Behavior Analysis in Practice (2020) 13:411–420
https://doi.org/10.1007/s40617-019-00361-6

RESEARCH ARTICLE

Procedures and Accuracy of Discontinuous Measurement of Problem Behavior in Common Practice of Applied Behavior Analysis

Linda A. LeBlanc[1,2] · Coby Lund[1] · Chris Kooken[1] · Janet B. Lund[1] · Wayne W. Fisher[3]

Published online: 3 June 2019

Abstract

Discontinuous measurement involves dividing an observation into intervals and recording whether a behavior occurred during some or all of each interval (i.e., interval recording) or at the exact time of observation (i.e., momentary time sampling; MTS). Collecting discontinuous data is often easier for observers than collecting continuous data, but it also produces more measurement error. Smaller intervals (e.g., 5 s, 10 s, 15 s) tend to produce less error but may not be used in everyday practice. This study examined the most common intervals used by a large sample of data collectors and evaluated the effect of these intervals on measurement error. The most commonly used intervals fell between 2 and 5 min. We then analyzed over 800 sessions to evaluate the correspondence between continuous and discontinuous data at each commonly used interval. Intervals of 3 min or less produced the greatest correspondence, and MTS outperformed interval recording.

Keywords Data collection · Discontinuous measurement · Partial interval · Momentary time sample · Observation interval

The systematic measurement of behavior is foundational to the field of behavior analysis (Baer, Wolf, & Risley, 1968; Cooper, Heron, & Heward, 2007; Sidman, 1960). Data collection and visual inspection of data are so integral to the field that procedures for measurement and visual inspection are frequently the subject of experimental analysis (Jessel, Metras, Hanley, Jessel, & Ingvarsson, in press; Machado, Luczynski, & Hood, 2019; Saini, Fisher, & Retzlaff, 2018). A practitioner's choices about how direct-observation data are collected impact other important decisions, such as the determination of the function of problem behavior in a functional analysis and decisions about when to implement or change interventions. LeBlanc, Raetz, Sellers, and Carr (2016) suggested that there are no valid circumstances under which applied behavior analysis should be practiced without the collection of meaningful data.

Continuous measurement refers to procedures that capture each event that occurs during an observation using a measure that fully represents at least one relevant response dimension (e.g., duration, intensity; Cooper et al., 2007; Johnston & Pennypacker, 2009). Frequency is a common measure of problem behavior for discrete, countable responses that are somewhat uniform with respect to the duration of each event (e.g., aggression, self-injurious behavior; see Beavers, Iwata, & Lerman, 2013, for a recent review of studies on functional analysis of problem behavior, including the most commonly used measures). Duration is often used for responses for which it is difficult to discern the beginning of one event from the end of the prior event (e.g., stereotypic behavior) or that vary with respect to the duration of each event (e.g.,

Research Highlights
- Practitioners often use discontinuous measurement intervals that are longer than recommended in the empirical literature.
- Intervals and samples longer than 3 min produced error on every metric examined, whereas those at or below 3 min were generally quite accurate.
- The default settings in an electronic data-collection system were often the ones used in programming.
- The level of problem behavior in the session systematically impacted the correlations between discontinuous and continuous measures with very small amounts of problem behavior producing the lowest correlations at all values.
- Practitioners should use small intervals or samples for discontinuous measurement and should not use discontinuous measurement of any type for very low amounts of problem behavior.

✉ Linda A. LeBlanc

[1] DataFinch Technologies, Atlanta, GA, USA

[2] Golden, USA

[3] Monroe Meyer Institute, University of Nebraska Medical Center, Omaha, NE, USA

tantrums), making duration an important response dimension to measure (Gardenier, MacDonald, & Green, 2004).

Discontinuous measurement refers to procedures that use a sampling observational procedure to estimate the amount of behavior that actually occurred (Johnston & Pennypacker, 2009). The most common type of discontinuous measurement procedure used for problem behavior in the published literature is partial-interval recording (PIR; Beavers et al., 2013), but momentary time sampling (MTS) is also a viable option. Partial-interval measures require the observer to score each interval in which any instance of the behavior occurred for any duration of time. The observational system can be arranged for the observer to score as soon as any behavior occurs during the interval or at the end of the interval. Time-sampling procedures require the observer to score whether the target behavior occurred at specified moments in time (e.g., at 30 s, 60 s, 90 s) without regard for whether it occurred between observation moments. Whole-interval measures require the observer to score each interval in which the problem behavior occurred during the entire interval. However, whole-interval recording measures underestimate the level of actual behavior and are not recommended for use when measuring problem behavior (Fiske & Delmolino, 2012; LeBlanc et al., 2016), as they may lead the clinician to overestimate the effectiveness of a treatment.

Continuous measurement is preferred over discontinuous measurement, when possible, as discontinuous measures include some amount of sampling error and can over- or underestimate the level of behavior (Fiske & Delmolino, 2012; Gardenier et al., 2004; Johnston & Pennypacker, 2009). This preference for continuous measurement systems is clearly reflected in the applied research literature on problem behavior (e.g., Beavers et al., 2013; over 60% of published studies on functional analysis used continuous measurement systems). However, discontinuous measures are often easier to implement, which can be helpful in applied settings when the data collector is also serving as the teacher or implementer, or when multiple or very high-rate behaviors are scored (Fiske & Delmolino, 2012; LeBlanc et al., 2016). In general, smaller intervals (e.g., 6 s, 10 s) introduce less error into the data-collection system than longer intervals do (Alvero, Struss, & Rappaport, 2008; Powell, 1984), but intervals that are too small (i.e., below 3 s; Hanley, Cammilleri, Tiger, & Ingvarsson, 2007) increase the difficulty of data collection and may result in poor interobserver agreement. In addition, the total observation duration (Mudford, Beale, & Singh, 1990) may differentially impact PIR and MTS data-collection systems for different levels of behavior (i.e., high rate/long duration, low rate/short duration; Harrop & Daniels, 1986; Saudargas & Zanolli, 1990). Thus, each of these various aspects of the design of the measurement system can impact whether the data that are collected are meaningful and appropriate for guiding clinical decision-making.

