CONTINUOUS RECORDING AND INTEROBSERVER AGREEMENT ALGORITHMS REPORTED IN THE JOURNAL OF APPLIED BEHAVIOR ANALYSIS *(1995–2005)*

OLIVER C. MUDFORD AND SARAH ANN TAYLOR

UNIVERSITY OF AUCKLAND, NEW ZEALAND

AND

NEIL T. MARTIN

TREEHOUSE TRUST, LONDON

We reviewed all research articles in 10 recent volumes of the *Journal of Applied Behavior Analysis* (*JABA*): Vol. 28(3), 1995, through Vol. 38(2), 2005. Continuous recording was used in the majority (55%) of the 168 articles reporting data on free-operant human behaviors. Three methods for reporting interobserver agreement (exact agreement, block-by-block agreement, and time-window analysis) were employed in more than 10 of the articles that reported continuous recording. Having identified these currently popular agreement computation algorithms, we explain them to assist researchers, software writers, and other consumers of *JABA* articles.

DESCRIPTORS: computers, continuous recording, interobserver agreement, observational data, recording and measurement

It has been over 30 years since Kelly's (1977) initial review of data-collection and interobserver agreement methods in the *Journal of Applied Behavior Analysis* (*JABA*). Kelly found that the majority (76%) of research articles published from 1968 to 1975 used pencil-and-paper methods for discontinuous recording of behavioral observations (e.g., interval recording and time sampling). Following Kelly's review, there have been a series of investigations of the merits of various methods of data collection in *JABA*, including bias of interval recording and random error of time sampling (e.g., Powell, Martindale, & Kulp, 1975) and problems with interobserver agreement (e.g., Repp, Deitz, Boles, Deitz, & Repp, 1976).

Discontinuous methods do not allow the basic dimensions of behaviors to be quantified accurately in standard scientific units (e.g., rates in responses per minute, durations, interresponse times, and latencies in seconds; Hanley, Cammilleri, Tiger, & Ingvarsson, 2007; Johnston & Pennypacker, 1993). Continuous recording is required for direct measurement of the basic dimensions of behaviors. The availability of handheld portable electronic data-entry and storage devices has increased the practicality and affordability of continuous recording for research and clinical purposes. The aims of this review were to determine the relative frequencies of continuous and discontinuous recording methods in *JABA* articles over 10 recent volumes (1995 to 2005) and to quantify variations in methods for assessment of the reliability (i.e., interobserver agreement and accuracy) of continuously recorded behavioral data.

Sarah Ann Taylor is now at Odyssey House, Auckland, New Zealand.

A portion of this paper was presented at the third international conference of the Association for Behavior Analysis, Beijing, China, November, 2005. Contributions to the funding of this research were received from the University of Auckland (UoA) Research Committee (first author) and the UoA Faculty of Science Summer Scholarship Programme (second author). We are grateful for explanatory correspondence with Wayne Fisher, Greg Hanley, Brian Iwata, and SungWoo Kahng.

Address correspondence to Oliver C. Mudford, Applied Behaviour Analysis Programme, Department of Psychology, University of Auckland (Tamaki Campus), Private Bag 92019, Auckland 1142, New Zealand (e-mail: o.mudford@auckland.ac.nz).

doi: 10.1901/jaba.2009.42-165

METHOD

Selection of Articles for Review

All research articles published in 10 years of *JABA*—Vol. 28(3), 1995, to Vol. 38(2), 2005—were examined by the first author. Reviews, discussion articles, and reports (abbreviated research articles) were not included. The second author acted as an independent reviewer throughout for the purposes of estimating interobserver agreement (described below).

Review Procedure

Review was conducted in four stages. First, all articles were retained for closer examination that reported at least some direct observation data, either in vivo or from video, of free-operant human behavior. Thus, papers that included only automatically (mechanically or electronically) recorded data were excluded, as were articles that reported on restricted-operant behaviors only (e.g., data from trial-by-trial teaching or from researcher-controlled bite-by-bite behaviors in eating-related studies) and two research articles concerning animal behaviors. This resulted in a total of 168 articles.

Second, the retained articles were reviewed to determine whether they reported continuous or discontinuous data. Continuous data collection was identified by applying the following definition: Researchers described observational records that contained second-by-second records of occurrences of discrete behaviors or the onsets and offsets of behaviors with duration, and the results were reported in standard units of measurement or their derivatives (e.g., responses per minute, percentage of observation session). Discontinuous methods were defined as data-collection procedures that recorded behaviors in time samples or intervals of more than 1 s. The 93 articles that included continuous data were reviewed further.

