JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1977, 10, 97-101 NUMBER 1 (SPRING) 1977

A REVIEW OF THE OBSERVATIONAL DATA-COLLECTION AND RELIABILITY PROCEDURES REPORTED IN THE JOURNAL OF APPLIED BEHAVIOR ANALYSIS

MICHAEL BRYAN KELLY[1]

THE UNIVERSITY OF KANSAS

The research published in the *Journal of Applied Behavior Analysis* (1968 to 1975) was surveyed for three basic elements: data-collection methods, reliability procedures, and reliability scores. Three-quarters of the studies reported observational data. Most of these studies' observational methods were variations of event recording, trial scoring, interval recording, or time-sample recording. Almost all studies reported assessment of observer reliability, usually total or point-by-point percentage agreement scores. About half the agreement scores were consistently above 90%. Less than one-quarter of the studies reported that reliability was assessed at least once per condition.

DESCRIPTORS: behavioral recording in JABA, observational data, observational technology, observer agreement, reliability

The present study surveyed the observation and reliability characteristics of experimental data published in the *Journal of Applied Behavior Analysis* (*JABA*) from 1968 through 1975. The survey revealed that 222 (76%) of the 293 research reports published in the 8-yr history of *JABA* presented data collected by the observation of human subject's behavior. Just 16% reported only mechanically produced data, and 8% reported only permanent-product data, such as written academic tasks. Ninety-four per cent of the (222) *JABA*-published manuscripts that reported observational data also reported assessment of the reliability of the data collected. In 9% of the (222) studies, reliability checks were conducted each session, and in 23% checks were made in each condition. The remaining studies conducted less-frequent checks or did not specify when reliability was checked.

The names used for the data-collection procedures reported varied widely and were not relied on for reviewing purposes. Instead, the author's procedural descriptions were analyzed for inclusion in categories described later in this text. In reviewing reliability scores, 90% agreement was selected as an arbitrary benchmark for characterizing the levels of agreement. Authors represented percentage agreement scores in many forms, including means, ranges, and modes. The lowest scores reported was the one noted.[2] The results of the survey are shown in Tables 1 and 2.

Event Recording

A method of collecting the data used in 29% of *JABA's* research reports was event recording

[1]The author thanks Don Bushell, Jr., Eugene A. Ramp, and the Follow Through graduate students for their support during the development of this review. The preparation of this manuscript was supported in part by a grant (OEG-0-8-522422-4433) to the University of Kansas Support and Development Center for Follow Through. Reprints may be obtained from M. B. Kelly, Department of Human Development, University of Kansas, Lawrence, Kansas 66045.

[2]Frequently, studies were found to have reported multiple observational or reliability procedures. Because of this and the rounding of percentages, the per cent of studies reported as employing specific categories of procedures or controls may not always total 100%. The reliability of the present review was estimated by having a second person review 10% of the studies. Percentage agreement was computed separately for the variables of the review. Measurement technique: point-by-point, 90% (chance = 71%); occurrence, 55% (chance = 3%). Need for Effective Percentage Agreement in interval and time-sample studies: 100%. Minimum reliability scores: point-by-point, 74%. Scheduling of assessment: point-by-point, 80%. Reliability method: occurrence, 69% (chance—less than 2%). I am indebted to Ms. Jill Becker for serving as reliability reviewer.

Table 1

JABA Data Collection Survey

Information Reported	*Percentage of Research Reports* 293 *Published*	222 *with Observational Data*
Observational Data	76%	
Only Mechanically Collected Data	16%	
Only Permanent-Product Data	8%	
Reliability		94%
Reliability on Every Session		9%
Reliability in Each Condition		23%
Event Recording		29%
Trial Scoring		35%
Interval Recording		20%
Time-Sample Recording		21%
Response Duration		9%
Other		6%
Unidentifiable		2%

Table 2

JABA Reliability Survey

Reliability Procedure	*Percentage of Studies Reporting* *Reliability Procedure*	*90-100% Agreement*
Event Recording		
Total	45%	52%
Point-by-Point	31%	25%
Correlational	11%	
Unidentifiable	5%	
No Reliability	8%	
Trial Scoring		
Total	5%	100%
Point-by-Point	58%	69%
Amount of Difference	1%	
Correlational	1%	
Unidentifiable	31%	54%
Other	3%	
No Reliability	1%	
Interval and Time Sample		
Total	9%	40%
Point-by-Point	54%	45%
Occurrence or Nonoccurrence	36%	85%[a]
Lack Occurrence	33%	
Lack Nonoccurrence	35%	
Correlational	2%	
Cofunctional	1%	
Unidentifiable	5%	
No Reliability	0%	
Response Duration		
Total	45%	66%
Point-by-Point	10%	0%
Amount of Difference	10%	
Correlational	15%	
Cofunctional	5%	
Unidentifiable	5%	
No Reliability	15%	

[a]Percentage of studies reporting agreement from 75 to 100%.

or simply counting the number of responses observed during an observational period. In some studies, the record of this counting was made as a series of simple "hatch" marks. In other studies, code symbols differentiated between subjects or behaviors observed.