The majority of systematic behavior-analytic research studies on PIR and MTS for problem behavior have investigated the accuracy of relatively short intervals (e.g., 1 s to 6 min) for relatively brief observation sessions (e.g., 6–10 min). Studies have consistently found that 5-s or 10-s intervals provide reasonable estimates of behavior with reasonable interobserver agreement and that PIR systematically overestimates behavior compared to MTS (Harrop & Daniels, 1986). Hanley et al. (2007) found little difference in estimation error for MTS intervals between 5 s and 120 s (i.e., 2 min). Guntner, Venn, Patrick, Miller, and Kelly (2003) also found that a 2-min MTS produced a reasonable estimate of behavior for students in a classroom setting, whereas 4-min and 6-min intervals resulted in data paths that differed substantially from continuous data. In a series of studies examining sensitivity to change, Rapp et al. found that 10-s PIR detects the most change in frequency events, whereas 10-s MTS detects the most change in duration events (Schmidt, Rapp, Novotny, & Lood, 2013). In addition, MTS intervals up to 60 s detected change for both measures for longer observation intervals of 30 min and 60 min (Devine, Rapp, Testa, Henrickson, & Schnerch, 2011). Kolt and Rapp (2014) evaluated data-collector preference for 10-s and 1-min intervals for both PIR and MTS. Therapists preferred a 1-min MTS procedure to either 10-s MTS or PIR value because it was easier (i.e., required less vigilance) and less stressful. Some participants stated that their preference for the longer MTS procedure was partially because they could potentially do other things in between recording, which is important when one is simultaneously fulfilling other responsibilities (e.g., implementing teaching procedures) while collecting data (Gardenier et al., 2004).

Although this information is useful in guiding the measurement selection of applied researchers in resource-rich settings, it may be less relevant to the experience of practitioners in the majority of human-service settings. In many practice settings, it may not be feasible to have an interval of 2 min or less, and the observation session may often be significantly longer than 5–15 min, as is the duration of observation for most published functional analysis and treatment studies (Beavers et al., 2013). For example, Morris et al. (in press) published a technical article describing procedures for creating a 60-min PIR data-collection system in Excel that can be used during all waking hours (e.g., 16–18 hr) in human-service settings, such as group homes. However, hour-long intervals may introduce significant estimation error into the obtained data, leading to faulty data-based decisions and a limited ability to detect treatment effects. Though the same instructions could be used to set the interval to any length, the authors' use of 60 min as an example in the instructions could lead practitioners who read the article to follow that example if they are not well versed in the literature on discontinuous measurement. A lack of familiarity with this literature and the desire to minimize the response effort for data collection could lead practitioners to

use intervals that are longer than those examined in published studies. This decision might be made in spite of the fact that long intervals might jeopardize the quality of the obtained data. The purpose of Study 1 was to examine a large database to identify the most commonly used intervals for PIR and MTS data in applied behavior analysis service settings. The purpose of Study 2 was to examine the correlations of discontinuous measurement with longer intervals with continuous measurement using electronic data from actual (i.e., not simulated) sessions as a means to determine the level of estimation error evident with the commonly used intervals.

Study 1

Method

Catalyst Product and Database Catalyst is a commercially available electronic data-collection tool designed to assist applied behavior analysts with the capture and analysis of large quantities of data. Catalyst users (e.g., Board Certified Behavior Analysts) create a profile for each patient and establish programs and data-collection procedures for problem behavior and/or skill acquisition programs. Data collection occurs using real-time data-stamping methodology so that data can be examined to the second that the data were collected, rather than just as a summary metric. For problem behavior, the user creates an operational definition, selects from among various continuous (e.g., frequency, duration) and discontinuous measurement systems (e.g., PIR, MTS), and programs the length of the interval for discontinuous systems (e.g., 10 s, 30 s, 2 min) from a drop-down menu in the portal (see Figure 1, top panel). All topographies of problem behavior with the same discontinuous measurement system have the same interval setting (i.e., the interval is set by patient rather than by topography). Typically, technicians (e.g., registered behavior technicians) then use a portable electronic device (e.g., iPad) to capture data throughout their ongoing therapy sessions. When a discontinuous measurement system is used, an auditory or vibratory stimulus (setting selected by user) signals the end of the interval and the technician then scores whether each problem behavior is currently occurring (i.e., MTS) or has occurred at all since the last signal or for the entire interval (i.e., PIR; see Figure 1, middle and bottom panels).

Customers of Catalyst could choose whether to have their data included in research analyses, and agencies that declined participation were excluded from all analyses. We included data from over 700 agencies and 30,000 patients. Approximately 80% of patients were male, and autism was the most commonly reported diagnosis. Approximately 88% of patients were age 15 or younger, but the sample included older adolescents and adults as well. Approximately 70% of patients had continuous measurement systems in place, but those who had discontinuous measurement systems in place were included in the analysis to identify the most commonly selected intervals for measuring problem behavior. Data were analyzed separately for PIR and MTS measures.