Third, the articles remaining were examined to determine whether obtained continuous data were analyzed to produce frequency (or rate) measures, duration measures, or both. Fourth, these articles were again scrutinized to ascertain what algorithms were used to assess the reliability of the data (e.g., block-by-block agreement; Bailey & Bostow, cited in Page & Iwata, 1986; Bailey & Burch, 2002), exact agreement (Repp et al., 1976), time-window analysis (MacLean, Tapp, & Johnson, 1985; Tapp & Wehby, 2000), or others.

Interobserver Agreement of Review

The interobserver agreement for the review process was assessed using a stratified procedure in which approximately 20% of articles at each level were selected randomly for independent examination by the second author. Percentage agreement at the first three levels of review was calculated by dividing number of agreements between reviewers by number of articles examined by both reviewers and converting this ratio to a percentage. Mean interobserver agreements were 100%, 96%, and 95% on 51, 36, and 21 articles subjected to assessment, respectively.

A different procedure was used for checking the first author's identification of interobserver agreement calculation methods, the fourth level of review. Although the time-window analysis was identifiable with 100% agreement, published descriptions of the block-by-block and exact agreement algorithms differed across articles such that agreement assessment was not considered as a valid surrogate for accuracy of identification of these algorithms. Therefore, senior authors whose work had used one or both of these algorithms and had been previously published in *JABA* were contacted by the first author. Confirmation of the calculation methods was obtained from each author. Although it was an unconventional method for assessing interobserver agreement of review, this provided an accurate means of confirming our identification of the algorithms.

RESULTS AND DISCUSSION

All 256 *JABA* research articles published from mid-1995 to mid-2005 were reviewed. Of

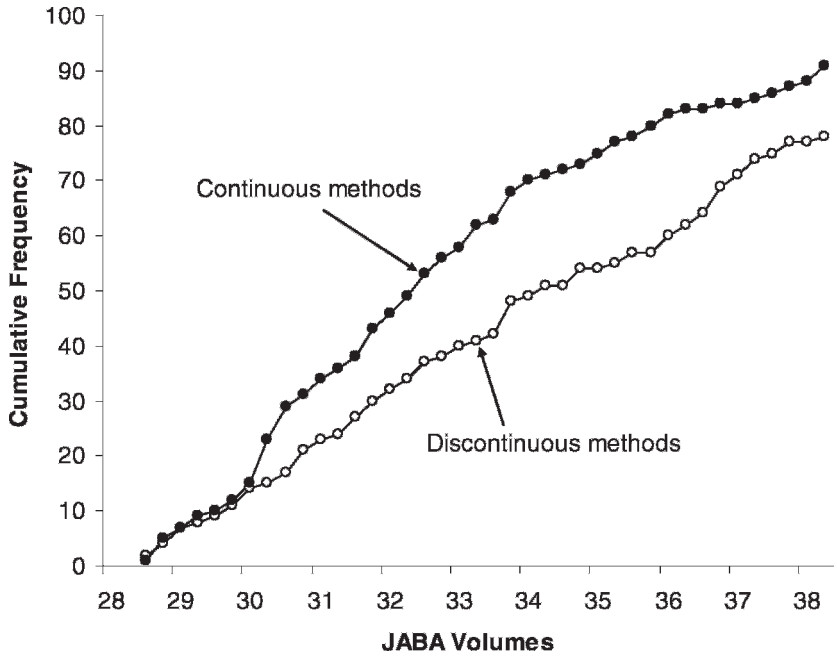

Figure 1. Cumulative frequencies of *JABA* research articles from 1995 to 2005, reporting data on free-operant human behaviors that were recorded by observers using only discontinuous recording procedures and those reporting continuously recorded data.

these, 168 reported direct observational data from free-operant human behavior. Of these 168, 93 articles (55%) reported continuously recorded data. Discontinuous methods for recording such behaviors have been superseded in published research applications of behavior analysis. Figure 1 shows the rates of use of continuous and discontinuous methods for observational recording in research articles across the 10 years reviewed. Among the 93 articles reporting continuously recorded free-operant human behaviors, 88 (95%) reported frequency measures (usually rate of responding per minute). Duration measures were reported in 33 articles (36%).

