*INFLUENCES OF RESPONSE RATE AND DISTRIBUTION ON THE
CALCULATION OF INTEROBSERVER RELIABILITY SCORES*

NATALIE U. ROLIDER

KENNEDY KRIEGER INSTITUTE

BRIAN A. IWATA

UNIVERSITY OF FLORIDA

AND

CHRISTOPHER E. BULLOCK

KENNEDY KRIEGER INSTITUTE

We examined the effects of several variations in response rate on the calculation of total, interval, exact-agreement, and proportional reliability indices. Trained observers recorded computer-generated data that appeared on a computer screen. In Study 1, target responses occurred at low, moderate, and high rates during separate sessions so that reliability results based on the four calculations could be compared across a range of values. Total reliability was uniformly high, interval reliability was spuriously high for high-rate responding, proportional reliability was somewhat lower for high-rate responding, and exact-agreement reliability was the lowest of the measures, especially for high-rate responding. In Study 2, we examined the separate effects of response rate per se, bursting, and end-of-interval responding. Response rate and bursting had little effect on reliability scores; however, the distribution of some responses at the end of intervals decreased interval reliability somewhat, proportional reliability noticeably, and exact-agreement reliability markedly.

Key words: interobserver agreement, reliability, response rate, response distribution

Response measurement in applied research usually involves data collection by human observers, which is likely to be more prone to error than machine transduction. As a result, evaluation of observers' consistency or reliability has become a standard feature of applied research and is accomplished by determining the extent of scoring agreement between the records of independent observers. A number of factors may influence reliability (see reviews by Kazdin, 1977; Page & Iwata, 1986); this study focuses on methods used to calculate reliability statistics and how they are influenced by response rate and distribution.

Four methods of quantifying observer reliability seem to be used most often in current research: total, interval, exact, and proportional

indices. Total reliability determines the extent to which observers record the same number of responses throughout a session (Bijou, Peterson, & Ault, 1968). Reliability is calculated by dividing the smaller number of recorded responses by the larger. This calculation is fast, simple, and sensitive to discrepancies in the overall level of responding recorded between observers. However, total reliability does not indicate whether observers recorded the same responses (Page & Iwata, 1986).

An alternative to total reliability is interval reliability, which involves dividing a session into smaller, equal-length time intervals and comparing observers' records during each interval (Bijou et al., 1968). An agreement is defined as both observers recording at least one response or not recording any responses in a given interval, and reliability is calculated by dividing the number agreement intervals by the number of intervals in the session. Interval

Address correspondence to Brian A. Iwata, 114 Psychology Building, University of Florida, Gainesville, Florida 32611 (e-mail: iwata@ufl.edu).

doi: 10.1901/jaba.2012.45-753

reliability restricts comparisons to smaller units of time, thereby increasing the likelihood that agreements reflect observation of the same responses. A limitation of the interval method, however, is that extremely high (or low) response rates may produce spuriously high agreement scores (Hopkins & Hermann, 1977).

A more stringent variation of interval reliability is exact-agreement reliability (Repp, Deitz, Boles, Deitz, & Repp, 1976), in which an agreement is defined as both observers recording the same number of responses in an interval. This method is the most conservative reliability estimate because any difference in data recording results in a complete disagreement. For example, an interval in which one observer records four responses while the other observer records five responses will be scored as a disagreement.

Another variation of interval reliability is proportional reliability (Bailey & Burch, 2002). First described by Bostow and Bailey (1969), proportional reliability involves dividing the smaller frequency by the larger in each observation interval and averaging the resulting fractions across all intervals. Because proportional reliability is based on the number of responses recorded, not just whether a response was recorded, it is a more stringent measure than interval reliability. At the same time, it allows "partial credit" for each response recorded in the interval, whereas exact-agreement reliability does not. Thus, proportional reliability seems to strike a balance between the less stringent interval method and the more stringent exact-agreement method.