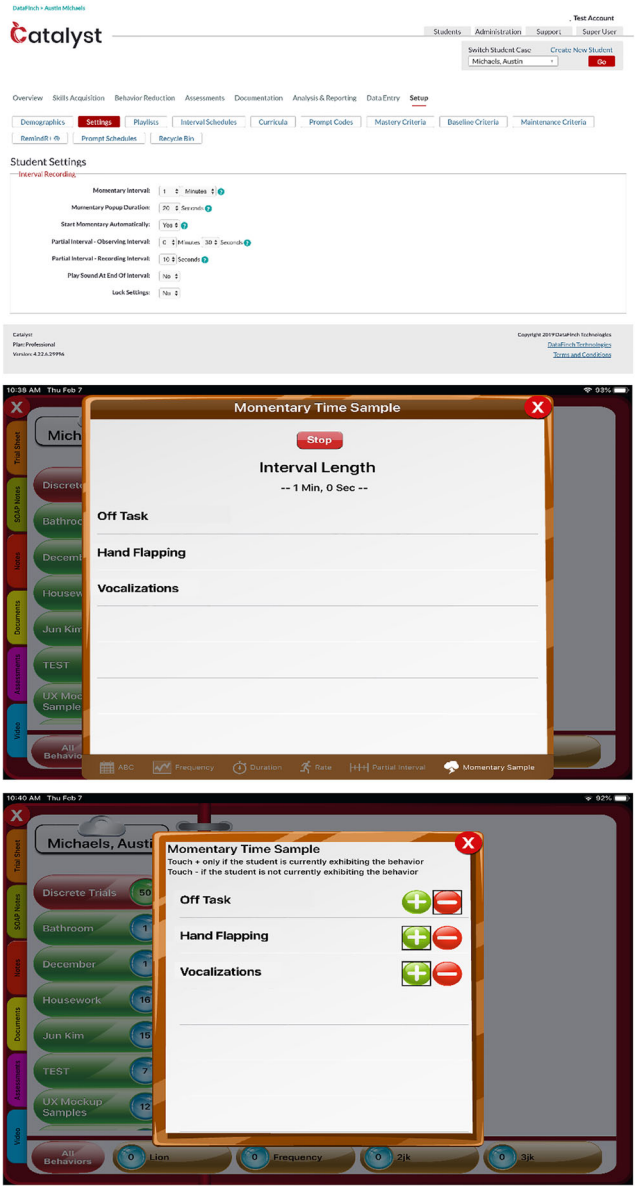

Fig. 1 The portal is used to create a discontinuous data-collection system in Catalyst (top panel), and the portable application is used to start the feature (middle panel) and collect the data (bottom panel)

Data Inclusion and Procedures We identified all patients in the database with either an MTS or PIR measure for problem behavior. We reviewed the data for those patients to ensure that multiple sessions with measures of problem behavior had been collected, and we excluded those patients with fewer than 50 unique data points. That is, we excluded patients where a discontinuous measurement system had been created but was not actually used for regular, ongoing measurement of

problem behavior. We captured the value of the MTS or PIR interval for each patient and imported that value into an Excel spreadsheet. The interval values were sorted from shortest to longest. For each value, the total number of patients with that assigned value was tabulated and then divided by the total number of patients to calculate the percentage of patients with that interval.

Results and Discussion

Table 1 shows the most commonly used intervals for discontinuous measurement procedures, with the most frequently used intervals depicted in bold for each measure. For PIR, the drop-down menu provided a default value of 5 min and the user could slide the cursor up or down to select an alternative value. The database included interval values for 1,834 patients. Forty-four unique interval values had been used, and the range of selected intervals was 1 s to 3,600 s (i.e., 1 hr). The most commonly used interval was the default in the drop-down menu (i.e., 5 min; 50%) with a steep drop-off in use to the second-most commonly used interval of 2 min (12%). A total of 36% of patients had an assigned interval value at or below the value identified in studies as the longest interval producing reasonable estimates of the actual occurrence of behavior (i.e., 2 min; Guntner et al., 2003; Hanley et al., 2007).

For MTS, the default selection in the drop-down menu was 2 min. Sampling interval values were available for 1,320 patients. Thirty-one unique values had been used, and the range of selected intervals was 1 s to 7,200 s (i.e., 2 hr). The most commonly used interval was the default value from the drop-down menu (i.e., 2 min; 37%) with a decrease to 19% for the second-most commonly used interval of 5 min. A total of 64% of patients had an assigned MTS interval value at or below the value identified in studies as the longest interval producing reasonable estimates of the actual occurrence of behavior (i.e., 2 min; Guntner et al., 2003; Hanley et al., 2007). The eight most commonly used intervals were the same for the two measures, although the rankings of these intervals varied (e.g., 15 min ranked third for PIR but tied for eighth for MTS).

Table 1 Ten most frequently used interval and sample sizes for discontinuous measurement procedures in the catalyst database

Interval/Sample size	Cases With PIR (% of 1,834)	Cases With MTS (% of 1,320)
10 s	69 (4%)	32 (2%)
15 s	26 (1%)	13 (1%)
20 s	24 (1%)	9 (1%)
30 s	84 (5%)	131 (10%)
60 s	222 (12%)	172 (13%)
120 s	230 (12%)	490 (37%)*
180 s	41 (2%)	73 (5%)
300 s	928 (50%)[a]	246 (19%)
600 s	91 (5%)	94 (7%)
900 s	105 (6%)	41 (3%)
1,800 s	14 (1%)	19 (1%)

[a] Indicates that this value was the default in the drop-down menu

Similar to the findings of Beavers et al. (2013) and Mudford, Taylor, and Martin (2009), Catalyst users employed continuous measurement procedures more often than discontinuous measures, and users implemented PIR more frequently than MTS. However, the most commonly used intervals in practice differed from those most commonly examined intervals in research studies. Only 2% of patients had an assigned MTS interval at or below 10 s, and only 4% of patients had an assigned PIR interval at or below 10 s (i.e., the most commonly used interval in studies on problem behavior published in the *Journal of Applied Behavior Analysis* and in studies on discontinuous measurement). The vast majority of patient intervals were set well below the 1-hr interval described by Morris et al. (in press), with six patients (0.4%) with an assigned MTS interval at 1 hr or longer and only six patients (0.3%) with an assigned PIR interval at 1 hr or longer.