In the case of hatch-mark recording, a "total reliability" percentage of agreement score could be obtained by dividing the larger session total of one observer into the smaller session total obtained by the other. When the code-symbols were employed with event recording, observers' records could be compared for "point-by-point agreement", computed by dividing the number of specific notations the observer's records agreed on by the total number of symbols they both recorded plus the number of symbols only one or the other recorded. Unfortunately, with total reliability, random or sequence errors tend to be cancelled out and the reliability figures become spuriously inflated (Thomas, Becker, and Armstrong, 1968). Computing point-by-point reliability overcomes this problem.

The third method used by *JABA* authors to estimate reliability for event-recorded data was to use correlation coefficients, typically one of the product-moment correlations. Generally, whether to use a correlational method instead of a percentage agreement method depends on the observational method and response measure.

Less than one-half of the event-recording studies reported total reliability, while one-third reported the more stringent point-by-point reliability. The arbitrary minimum percentage score reviewed for either method was 90%, although scores can be somewhat inflated with total reliability. In fact, 52% of the total reliability studies met the 90% minimum, while only 25% of

the point-by-point studies did. Correlational reliability was reported in only 11% of the event-recording studies. In 5% of the event-recording studies the method of computing reliability could not be determined, while 8% of the event studies reported no reliability at all.

Trial Scoring

A second method, used in 35% of the *JABA* reports, was trial scoring which requires that an observer record the subject's responses as correct or incorrect trials, usually in relation to some stimulus. Only 5% of the trial-scoring studies reported total reliability, and all of these reported agreement above 90%. Point-by-point reliability was reported in 58% of the trial scoring studies, with 69% reporting better than a 90% minimum agreement score. A small number of studies reported correlational reliability, reported the second observer's data instead of reliability scores, or reported no reliability. Almost one-third of the trial-scoring studies reported reliability without indicating how it was computed. Scores above 90% were given for 54% of this group of studies.

Interval Recording and Time-Sample Recording

For interval recording, reported in 20% of the studies, the observational period is divided into brief segments or intervals and the observer notes whether a response occurs *during each interval.* Perhaps the most common interval length is 10 sec, although longer lengths are also used, such as 20 sec, 2 min, or 15 min. The study was categorized as using interval recording if it included an observational period divided into intervals, if the intervals were consecutive, and if the response was recorded once per interval when it occurred during any part of the interval. If there were breaks in the recording during which responses were not recorded, the study was classified as having used time-sample recording. Time-sample recording, used in 21% of the studies, is similar to interval recording, in that the observational period is divided into brief intervals. Time-sample recording differs in that the observer records the occurrence of a behavior only during designated portions of the observational period, such as at the end of 10-sec intervals if the behavior was observed during the tenth second of the interval (*e.g.,* Bailey, Wolf, Phillips, 1970) or if the occurrence of the response was observed over a 10-sec (*e.g.,* Pendergrass, 1972) or 20-sec (*e.g.,* Ward and Baker, 1968) interval, then recorded during the next 10 sec. Since both interval and time-sample recording employ time intervals that can be used to identify each particular observation and are amenable to the same reliability procedures, they were surveyed jointly.

When calculating percentage agreement with interval or sample data, either total or point-by-point (in this case, interval-by-interval) reliability could be used. Several authors have stated that point-by-point reliability could produce inflated reliability figures when the response under study occurred at either a very high or very low rate (Bijou, Peterson, Ault, 1968; Hawkins and Dotson, 1975; Kifer, Note 1). Generally, occurrence reliability is called for when the behavior under observation occurred at a low rate, and nonoccurrence when the behavior occurred at a high rate (Jensen, 1959).

Nine per cent of the interval recording and time-sampling recording studies reported total reliability, while 54% reported point-by-point reliability. Less than half of both of these groups of studies met the 90% minimum level of agreement. With an arbitrary cut-off point of 20% occurrence or less calling for occurrence and 80% or more calling for nonoccurrence reliability, it was determined from the author's data that 35% of the studies should have employed nonoccurrence reliability but did not, and 33% should have employed occurrence reliability but did not. In fact, 36% reported either form of effective percentage agreement. Nearly all of these met an arbitrary 75% minimum percentage of agreement.[3]

[3]Effective percentage agreement scores may, of course, be very low if the rate of behavior is extreme. If a response were recorded twice in 10 intervals by

Two per cent of the interval and time-sample studies reported correlational reliability. A few reported cofunctional reliability, defined by Goldiamond (1968, p. 117) as two observers' data, graphed over the course of a study, forming similar functions.[4] Some interval and sample studies gave insufficient information to determine how agreement scores were computed.