Repp, Dietz, *et al.* (1976) compared three methods of calculating interobserver reliability for five responses recorded by two observers: total (described as "whole session"), interval (described as "category"), and exact agreement. The authors' main focus was on the influence of observation interval length (5, 10, 20, and 30 s); thus, all responses occurred at moderate rates ($M = 3.3$ responses per minute). They

found that interval reliability increased and exact-agreement reliability decreased as a function of increased interval length, resulting in larger discrepancies between the two indices. It should be noted that longer intervals (e.g., 20 or 30 s) are not typically used in current applied behavior-analytic research; 87% of the studies published in the *Journal of Applied Behavior Analysis* from 1999 through 2005 in which time-based data were collected used interval lengths of 10 s or less. Thus, we used 10-s intervals in the present study as the basis for comparison.

Given a constant interval length, the response characteristic most likely to influence interobserver reliability is response rate. Although several studies have examined the effect of response rate on outcomes that result from the use of different data-collection methods (Powell & Rockinson, 1978; Repp, Roberts, Slack, Repp, & Berkler, 1976), only one study to date has examined the influence of response rate on interobserver reliability indices (Mudford, Martin, Hui, & Taylor, 2009).

Mudford *et al.* (2009) compared exact and proportional (described as "block-by-block") reliability to time-window analysis, in which an agreement is scored if both observers' data records contain a response within $\pm t$ s of one another. Twelve observers recorded data from six video samples of client-therapist interactions, and focused on one target response during each session that varied in response rate (three samples) or duration (three samples). Response rates were 4.8, 11.3, and 23.5 per minute for the low-, medium-, and high-rate responses, respectively. Results showed that exact and proportional reliability were similarly high for the low-rate response ($M_s = 78.3\%$ and 85.3% , respectively). However, exact-agreement reliability was noticeably lower than proportional reliability for the medium-rate ($M_s = 59.5\%$ and 76.8% , respectively) and high-rate responses ($M_s = 50.3\%$ and 88% , respectively). These results suggest that reliability calculations are influenced by the rate of a

response, but they did not determine whether lower exact-agreement results were a function of response rate per se or some other characteristic of high-rate responding such as periodic bursting.

The purpose of this study was to extend the results of Mudford et al. (2009) by comparing interobserver reliability scores based on four calculation methods with data sets that varied in response rates (Study 1) and subsequently conducting a more fine-grained analysis to identify response-distribution characteristics that may contribute to lower reliability scores (Study 2).

STUDY 1

METHOD

Subjects

Five undergraduate students who had been trained previously in data-recording methods served as observers. They participated on a voluntary basis as an alternative to their normal responsibilities as observers for other ongoing studies.

Response Apparatus

Naturally occurring responses vary across a number of dimensions and, as such, are not well suited to the manipulation of any one dimension. Therefore, responses in the present study consisted of automated "data" that were computer generated with software (Visual Basic) that allowed precise control over response rate, duration, and distribution within and across sessions. All sessions lasted 10 min and began with the computer signaling a countdown to synchronize observers, after which the screen turned white to indicate that the session had begun. A response consisted of the appearance of a colored geometric shape (e.g., blue square) on the screen for 0.5 s. Different-colored shapes represented different target responses.

Responses were programmed to occur at three different rates so that reliability could be compared across a range of values. The mean

rate was 1 response per minute (range, 0 to 2) for low-rate responses, 5 (range, 2 to 8) for moderate-rate responses, and 25 (range, 15 to 35) for high-rate responses. Six types of sessions were generated; three involved only one target response (low-, moderate-, and high-rate sessions) and three involved two target responses (low/moderate-, moderate/moderate-, and high/moderate-rate sessions). For sessions with two target responses, the main objective was to compare reliability scores for the low-, moderate-, and high-rate responses; the second response was added at a moderate rate to increase the complexity of scoring.

Procedure

The five observers used handheld computers to record data for each of the six types of sessions, which were presented separately (each session was viewed by each observer independent of other sessions and other observers) in the following order: low rate, moderate rate, high rate, low/moderate rate, moderate/moderate rate, high/moderate rate. Prior to each session, the experimenter provided the observer with instructions on the key assignments for responses. During the session, the observer sat 1.5 m away from the computer screen. All data were saved on disk for later comparison with the automated data set generated by the computer program.