The default values in the drop-down menus were the most commonly selected values for their respective measurement types (i.e., 50% of PIR users selected the 5-min default; 37% of MTS users selected the 2-min default). Users might select the default value because the response effort is lower for this selection than for any other (i.e., a click rather than a scroll and click). Users might also select based on an assumption that the default values are empirically selected. If the user does not understand the relation between the length of the interval and the amount of estimation error, he or she might not recognize the gravity of choosing based on response effort or potentially faulty assumptions. Based on these findings, Catalyst developers changed the PIR default to 2 min and left the MTS default value unchanged (i.e., both are 2 min). Since implementing that change for the PIR default value, 60% of users have used this new default interval for IR, making it now the most commonly selected PIR interval, whereas the MTS interval that remained the same is still the most commonly used MTS interval. Thus, both default selections are now empirically selected and are within the range of intervals that have proven to be highly correlated with continuous measurement procedures.

Results of Study 1 show that the observation intervals most practitioners use for collecting behavioral data using discontinuous measures do not correspond to the results and recommendations of empirical research on discontinuous data collection. That is, most patients in the current database had PIR observation intervals that exceeded 2 min, the recommended maximum interval by Hanley et al. (2007) and Guntner et al. (2003). As such, additional research is needed on the impact of the longer intervals commonly used in practice on the

accuracy of discontinuous measurement. These longer intervals are likely selected to make data collection easier for technicians, but some of these intervals may introduce significant estimation error. Prior studies have suggested that the level of behavior may impact the accuracy of discontinuous measurement (Harrop & Daniels, 1986; Saudargas & Zanolli, 1990), so an analysis of these longer intervals should take into account the level of problem behavior. We attempted to address these issues in Study 2.

Study 2

Study 2 examined the accuracy of the most commonly selected interval values (i.e., correspondence to continuous measurement procedures) from Study 1 using archival data from actual (i.e., not simulated) sessions. Sessions with continuous measurement procedures (i.e., duration) were selected from the database, and those data were reanalyzed using the most common interval values for PIR and MTS data collection, including relatively understudied intervals. The resulting values for PIR and MTS data were compared to the values obtained with the original continuous measurement systems to examine the correspondence of the measures at different interval lengths, for different session lengths, and for different levels of problem behavior.

Method

Identification of Sessions The Catalyst database included millions of data points with problem behavior, but the database did not link all data points to a specific "session." That is, in the Catalyst database, a session has a time-stamped beginning of the collection of data and a time-stamped ending of the collection of data with either (a) time-stamped instances of problem behavior during the session or (b) no instances of problem behavior recorded during the session. Therefore, we searched the Catalyst database for sessions that met the following criteria. First, the session had to include data on a relatively common topography of problem behavior for individuals with autism and other disabilities (e.g., aggression, self-injury, stereotypy, tantrums). Second, the session had to be at least 1 hr in duration to allow multiple scoring opportunities at the longer intervals. Third, the session had to contain data on problem behavior (i.e., sessions in which the observer recorded zero instances of behavior were excluded).

The search resulted in the identification of 878 sessions that met all three criteria. The sessions ranged from 1.0 to 11.0 hr in duration. Over 90% of sessions were under 4 hr in duration: 1–1.99 hr (n = 176), 2–2.99 hr (n = 394), or 3–3.99 hr (n = 236). There were 30 or fewer sessions in each of the remaining duration increments, and the fewest sessions in the range of 8 hr in duration or greater (e.g., 8–8.99 hr, n = 0; 11–11.99 hr, n = 2). The percentage duration of problem behavior (i.e., the continuous measure that was captured during services) in these sessions ranged from .0001% to 67%. Over 90% of the sessions had less than 10% duration of problem behavior, and 51% of sessions had less than 1% duration of problem behavior.

The most commonly represented topographies (62%) of problem behavior were tantrums (n = 361 sessions) and noncompliance (n = 186 sessions). Other commonly occurring topographies included protesting, crying, negative vocalizations, being off-task, and hiding. The least commonly occurring were running, self-injury, property destruction, self-stimulatory behavior, and aggression. The smaller number of sessions with topographies such as self-injury and aggression is likely due to the fact that other measures besides duration (e.g., rate, PIR) were often used for these topographies. A full description of the number of sessions at each length for each topography is available from the first author upon request.

Scoring Procedures We analyzed the data from each session to generate a percentage duration measure (i.e., continuous), as well as both PIR and MTS (i.e., discontinuous) measures for each of the following interval lengths: 10 s, 15 s, 20 s, 30 s, 60 s, 120 s, 180 s, 300 s, 600 s, 900 s, and 1,800 s. The third author created an algorithm that established the beginning of the session with the session-start time stamp (e.g., 10:01:00 a.m.) as the first second and the session-end time stamp (e.g., 12:09:53 p.m.) as the last second of the session. The algorithm scored each individual second as either containing problem behavior or not based on the time stamping of each event of problem behavior in that session, and a resulting percentage duration was generated (i.e., the continuous measure used for every comparison). The first author manually scored three short sessions in an Excel spreadsheet and compared the results to those generated by the algorithm before the algorithm was used on the primary data set. The two sets of results matched exactly for each session.

For PIR, the algorithm scored each designated interval (e.g., 30 s) as including or not including problem behavior for at least one second. For the example provided previously (e.g., session-start time stamp = 10:01:00 a.m., session-end time stamp = 12:09:53 p.m.), when scored in 10-s intervals, a time-stamped event of problem behavior beginning at 10:02:00 and ending at 10:02:37 would have resulted in the seventh, eighth, ninth, and tenth intervals of the session scored as containing problem behavior. For the same session scored with 30-s intervals, the third and fourth intervals would have been scored as containing problem behavior. We discarded incomplete intervals (i.e., 3 s remaining after the last complete scorable interval). We divided the number of intervals scored as containing problem behavior by the number of intervals in an observation session to generate a percentage of intervals with problem behavior.