Response Duration

Response duration, reported in 9% of the studies, describes the amount of time the subject responded. The larger amount of time, collected by either observer, can be divided into the smaller amount, collected by the remaining observer, yielding a total reliability score. If the response-duration data were collected for individual, identified instances of a response, instead of for session totals, then more stringent point-by-point reliability could be calculated (*e.g.,* Scott and Bushell, 1974). Two other types of reliability reported in *JABA* response-duration studies were the amount of difference between two observers' scores (*e.g.,* Hopkins, Schutte, and Garton, 1971), and correlational methods of estimating observer agreement (*e.g.,* Skiba, Pettigrew, and Alden, 1971).

Of all the *JABA* response-duration studies, 45% reported total reliability and 10% reported point-by-point reliability. Two-thirds reporting total reliability met the 90% agreement level, while none of the point-by-point studies did so. A few studies each reported either amount of difference, correlational, or cofunctional reliability. One response-duration study reported 100% agreement, with no reference to the type of reliability utilized. Fifteen per cent of the response-duration studies reported no reliability whatsoever.

Other

Six per cent of the 222 observational studies in *JABA* reported methods of data collection that did not fit any of the preceding categories. These included weighing items and reading decibel meters. Two per cent of the *JABA* observational reports did not describe the methods employed in the data-collection process.

CONCLUSION

The *JABA* "Preparation of Manuscripts" statement points out that "in many instances human observation may be the only current recording technique. In such cases, however, an analysis of reliability between independent observers should be included (1969, p. 1)." This review demonstrated that 76% of the *JABA* reports reviewed did indeed employ human observation as a recording techniques, and 94% of these studies reported reliability analysis. Variations on a handful of techniques dominated the collection of procedures authors have used.

REFERENCE NOTE

1. Kifer, R. E. *A review of reliability in applied behavioral research.* Unpublished manuscript. 1974. (Available from the Department of Human Development, University of Kansas, Lawrence, Kansas 66045).

REFERENCES

Bailey, J. S., Wolf, M. M., and Phillips, E. Home-based reinforcement and the modification of predelinquents' classroom behavior. *Journal of Applied Behavior Analysis,* 1970, **3,** 223-233.

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.

one observer, but only once (in agreement) by the second, occurrence reliability would be only 50%. It should be kept in mind that, at 10% to 20% response occurrence levels, *chance* occurrence agreement would range from 1% to 4%.

[4]Cofunctional plots offer the advantages of showing on which sessions, and hence how frequently and at what levels of responding, reliability checks are made. Its principal disadvantage is that it includes no check on agreement of individual observations (similar in this respect to total and amount of difference reliability). The reliability observer, knowing what level of responding had become typical or was expected under a given condition, could merely report data within the expected range and the result would be cofunctional agreement.

Goldiamond, I. Stuttering and fluency as manipulatable operant response classes. In L. Krasner and L. P. Ullmann (Eds.), *Research in behavior modification,* New York: Holt, Rinehart & Winston, 1968. Pp. 106-156.

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.

Hopkins, B. L., Schutte, R. C., and Garton, K. L. The effects of access to a playroom on the rate and quality of printing and writing of first- and second-grade students. *Journal of Applied Behavior Analysis,* 1971, **2,** 77-78.

Jensen, A. R. The reliability of projective techniques: Methodology. *Acta Psychologica,* 1959, **16,** 32-67.

Pendergrass, V. Timeout from positive reinforcement following persistent, high-rate behavior in retardates. *Journal of Applied Behavior Analysis,* 1972, **1,** 85-92.

Preparation of manuscripts, *Journal of Applied Behavior Analysis,* 1969, **1,** 1-2.

Scott, J. and Bushell, D. The length of teacher contacts and students off-task behavior. *Journal of Applied Behavior Analysis,* 1974, **1,** 39-44.

Skiba, E., Pettigrew, L., and Alden, S. A behavioral approach to the control of thumbsucking in the classroom. *Journal of Applied Behavior Analysis,* 1971, **2,** 121-128.

Thomas, D., Becker, W., and Armstrong, M. Production and elimination of disruptive classroom behavior by systematically varying teacher's behavior. *Journal of Applied Behavior Analysis,* 1968, **1,** 35-45.

Ward, M. and Baker, B. Reinforcement therapy in the classroom. *Journal of Applied Behavior Analysis,* 1968, **4,** 323-328.

Received 17 March 1975.
(Final acceptance 15 May 1976.)