Data Analysis

Reliability usually is assessed by comparing two observers' data records under the assumption that a high index of reliability reflects accurate data collection. Comparisons in the present study were made between each observer's record and the computer-generated data set, which served as a preestablished standard for comparison; thus, the index of reliability in this study was identical to an index of accuracy.

Reliability for each observer for each session type was calculated in four different ways. Total reliability was calculated by dividing the smaller frequency of observed events across the entire session by the larger. Interval reliability was assessed by first partitioning the session into

consecutive 10-s intervals and then comparing the records. Intervals in which both the observer and the data set recorded at least one response or did not record any responses were scored as agreements, and reliability was calculated by dividing the number of agreement intervals of by the total number of intervals. Exact reliability was assessed by comparing the number of recorded events for both records in each 10-s interval. Intervals that contained the same number of recorded events were scored as agreements, and reliability was calculated by dividing the number of agreement intervals by the total number of intervals. Proportional reliability was assessed by comparing the number of recorded events in each 10-s interval. Reliability was calculated by dividing the smaller frequency by the larger in each interval and averaging the resulting fractions across intervals.

Reliability for each of the above calculations was assessed by having a second scorer (a graduate student) independently calculate reliability scores for 30% of sessions. Agreement on the calculated scores was 100% for these sessions.

RESULTS

Figure 1 shows individual reliability scores for each observer for each of the six session types, and Table 1 summarizes the results as mean percentages across observers. All calculation methods yielded high reliability scores for the low-rate session. Total reliability remained high for the moderate-rate session, whereas interval, proportional, and exact-reliability scores decreased noticeably for Observers 1 and 2 and slightly for Observers 3, 4, and 5. Interestingly, total, interval, and proportional reliability remained high for the high-rate session ($M_s = 98.9\%$, 100% , and 94.6% , respectively), whereas exact agreement decreased noticeably for every observer ($M = 77.3\%$).

Reliability scores for the low-rate response of the low/moderate-rate session were high for the

most part based on all calculations, although the total indices were somewhat lower for Observers 1, 3, and 4. Total reliability remained high for the first moderate-rate response of the moderate/moderate-rate session, whereas scores for the interval, proportional, and exact methods were somewhat lower for three of the five observers. Exceptions were Observer 4, whose interval score was slightly higher than the total score, and Observer 5, whose score was 100% based on all calculations. (Although not shown in the figure, mean scores for the second moderate-rate response of the sessions that involved two target responses were similar to what was obtained for the first moderate-rate response across all calculations.) Total reliability for the high-rate response of the high/moderate-rate session remained high ($M = 95\%$), interval reliability increased to 100% for all observers, and the proportional and exact methods showed consistently small ($M = 91.2\%$) and large ($M = 64.3\%$) decreases, respectively.

Overall, these data show that total reliability was relatively unaffected by response rate and that proportional reliability was mildly influenced by response rate. The interval and exact methods yielded the highest and lowest scores, respectively, for high-rate responses.

STUDY 2

Three characteristics of high-rate responding may influence reliability scores. First, an increase in response rate per se (i.e., more responses) may produce more recording errors. Second, high-rate responding may be characterized by irregular interresponse times that create periodic response bursts, which may be more difficult to recorded accurately. Third, high-rate responding may produce a situation in which responses frequently occur at the end of an interval, which may decrease reliability if one observer records the response at the end of one interval and the other observer records it at the beginning of the subsequent interval. Under

Figure 1. Reliability scores for each observer for low-, moderate-, and high-rate responses during single-response sessions (left) and during sessions in which a second moderate-rate response occurred (right) in Study 1.

typical conditions, including those used in Study 1, all of these effects may occur but are uncontrolled and confounded. The purpose of Study 2 was to isolate these three effects to determine their influence on reliability scores.