For MTS, the program scored each designated time sample (e.g., 15 s) as having problem behavior occurring or not at exactly that second. For the session example provided previously (e.g., session-start time stamp = 10:01:00 a.m., session-end time stamp = 12:09:53 p.m.), scored in 10-s time samples, a time-stamped event of problem behavior beginning at 10:02:00 and ending at 10:02:37 would have resulted in the seventh, eighth, and ninth samples of the session scored as having problem behavior occurring. For the same session scored with 30-s time samples, the second and third intervals would have been scored as having problem behavior occurring. We discarded additional seconds after the last scorable sample. We divided the number of samples scored as having problem behavior occurring by the total number of samples in an observation session to generate a percentage of samples with problem behavior.

Analysis of Effects of Interval Length We compared the calculated percentage of either intervals or time samples to the percentage duration described previously for each of the 878 observation sessions at each of the designated observation intervals (e.g., 30 s, 60 s) for each type of discontinuous measure calculation (i.e., IR, MTS). We conducted comparisons using three statistics: a correlation, a mean difference score, and a mean agreement score. We calculated the correlation using the CORREL function of Excel to compare the percentage duration measure to the percentage of intervals or samples measure. Scores closer to 1.0 indicate greater correspondence between the two measures. We calculated the mean difference score as the absolute value of the difference between the percentage of intervals or samples and the percentage duration. Scores closer to 0 indicate greater correspondence between the two measures. We calculated the mean agreement score by dividing the lower obtained value of the percentage duration and percentage of intervals/samples and dividing it by the larger obtained value. This calculation takes into account the proportion of the total amount of behavior represented in the measurement error. Scores closer to 1.0 indicate greater correspondence between the two measures.

Analysis of the Effects of Session Duration and Problem Behavior ***Session duration.*** We examined the potential effects of session duration on the accuracy of discontinuous measurement by arranging sessions from longest to shortest duration and then portioning them into quartiles. That is, the top quartile (Quartile 1) included the longest sessions, whereas the bottom quartile (Quartile 4) included the shortest sessions. For each quartile, we computed a correlation for the percentage duration of problem behavior and the percentage of intervals with problem behavior (PIR scoring) and for the percentage duration of problem behavior and the percentage of time samples with problem behavior (MTS scoring).

Amount of problem behavior. We also examined the potential effects of the amount of problem behavior on the accuracy of the discontinuous measurement by arranging sessions from the highest percentage duration of problem behavior (top quartile, Quartile 1) to the lowest (bottom quartile, Quartile 4). For each quartile, we computed a correlation for the percentage duration of problem behavior and the percentage of intervals with problem behavior (PIR scoring) and for the percentage duration of problem behavior and the percentage of time samples with problem behavior (MTS scoring).

Results and Discussion

Figure 2 (top panel) depicts the average correlation between the percentage duration (i.e., continuous measurement) score from the Catalyst database with each obtained discontinuous

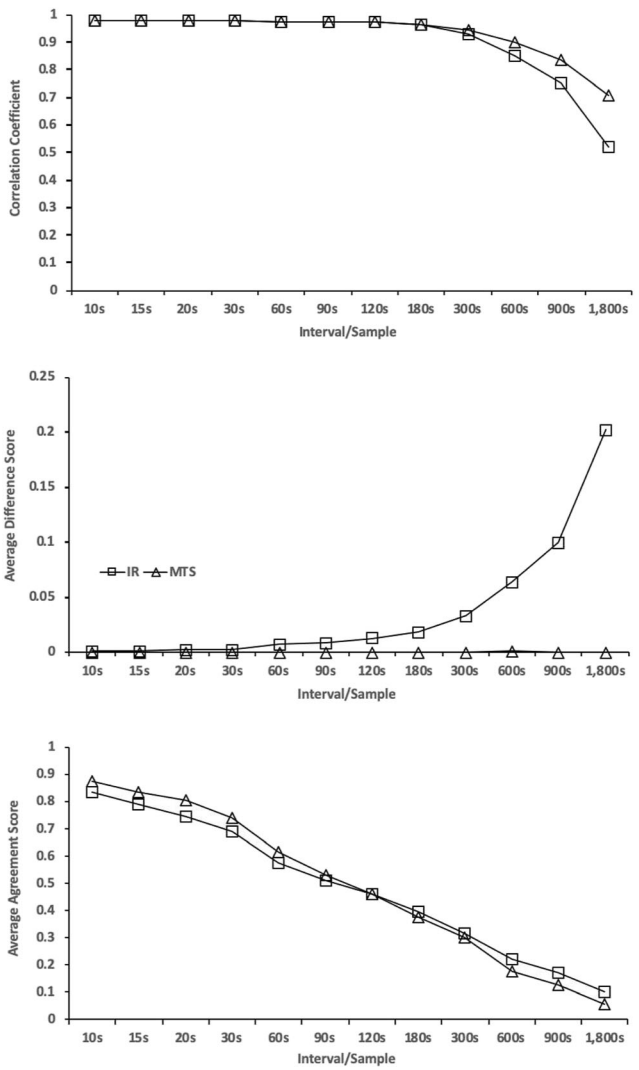

Fig. 2 Correspondence between discontinuous measurement and continuous measurement calculated as a correlation coefficient (top panel), average difference score (middle panel), and average agreement score (bottom panel)

measurement score generated by parsing the sessions into various intervals and time samples. The correlations for both PIR and MTS remain above .96 for values up to 180 s. At 300 s, the correlations for both measurement systems begin to decrease with a more rapid drop for the PIR measurement system. The lowest correlation is for PIR scoring at 1,800 s (r = .52). The average difference score (i.e., |percentage duration − percentage intervals or samples|) is depicted in Figure 2 (middle panel) for both PIR and MTS, for each value. The average difference score remains negligible at all intervals for MTS scoring, likely because the type of error produced is nonsystematic, so the overestimates and underestimates tend to cancel each other out over a very large number of sessions. However, the difference scores systematically increase for PIR with increasing values of the PIR interval. Although the continuous measure always remains the same, larger PIR intervals result in a higher percentage-of-intervals measure because behavior scored in two smaller intervals would be more likely to fall within a single larger interval (e.g., responses occurring at Second 8 and Second 18 would be scored in separate intervals when using 10-s PIR but would be scored in the same interval when using 20-s IR). This smaller number of intervals with problem behavior (i.e., the numerator) is also being divided by a smaller total number of intervals (i.e., the denominator) resulting in an increasing percentage.