All articles that contained continuously recorded data reported interobserver agreement data; none reported observer accuracy measures. Thus, interobserver agreement has continued to be the method by which the quality of behavioral data is assessed (as in Kelly, 1977). Three methods for computing agreement predominated (i.e., were reported over 10 times) in the articles reviewed: block-by-block agreement, exact agreement, and time-window analysis. Figure 2 shows the cumulative frequency of articles that reported using the three methods. The data should not be interpreted to suggest that one method is preferable because it was used more often than another (e.g., the block-by-block method was reportedly used 46 times, three times more often than the time-window analysis method). The frequency of use may be indicative of the publication rates of research groups that chose to employ the different methods. It was noted during review that computational methods for interobserver agreement were not always fully described or consistently named. Therefore, we provide a detailed explanation of the three most popular algorithms identified during our review.

The exact and block-by-block agreement methods were developed for use with discontinuously recorded data. They are similar in that

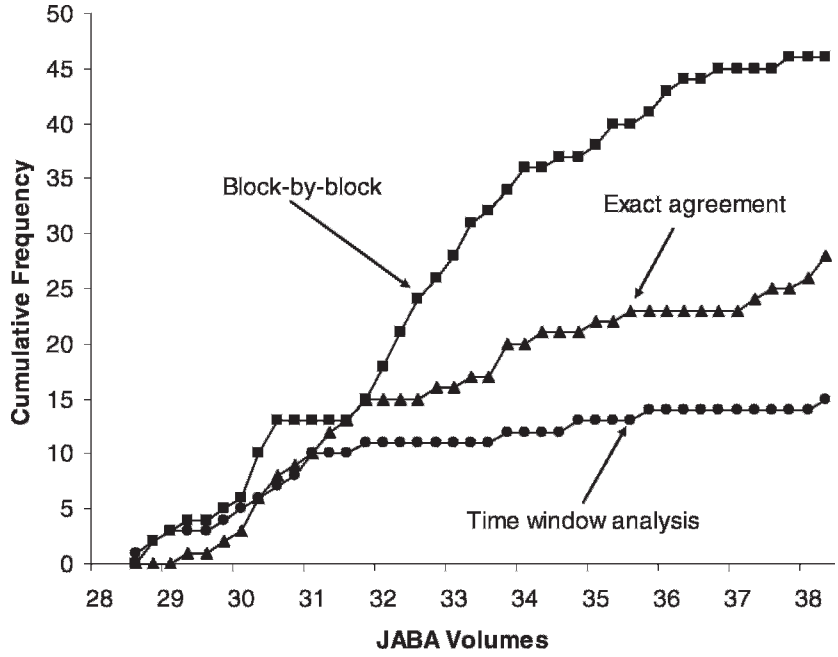

Figure 2. Cumulative frequencies of the three commonly reported interobserver computation methods for continuous recording in *JABA* research articles from 1995 through 2005.

the second-by-second data streams from two observers have 10-s intervals superimposed. The level of analysis for discrete data (events) is the number of occurrences of the behavior recorded in a 10-s interval. With duration measures, the number of seconds within a 10-s interval that the behavior was recorded as occurring is counted for each observer (e.g., Hagopian, Contrucci-Kuhn, Long, & Rush, 2005; Rapp, Vollmer, St. Peter, Dozier, & Cotnoir, 2004).

The exact agreement method is described fully in Piazza, Hanley, and Fisher (1996, p. 440):

> Exact agreement coefficients were calculated by partitioning each session into 10-s intervals. In each interval, two observers could agree on the exact number of behaviors that occurred, agree that behavior did not occur, or disagree about the exact number of behaviors that occurred (disagreement). … Coefficients were calculated by dividing the number of agreements by the sum of agreements plus disagreements and multiplying by 100%.

Repp et al. (1976) identified this formula as the exact agreement (all intervals) method because it includes agreements on non occurrence in the calculation.

In the block-by-block method, the smaller of the two observers' totals in a 10-s interval is divided by the larger. This provides a score between 0 and 1 for every interval. For the purposes of calculating agreement when both observers scored no occurrences during an interval, such intervals are scored as 1. Scores are summed across all intervals and divided by the number of intervals, and the result is converted to a percentage to provide a percentage agreement index. This description of calculating block-by-block agreement applies to all research articles in our review identified as using the method, and it was confirmed by users of the algorithm. As used, the method deviates from the computation explained by Bailey and Burch (2002), in which intervals of agreement on nonoccurrence (zero divided by zero) are ignored. The method in common use could be labeled block-by-block (all intervals) method to differentiate it from similar algorithms.