METHOD

Subjects

Ten trained observers drawn from the same general pool as those from Study 1 participated.

Procedure

Data for a single target response were generated and scored in a manner identical to that described in Study 1. All sessions lasted 10 min. Five observers collected data on two sessions that involved a manipulation of response rate only. In the moderate-rate session, responses were programmed to occur at a rate of 6 per minute, and all intervals contained one response. In the high-rate session, responses

Table 1
Mean Reliability Scores Across Observers for All Sessions in Studies 1 and 2

Session type	Total agreement	Interval agreement	Proportional agreement	Exact agreement
Study 1				
Low	98.0	100	99.8	99.7
Moderate	98.0	91.0	89.3	87.0
High	98.9	100	94.6	77.3
Low/moderate ^a	93.0	99.0	98.9	98.7
Moderate/moderate ^a	95.7	89.7	87.0	83.7
High/moderate ^a	95.0	100	91.2	64.3
Study 2				
Moderate	99.3	99.7	99.5	99.3
High	99.6	100	99.2	96.3
Constant	98.3	100	98.9	98.2
Burst	98.8	99.0	98.3	96.6
Middle of interval	99.0	99.7	99.4	98.6
Half and half	99.3	86.3	71.9	53.7

^a Scores presented are for the low-, moderate-, and high-rate responses that occurred along with a second moderate-rate response.

were programmed to occur at a rate of 24 per minute, and all intervals contained four responses. In both sessions, no responses occurred during the last 3 s of the interval. Thus, these sessions were programmed to manipulate rate but to eliminate the influence of both bursting and end-of-interval responding. The sessions were presented separately; moderate preceded high.

Five different observers collected data on four sessions that involved response bursting and end-of-interval manipulations. All responses were programmed to occur at a moderate rate (8 per minute; range, 4 to 18). In the response-bursting manipulation, response intervals in the constant session contained one to three responses, whereas response intervals in the burst session contained four to seven responses. To eliminate the potential influence of end-of-interval responding in the manipulation of bursting, no responses in the constant and burst conditions occurred during the last 3 s of an interval. In the end-of-interval manipulation, no responses in the middle-of-interval session occurred during the last 3 s of any response interval. Responses in the half-and-half session conformed to the same rule for half of the response intervals; for the other half, one response always occurred at the 9th s of the

interval (i.e., just before the end of the interval). We originally programmed a response at the end of every interval but found in pilot work that this arrangement could yield a spuriously high reliability. That is, if one observer consistently scored the response in the subsequent interval, that “late” entry would compensate for the other observer’s “early” entry of a response in the following interval. Programming the late response in half of the intervals prevented this problem. The observers collected data on each of the four sessions, which were presented separately in the following order: middle of interval, half and half, constant, and burst.

Data Analysis

Total, interval, exact, and proportional indices were calculated for all sessions as described in Study 1. Reliability of the calculations was assessed by having a graduate student independently calculate reliability scores for 33% of sessions. Reliability for the calculated scores was 100% for these sessions.

Results

Figure 2 shows individual agreement scores for each observer for each of the six session types, and Table 1 summarizes the results as mean percentages across observers. All calcula-

Figure 2. Reliability scores for each observer for all sessions in Study 2.

tion methods yielded high reliability scores in the moderate and high sessions, indicating that higher response rates per se did not influence reliability. Similarly, all calculation methods yielded high reliability scores in the constant and burst sessions, indicating that response bursting did not seem to influence reliability, at least for these observers. Reliability scores for the middle-of-interval session also were uniformly high based on all calculations. The only noticeable discrepancy among reliability indices was observed for the half-and-half session. Total reliability remained high ($M = 99.3\%$), whereas successively lower scores were obtained with interval, proportional, and exact methods ($M_s = 86.3\%$, 71.9% , and 53.7% , respectively). This effect was extremely consistent for all observers but was most pronounced for Observers 6, 7, and 10.