The bottom panel of Figure 2 depicts the average agreement score (i.e., smaller value/larger value), which takes into account the amount of problem behavior occurring in each session by illustrating the proportion of the total amount of problem behavior represented in the difference score. That is, a difference score of 1% represents 10% of the value of a total percentage duration of 10% (i.e., 9/10 = .90) but only 5% of the value of the total percentage duration of 20% (i.e., 19/20 = .95). Scores closer to 1 represent better agreement, which is evident for the smallest units for each type of measurement (i.e., PIR = .83; MTS = .88), with a similarly steady decline in agreement for both types of measurement as the scoring interval or sample increases.

The specific values for all three metrics at each IR or MTS value are available from the first author upon request. All three of the measures indicate that intervals greater than 2 to 3 min show progressively poorer correspondence between continuous and discontinuous measures. This supports the findings of other studies (Guntner et al., 2003; Hanley et al., 2007) that 2 min may be the largest interval or sample that produces data that correspond reasonably with continuous measures. Certainly, very long intervals, such as 30–60 min, are contraindicated. In addition, the MTS measurement system generally produced closer correspondence to continuous measurement than PIR across a large number of sessions, which supports the findings of prior studies (Alvero et al., 2008; Gardenier et al., 2004). The agreement measure, which accounts for the amount of problem behavior in the session, showed the most linear decline in agreement as the interval or sample size increases with similar effects for both metrics.

The analysis of the effects of session duration and amount of problem behavior on agreement between the two measures

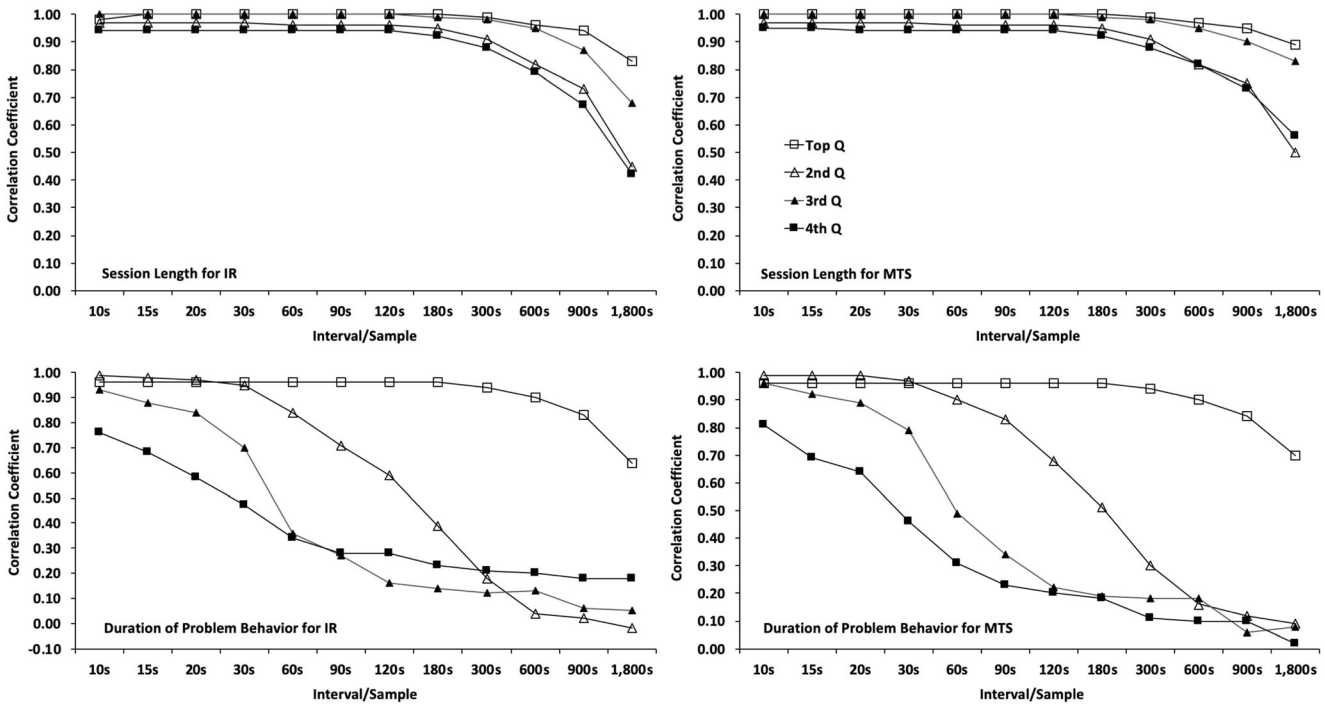

Fig. 3 Correlation coefficients with continuous data are presented by quartile with the top quartile representing the sessions with the longest durations (top panels) and highest percentage duration of problem behavior (bottom panels). Data for PIR analysis are in the left column, and data for MTS are in the right column

is presented in Figure 3. The duration of the session (top-left and top-right panels) did not systematically impact the correlations between the two measures, with Quartiles 1 and 3 having higher correlations, and Quartiles 2 and 4 having lower ones even though the session duration for Quartile 3 was lower than for Quartile 2. The durations of all sessions included in this analysis (i.e., 1–11 hr) are longer than the majority of most prior studies examining discontinuous measurement or using it as a dependent variable (e.g., Meany-Daboul, Roscoe, Bourret, & Ahearn, 2007; Mudford et al., 1990; Sanson-Fisher, Poole, & Dunn, 1980; Saudargas & Zanolli, 1990).