Computation of percentage agreement using time-window analysis was devised for continuously recorded data. One-second intervals are

imposed on two observers' data streams, and second-by-second comparisons are made between them. When both records show an event (for discrete behaviors) or a second of ongoing occurrence (for behaviors measured with duration), this is counted as an agreement. Any second in which only one record contains an event or occurrence of behavior is a disagreement. Percentage agreement is calculated by dividing the number of agreements by the number of agreements plus disagreements. MacLean et al. (1985) recognized that their algorithm was overly stringent for data on discrete events. Consequently, they recommended allowing tolerance for counting agreements by expanding the definition of an agreement to include observations when one observer recorded an event within $\pm$ t seconds of the other observer. In research articles sampled, t has varied from 1 s (e.g., Romaniuk et al., 2002) to 5 s (e.g., Lalli, Mauro, & Mace, 2000, Experiment 3).

Compared with the extensive methodological studies on discontinuous recording, there has been little research effort to comprehend, evaluate, or guide selection of methods for assessment of data quality with continuous recording. There have been recommendations for evaluating interobserver agreement with continuous data (e.g., Hollenbeck, 1978; MacLean et al., 1985) but no methodological studies have compared different methods in use. The results of this review suggest that continuous recording is a timely topic for methodological study.

REFERENCES

Bailey, J. S., & Burch, M. R. (2002). *Research methods in applied behavior analysis*. Thousand Oaks, CA: Sage.

Hagopian, L. P., Contrucci-Kuhn, S. A., Long, E. S., & Rush, K. S. (2005). Schedule thinning following communication training: Using competing stimuli to enhance tolerance to decrements in reinforcer density. *Journal of Applied Behavior Analysis*, *38*, 177–193.

Hanley, G. P., Cammilleri, A. P., Tiger, J. H., & Ingvarsson, E. T. (2007). A method for describing preschoolers' activity preferences. *Journal of Applied Behavior Analysis*, *40*, 603–618.

Hollenbeck, A. R. (1978). Problems of reliability in observational research. In G. P. Sackett (Ed.), *Observing behavior: Vol. 2. Data collection and analysis methods* (pp. 79–98). Baltimore: University Park Press.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Lalli, J. S., Mauro, B. C., & Mace, F. C. (2000). Preference for unreliable reinforcement in children with mental retardation: The role of conditioned reinforcement. *Journal of Applied Behavior Analysis*, *33*, 533–544.

MacLean, W. E., Tapp, J. T., & Johnson, W. L. (1985). Alternate methods and software for calculating interobserver agreement for continuous observation data. *Journal of Psychopathology and Behavioral Assessment*, *7*, 65–73.

Page, T. J., & Iwata, B. A. (1986). Interobserver agreement: History, theory and current methods. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 99–126). New York: Plenum.

Piazza, C. C., Hanley, G. P., & Fisher, W. W. (1996). Functional analysis and treatment of cigarette pica. *Journal of Applied Behavior Analysis*, *29*, 437–450.

Powell, J., Martindale, A., & Kulp, S. (1975). An evaluation of time-sample measures of behavior. *Journal of Applied Behavior Analysis*, *8*, 463–469.

Rapp, J. T., Vollmer, T. R., St. Peter, C., Dozier, C. L., & Cotnoir, N. M. (2004). Analysis of response allocation in individuals with multiple forms of stereotyped behavior. *Journal of Applied Behavior Analysis*, *37*, 481–501.

Repp, A. C., Deitz, D. E. D., Boles, S. M., Deitz, S. M., & Repp, C. F. (1976). Technical article: Differences among common methods for calculating interobserver agreement. *Journal of Applied Behavior Analysis*, *9*, 109–113.

Romaniuk, C., Miltenberger, R., Conyers, C., Jenner, N., Jurgens, M., & Ringenberg, C. (2002). The influence of activity choice on problem behaviors maintained by escape versus attention. *Journal of Applied Behavior Analysis*, *35*, 349–362.

Tapp, J., & Wehby, J. H. (2000). Observational software for laptop computers and optical bar code readers. In T. Thompson, D. Felce, & F. J. Symons (Eds.), *Behavioral observation: Technology and applications in developmental disabilities* (pp. 71–81). Baltimore: Brookes.

Received January 18, 2007
Final acceptance January 25, 2008
Action Editor, Mark Dixon