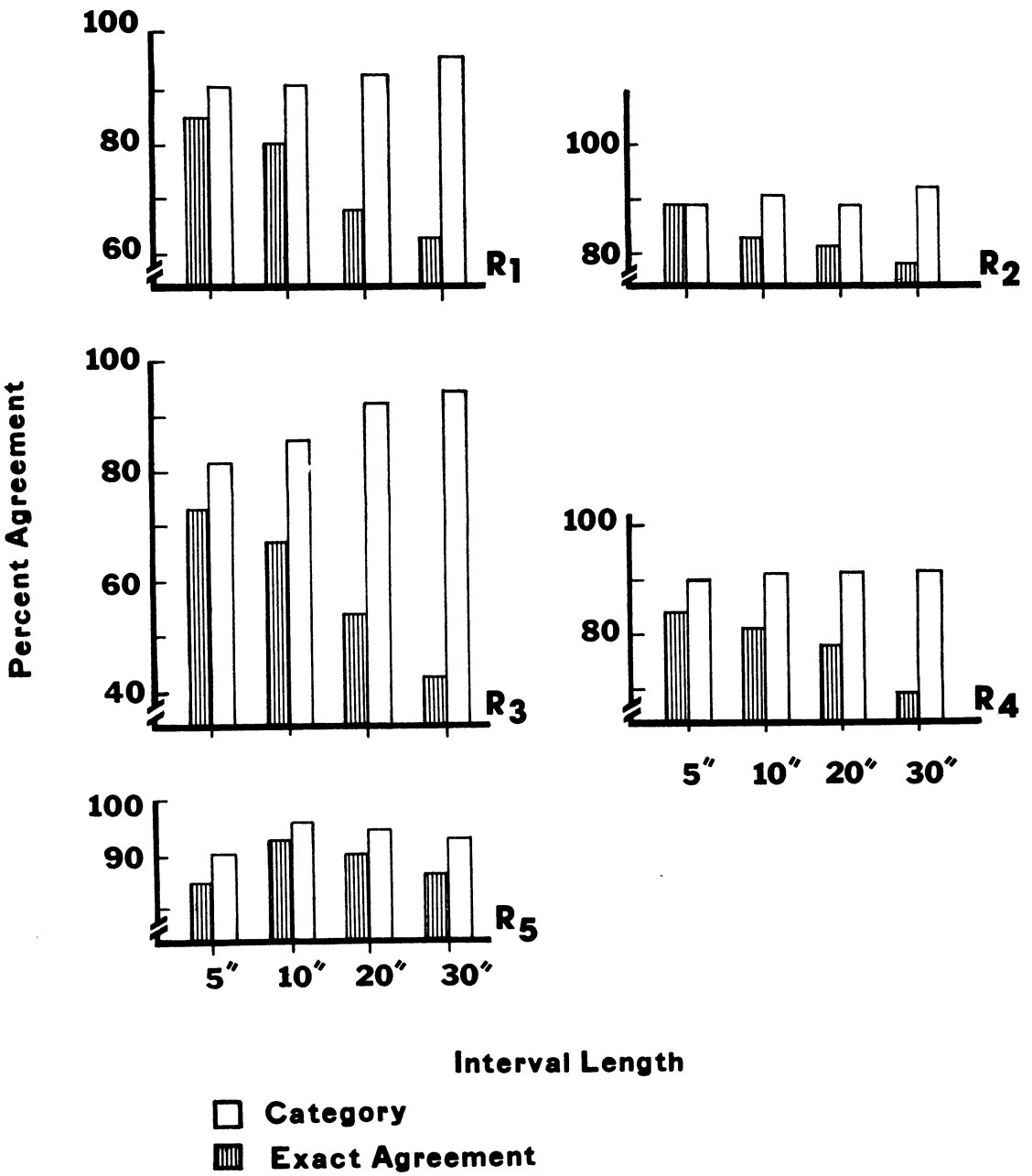

Fig. 2. The mean per cent of interobserver agreement calculated by the Category and Exact-Agreement methods for each response class. Data were collapsed across the time intervals and across the other methods.

Interval Size

Generally, as the interval size increased, the difference between the interobserver agreement percentages derived from the Exact-Agreement and Category methods increased. Figure 2 indicates that the increased differences resulted from two factors. The first was that as the interval size increased, the agreement percentages calculated by the Exact-Agreement method tended to decrease. As the interval size increased, the number of responses recorded by each observer in an interval increased, and the opportunity for and, hence, probability of disagreement between observers increased. The trend held in all but one case, 5-sec and 10-sec intervals for R_5. The reason for this exception was that errors could cancel each other. For example, Observer A might record two responses in the first 5 sec and no responses in the next 5 sec, while Observer B might do the obverse. With 5 sec as the interval size, both intervals would be scored as intervals of disagreement, and the agreement score would be 0%. However, with 10 sec as the interval size, the interval would be scored as one of agreement, and the agreement score would be 100%.

The second factor was that the agreement percentages calculated by the Category method generally increased as the interval size increased. The tapes indicated that as interval size increased, the probability of both observers recording at least one response increased. As this method defined agreement as an interval in which both observers recorded at least one response, the interobserver agreement percentage increased as interval size increased. This trend, however, did not occur in all cases (10, 20, and 30 sec for R_5, 30 sec for R_4). The reason for these exceptions appears to be the case in which neither observer recorded responding in a smaller interval while one observer recorded responding in the larger interval.

Whole Session

The Whole-Session method resulted in high, but not the highest, interobserver agreement percentages. Since the calculation of interobserver agreement by this method does not have the same basis as the other methods, there is no reason to expect similarity in the results of this and the other methods. Some investigators (*e.g.,* Johnson and Bolstad, 1973) have argued that this is the least justifiable of the common methods, since it does not indicate whether one observer recorded the same response (or at least responding during the same time period) as the other observer. Because of the argument that it is the least justifiable method, there is a common assumption that it would produce the most liberal interobserver agreement percentages. While the method may not be valid, the lack of an exaggerated agreement percentage is interesting and perhaps unexpected.

GENERAL DISCUSSION

Previous research has indicated several reasons for inaccuracy in recording data and in calculating interobserver agreement, and most of these reasons have been based on the observer's behavior. The present study has defined the behavior of another individual, the experimenter, as a cause of variation in reported interobserver agreement scores. The data indicated that the method chosen by the experimenter for calculating interobserver agreement has an effect on the percentages reported. In reporting agreement percentages, one can ensure higher scores by using the intervals in the recording period. Experimenters concerned with reporting more conservative scores can do so by calculating interobserver agreement based only upon those intervals in which responding occurred. In addition, more conservative agreement percentages also can be presented by using the Exact-Agreement method instead of the Category method (a common variation of which is usually labelled "time block"). However, this is not an argument for the exclusive use of the Exact-Agreement method, even though it is the most conservative, as the category method may occasionally be more appropriate (*e.g.,* when observers are recording multiple responses and using only pencil and paper).

The method of calculating interobserver agreement can have a considerable effect on the scores reported (overall means across responses on the same data varied from 64% to 94%). While 64% agreement may be insufficient for most experimenters and 94% agreement sufficient, the percentages themselves are misleading because they are a function of the method used to calculate interobserver agreement, rather than a function of data. Differences that one could reasonably expect would have arisen only from differences in the response data or from the comprehensiveness of the response definition may have arisen from differences in methods for calculating agreement, and one experimenter's report of 94% may reflect no more agreement between observers than another experimenter's report of 64%.

REFERENCES

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 practice.* Champaign, Illinoise: Research Press, 1973. Pp. 7-67.