The amount of problem behavior in the session (bottom-left and bottom-right panels) systematically impacted the correlations between the two measures, with Quartile 1 having the highest correlations, followed by Quartile 2, then Quartile 3, then Quartile 4. That is, a greater amount of problem behavior captured in the session resulted in a higher correlation between the two measures, whereas a lower amount of problem behavior (Quartiles 2–4) was associated with lower correlations and a much steeper decrease in the value of the correlation as the interval size increased. These differences generally increased as the size of the interval or sample increased. For example, differences between Quartile 1 and Quartile 2 were minimal until the interval size reached 60 s (lower-left panel), but the differences between Quartile 1 and Quartile 2 increased progressively as the interval size increased above 60 s. The correlations for the third and fourth quartiles (i.e., sessions with .0089% duration and less) are below .9, with intervals as small as 15 s and 10 s for PIR scoring and 20 s and 10 s for MTS scoring. Thus, with very low levels of problem behavior, even very small intervals produce unsatisfactory correspondence between continuous and discontinuous measurement.

General Discussion

These studies examined discontinuous measurement procedures using a large database of data collected across multiple human-service settings. As in published studies (Beavers et al., 2013), Catalyst users implemented continuous measurement procedures more often than discontinuous measurement. The current users often implemented larger intervals or time samples than those used or recommended in published studies (Alvero et al., 2008; Powell, 1984). In addition, Catalyst users implemented PIR more often than MTS procedures, even though MTS procedures do not require constant observation and prior studies have found that MTS produces better correspondence with continuous measurement than IR. These findings suggest that practitioners who use discontinuous measurement procedures may select and design measures in ways that are not supported by the empirical literature (e.g., the interval is too long). In addition, our results suggest that Catalyst users often selected the default value for the length of the PIR and MTS intervals, which suggests that using findings from empirical research to select default values for technological aids like Catalyst may improve the accuracy of discontinuous measurement procedures for users.

The subsequent analysis of the correspondence between continuous and discontinuous measurement procedures supports prior findings in several ways. First, smaller intervals and samples consistently produced higher correspondence with continuous data than longer intervals and samples. Second, intervals and samples up to 2 min produced excellent correspondence (i.e., correlation, difference score) for both PIR and MTS, with a progressively steeper decrement in correspondence at values of 5 min and longer. Third, the amount of problem behavior occurring during the observation systematically impacted the correlation between the two measures, with very low levels of problem behavior resulting in lower correlations at every interval and sample, including 10 s. Conversely, higher levels of problem behavior produced correlations between discontinuous and continuous data above .80 until the intervals or samples exceeded 900 s. The agreement metric, which takes into account the level of problem behavior, provides a slightly different view of the correspondence between the measurement systems and suggests that there is a linear and similar decline observed for both PIR and MTS procedures. Unlike some other studies, the duration of the observation did not systematically impact agreement between the two measures. However, this may be due to the fact that the observations included in this analysis tended to be quite long relative to prior studies.

Several cautions and limitations regarding the interpretation of these data are worthy of note. First, these data are reflective of the users of this electronic data-collection software. Although this is a large sample of users, the data may not accurately reflect discontinuous data-collection procedures of all practicing behavior analysts. Second, we conducted comparisons between continuous and discontinuous measurement by reanalyzing continuous data collected and stored within the database rather than by simultaneously scoring the same observation using the two different systems. Thus, any change in error (e.g., less error because the data are easier to collect) associated with collecting data using a PIR or MTS procedure are not captured in this analysis. As such, it is likely that these data represent the upper limit of agreement between continuous and discontinuous data collection.

In summary, these studies examined a large data set to describe the data-collection practices of behavior analysts and whether those practices are generally in alignment with the published literature. Future studies could examine practices related to other aspects of programming (e.g., use of schedule thinning, use of prompt sequences, number of active programs per hour of programming, selection of measures for different topographies of behavior) or decision-making (e.g., criteria for changing programming, mastery criteria) to

determine whether common practices comport with research findings. Large data sets offer the opportunity to conduct analyses that typically have not been possible for behavior analysts. These analyses may help us better understand how everyday practice decisions impact the quality of behavior-analytic programming.

Author Note The authors thank Jamila Pitts for assistance with coding data from research articles.

Compliance with Ethical Standards

Conflict of Interest The authors of this manuscript are employed by or serve on the advisory board of DataFinch Technologies (i.e., Catalyst).

Informed Consent All agencies whose data were included in the analysis provided consent.

References

Alvero, A. M., Struss, K., & Rappaport, E. (2008). Measuring safety performance: A comparison of whole, partial, and momentary time-sampling recording methods. *Journal of Organizational Behavior Management, 27*, 1–28. https://doi.org/10.1300/J075v27n04_01.

Baer, D. M., Wolf, M. M., & Risley, T. (1968). Current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1*, 91–97. https://doi.org/10.1901/jaba.1968.1-91.

Beavers, G. A., Iwata, B. A., & Lerman, D. C. (2013). Thirty years of research on the functional analysis of problem behavior. *Journal of Applied Behavior Analysis, 46*, 1–21. https://doi.org/10.1002/jaba.30.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River, NJ: Pearson.

Devine, S. L., Rapp, J. T., Testa, J. R., Henrickson, M. L., & Schnerch, G. (2011). Detecting changes in simulated events using partial-interval recording and momentary time sampling III: Evaluating sensitivity as a function of session length. *Behavioral Interventions, 26*, 103–124. https://doi.org/10.1002/bin.328.

Fiske, K., & Delmolino, L. (2012). Use of discontinuous methods of data collection in behavioral intervention: Guidelines for practitioners. *Behavior Analysis in Practice, 5*(2), 77–81. https://doi.org/10.1007/BF03391826.