GENERAL DISCUSSION

Variations in response rate had different effects on four common reliability calculation methods (Study 1). Unlike total and interval reliability, proportional reliability showed sensitivity to response rate but was not affected as detrimentally by response rate as was exact reliability. When the characteristics of high-rate responding were examined more closely (Study 2), it appeared that end-of-interval responding, rather than response rate per se or response bursting, exerted greater influence over reliability scores. As was the case in Study 1, the exact-agreement method was most adversely affected by end-of-interval responses.

Results of Study 1 show that total reliability was relatively unaffected by response rate; only slight decreases in total reliability were observed in the high-rate condition. This finding was not surprising given that the sole basis for agreement is the total number of responses recorded during a session. By contrast, interval reliability was the only index that produced uniformly high scores in the high-rate condition. This result was consistent with previous research that has shown that interval agreement produces

spuriously high scores under extreme rates of responding (e.g., Hopkins & Hermann, 1977). It is important to note that we examined the upper but perhaps not the lower range of the response-rate continuum. All of the calculation methods in the low-rate (1 response per minute) condition yielded high scores; thus, it is possible that interval reliability would have been higher relative to the other calculation methods had an even lower response rate been used.

Proportional reliability scores were mildly influenced by response rate, with obtained values always between interval and exact scores, whereas the exact method was most affected by response rate (decreasing as response rates increased). Collectively, these findings suggest that proportional reliability should be the preferred index because it is neither insensitive to nor overly influenced by response rate and it more accurately reflects relative patterns of scoring between observers.

Results of Study 2 showed that high (constant) response rates and response bursting had little effect on reliability scores. When the location of responses within the interval was manipulated so that half of the response intervals contained a response near the end of the interval, only total reliability was unaffected. Progressively larger effects of end-of-interval responding were observed with interval, proportional, and exact reliability.

End-of-interval responding had no effect on total reliability because the calculation is based on overall session rates; therefore, distribution within intervals is irrelevant. When an end-of-interval response is scored in a subsequent interval by one observer, the calculation of interval reliability is affected to a greater or lesser degree depending on whether other responses occurred or did not occur in the two intervals. The calculation of proportional reliability is always affected because the response totals in the two intervals in question are different for the two observers, reducing the fraction in both intervals by $1/x$ ($x =$ the

number of responses scored in the interval). Finally, the most pronounced effect occurs with the exact calculation because two intervals are scored as disagreements, even though they are "off" by only one response.

A potential limitation of the present study is the somewhat arbitrary designation of response rates for the low-, moderate-, and high-rate sessions. It is possible, for example, that response rates as high as 25 per minute are observed infrequently and may be considered by some to be unrepresentative of a high rate of responding. However, in the absence of a set standard of classifying response rates, it was necessary to sample various response rates that included the low, middle, and high ends of a continuum. Future studies may examine a larger sample of response-rate parameters that range from very low (e.g., 0.1 response per minute) to very high (e.g., 40).

Another potential limitation is that all of the subjects in this study were trained observers who had at least 3 months of data-collection experience prior to their performance on the experimental task. Thus, it is possible that different results would be obtained with newly trained observers who may have more difficulty scoring high-rate (constant) responding and response bursts, although discrepancies among newly trained observers would reflect skill deficiencies rather than typical recording errors. However, differential results were obtained for the five experienced observers in the end-of-interval response condition. Reliability for Observers 8 and 9 was not affected greatly by end-of-interval responding, whereas the data for Observers 6, 7, and 10 showed that end-of-interval responses could influence reliability of even trained observers. The extent to which high-rate responding and response bursting influence scoring accuracy for novice observers should be examined in future research.

Finally, the generality of the present findings to the observation of human behavior during *in vivo* sessions, which may involve finer discriminations, detection of responding by more than

one sensory modality (i.e., scoring based on visual or auditory response characteristics), more distracters (i.e., the presence of nontarget behaviors and other individuals in the session environment), and potentially more target responses, was not examined. It is important to note, however, that the computer-generated data in this study permitted the isolation and precise control of single variables and ensured the uniformity of all response dimensions in all conditions. Future studies may examine the generality of the present findings by comparing reliability scores between videotaped sessions and computer-generated data that match video sessions in rate and distribution of target responses.