Gardenier, N. C., MacDonald, R., & Green, G. (2004). Comparison of direct observational methods for measuring stereotypic behavior in children with autism spectrum disorders. *Research in Developmental Disabilities, 25*, 99–118. https://doi.org/10.1016/j.ridd.2003.05.004.

Guntner, P. L., Venn, M. L., Patrick, J., Miller, K. A., & Kelly, L. (2003). Efficacy of using momentary time samples to determine on-task behavior of students with emotional/behavioral disorders. *Education and Treatment of Children, 26*, 400–412.

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. https://doi.org/10.1901/jaba.2007.603-618.

Harrop, A., & Daniels, M. (1986). Methods of time sampling: A reappraisal of momentary time sampling and partial interval recording. *Journal of Applied Behavior Analysis, 19*, 73–77. https://doi.org/10.1901/jaba.1993.26-277.

Jessel, J., Metras, R., Hanley, G. P., Jessel, C., & Ingvarsson, E. T. (in press). Evaluating the boundaries of analytic efficiency and control: A consecutive controlled case series of 26 functional analyses. *Journal of Applied Behavior Analysis*. https://doi.org/10.1002/jaba.544.

Johnston, J. M., & Pennypacker, H. S. (2009). Observing and recording. In *Strategies and tactics of behavioral research* (3rd ed., pp. 115–138). New York, NY: Routledge.

Kolt, L. D., & Rapp, J. T. (2014). Assessment of therapists' preferences for discontinuous measurement systems. *Behavioral Interventions, 29*, 304–314. https://doi.org/10.1002/bin.1392.

LeBlanc, L. A., Raetz, P. B., Sellers, T. P., & Carr, J. E. (2016). A proposed model for selecting measurement procedures for the assessment and treatment of problem behavior. *Behavior Analysis in Practice, 9*, 77–83. https://doi.org/10.1007/s40617-015-0063-2.

Machado, M. A., Luczynski, K. C., & Hood, S. A. (2019). Evaluation of the accuracy, reliability, efficiency and acceptability of fast forwarding to score problem behavior. *Journal of Applied Behavior Analysis, 50*, 315–334. https://doi.org/10.1002/jaba.510.

Meany-Daboul, M. G., Roscoe, E. M., Bourret, J. C., & Ahearn, W. H. (2007). A comparison of momentary time sample and partial-interval recording for evaluating functional relations. *Journal of Applied Behavior Analysis, 40*, 501–514. https://doi.org/10.1901/jaba.2007.40-501.

Morris, C., et al. (in press). Using Microsoft Excel® to build a customized partial-interval data collection system. *Behavior Analysis in Practice*.

Mudford, O. C., Beale, I. L., & Singh, N. N. (1990). The representativeness of observational samples of different durations. *Journal of Applied Behavior Analysis, 23*, 323–331. https://doi.org/10.1901/jaba.1990.23-323.

Mudford, O. C., Taylor, S. A., & Martin, N. T. (2009). Continuous recording and interobserver agreement algorithms reported in the. *Journal of Applied Behavior Analysis, 42*, 165–169. https://doi.org/10.1901/jaba.2009.42-165.

Powell, J. (1984). On the misrepresentation of behavioral realities by a widely practiced direct observation procedure: Partial interval (one-zero) sampling. *Behavioral Assessment, 6*, 209–219.

Powell, J., Martindale, A., & Kulp, S. (1975). An evaluation of time-sample measures of behavior. *Journal of Applied Behavior Analysis, 8*, 463–469. https://doi.org/10.1901/jaba.1975.8-463.

Powell, J., Martindale, B., Kulp, S., Martindale, A., & Bauman, R. (1977). Taking a closer look: Time sampling and measurement error. *Journal of Applied Behavior Analysis, 10*, 325–332. https://doi.org/10.1901/jaba.1975.8-463.

Rapp, J. T., Colby, A. M., Vollmer, T., Roan, H. S., Lomas, J., & Britton, L. N. (2007). Interval recording for duration events: A re-evaluation. *Behavioral Interventions, 22*, 319–345. https://doi.org/10.1002/bin.239.

Rapp, J. T., Colby-Dirksen, A. M., Michalski, D. N., Carroll, R. A., & Lindenberg, D. N. (2008). Detecting changes in simulated events using partial-interval recording and momentary time sampling. *Behavioral Interventions, 23*, 237–269. https://doi.org/10.1002/bin.269.

Saini, V., Fisher, W. W., & Retzlaff, B. J. (2018). Predictive validity and efficiency of ongoing visual-inspection criteria for interpreting functional analyses. *Journal of Applied Behavior Analysis, 51*, 303–320. https://doi.org/10.1002/jaba.450.

Sanson-Fisher, R. W., Poole, A. D., & Dunn, J. (1980). An empirical method for determining an appropriate interval length for recording behavior. *Journal of Applied Behavior Analysis, 13*, 493–500. https://doi.org/10.1901/jaba.1980.13-493.

Saudargas, R. A., & Zanolli, K. (1990). Momentary time sampling as an estimate of percentage time: A field validation. *Journal of Applied Behavior Analysis, 4*, 533–537. https://doi.org/10.1901/jaba.1990.23-533.

Schmidt, M. G., Rapp, J. T., Novotny, M., & Lood, E. (2013). Detecting changes in non-simulated events using partial-interval recording and momentary time sampling: Evaluating false positives, false negatives, and trending. *Behavioral Interventions, 28*, 58–81. https://doi.org/10.1002/bin.1354.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology.* New York, NY: Basic Books.

Wirth, O., Slaven, J., & Taylor, M. A. (2013). Interval sampling methods and measurement error: A computer simulation. *Journal of Applied Behavior Analysis, 47*, 83–100. https://doi.org/10.1002/jaba.93.