Results of the present research have several implications. First, proportional reliability seems to be a preferred index of reliability for time-based frequency recording because, unlike interval reliability, proportional reliability does not produce spuriously high scores for high-rate responding, and, unlike exact reliability, it is not overly influenced by single-response discrepancies. Second, in light of the deleterious effects produced by end-of-interval recording errors, observer training should incorporate procedures to improve rapid recording and include a specific check for this skill as a performance criterion. Finally, differences among reliability indices are influenced to some degree by the number of intervals that serve as the basis for calculation. For example, end-of-interval responding will have a greater effect on calculations based on a larger rather than a smaller number of intervals (given the same total session duration) simply because there are more opportunities for observers to score responses in adjacent intervals. The use of 10-s intervals for calculating reliability is based on the traditional 10-s interval recording procedure that is common in applied research. When the dependent variable is reported as the percentage of intervals during which responding occurred, the unit of measurement (10-s interval) used for reliability calculation matches

that used for data calculation. That is, agreement is assessed for the measure that is reported. By contrast, when data are taken on response frequency, the dependent variable usually is expressed as responses per minute, not responses per 10-s interval, and the relevant question from the standpoint of reliability is whether observers agree on the number of responses scored in a given minute. Thus, we suggest that the most appropriate unit (interval) for calculating exact and proportional reliability for frequency data is 1 min instead of 10 s.

REFERENCES

- Bailey, J. S., & Burch, M. R. (2002). *Research methods in applied behavior analysis*. Thousand Oaks, CA: Sage.
- Bijou, S. W., Peterson, R. F., & Ault, M. H. (1968). A method to integrate descriptive and experimental field studies at the level of data and empirical concepts. *Journal of Applied Behavior Analysis, 1*, 175–191. doi:10.1901/jaba.1968.1-175
- Bostow, D. E., & Bailey, J. S. (1969). Modification of severe disruptive and aggressive behavior using brief timeout and reinforcement procedures. *Journal of Applied Behavior Analysis, 2*, 31–37. doi:10.1901/jaba.1969.2-31
- Hopkins, B. L., & Hermann, J. A. (1977). Evaluating interobserver reliability of interval data. *Journal of Applied Behavior Analysis, 10*, 121–126. doi:10.1901/jaba.1977.10-121
- Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis, 10*, 141–150. doi: 10.1901/jaba.1977.10-141
- Mudford, O. C., Martin, N. T., Hui, J. K. Y., & Taylor, S. A. (2009). Assessing observer accuracy in continuous recording of rate and duration: Three algorithms compared. *Journal of Applied Behavior Analysis, 42*, 527–539. doi:10.1901/jaba.2009.42-527
- Page, T. J., & Iwata, B. A. (1986). Interobserver agreement: History, theory, and current methods. In A. Poling, R. W. Fuqua, & R. Ulrich (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 99–126). New York, NY: Plenum.
- Powell, J., & Rockinson, R. (1978). On the inability of interval time sampling to reflect frequency of occurrence data. *Journal of Applied Behavior Analysis, 11*, 531–532. doi:10.1901/jaba.1978.11-531
- Repp, A. C., Deitz, D. E. D., Boles, S. M., Deitz, S. M., & Repp, C. F. (1976). Differences among common methods for calculating interobserver agreement. *Journal of Applied Behavior Analysis, 9*, 109–113. doi:10.1901/jaba.1976.9-109
- Repp, A. C., Roberts, D. M., Slack, D. J., Repp, C. F., & Berkler, M. S. (1976). A comparison of frequency, interval, and time-sampling methods of data collection. *Journal of Applied Behavior Analysis, 9*, 501–508. doi: 10.1901/jaba.1976.9-501

Received September 8, 2006
 Final acceptance July 30, 2012
 Action Editor, Henry Roane