A Comparison of Methods for Collecting Data on Performance During Discrete Trial Teaching

Dorothea C. Lerman ^a, Laura Harper Dittlinger ^b, Genevieve Fentress ^b & Taira Lanagan ^c

^a University of Houston, Clear Lake

^b Texana Behavior Treatment and Training Center

^c Center for Autism and Related Disorders, Tarzana, CA

ABSTRACT

Therapists of children with autism use a variety of methods for collecting data during discrete-trial teaching. Methods that provide greater precision (e.g., recording the prompt level needed on each instructional trial) are less practical than methods with less precision (e.g., recording the presence or absence of a correct response on the first trial only). However, few studies have compared these methods to determine if less labor-intensive systems would be adequate to make accurate decisions about child progress. In this study, precise data collected by therapists who taught skills to 11 children with autism were re-analyzed several different ways. For most of the children and targeted skills, data collected on just the first trial of each instructional session provided a rough estimate of performance across all instructional trials of the session. However, the first-trial data frequently led to premature indications of skill mastery and were relatively insensitive to initial changes in performance. The sensitivity of these data was improved when the therapist also recorded the prompt level needed to evoke a correct response. Data collected on a larger subset of trials during an instruction session corresponded fairly well with data collected on every trial and revealed similar changes in performance.

Keywords: continuous recording, data collection, discontinuous recording, discrete-trial teaching

Measurement of behavior and progress monitoring are essential practices in applied behavior analysis. Practitioners routinely inspect data collected during behavioral programs to evaluate the effectiveness of interventions, guide decisions about program modifications, and determine when goals are mastered. As such, it is important that practitioners have good observation systems in place. A good observation system is objective, meaning that it focuses on observable, quantifiable behavior; reliable, meaning that it produces consistent results across time and users; accurate, meaning that it provides a true representation of the behavior of interest; and sensitive, meaning that it reveals changes in behavior. The latter two characteristics are the concern of this article.

Teachers, parents, and other educators of individuals with autism are increasingly using practices drawn from the behavior-analytic literature to address skill deficits. One commonly used approach is called “discrete-trial teaching.” As part of discrete-trial teaching, the instructor presents multiple learning opportunities of a single skill during instructional sessions, provides immediate consequences for correct and incorrect responses, and records the outcome of trials to monitor progress. Results of a recent survey suggest that practitioners are using a variety of methods to collect data on the learner’s performance during discrete-trial teaching (Love, Carr, Almason, & Petursdottir, 2009). These methods vary in terms of frequency and specificity.

For example, the instructor may record the learner’s performance on every learning trial during teaching sessions, an approach called “continuous recording.” Alternatively, the instructor may record the outcome of just a sample of instructional trials (e.g., the first trial or first few trials of each teaching session), an approach called “discontinuous recording.” Relative to discontinuous recording, continuous recording may provide a more sensitive measure of changes in performance, minimize the impact of correct guesses on overall session data, and lead to more stringent mastery criteria (Cummings & Carr, 2009). On the other hand, discontinuous recording is more efficient and may be easier for practitioners to use than continuous recording. When data are collected on

just the first instructional trial, discontinuous recording also reveals the level of performance in the absence of immediately preceding learning trials or prompts.

The instructor also may vary the specificity of the data collected during continuous and discontinuous recording. For example, the instructor may simply record whether an independent (unprompted) versus a prompted response occurred on an instructional trial (i.e., nonspecific recording). Alternatively, if the learner required a prompt, the instructor may also record the prompt level that occasioned a correct response (i.e., specific recording). Like continuous recording, specific recording may increase the sensitivity of the measurement system to changes in the learner's performance. Instructors, however, may find nonspecific recording easier to use because less information is documented.

Effortful measurement systems could compromise data reliability and treatment integrity, as well as lengthen the duration of instructional sessions. This issue is even more important in light of the increasing numbers of teachers, therapists, parents, and other educators who are learning how to use discrete-trial teaching. As such, further research is warranted on the accuracy and utility of efficient measurement systems. Collecting data intermittently (discontinuous recording) and limiting the specificity of the data collected (nonspecific recording) may be viable alternatives to more effortful systems. However, only two studies thus far have addressed this question within the context of discrete-trial teaching, and the outcomes were inconsistent (Cummings & Carr, 2009; Najdowski et al., 2009).

Cummings and Carr (2009) compared continuous (trial-by-trial) versus discontinuous (first-trial-only) recording on the acquisition and maintenance of skills. Participants were six children with autism spectrum disorders between the ages of 4 and 8 years. The therapist taught a total of 100 skills, half of which were randomly assigned to continuous measurement and half of which were randomly assigned to discontinuous measurement. The therapist taught each skill during 10-trial sessions and recorded whether the participant exhibited a correct or incorrect response on each trial (for continuous recording) or on the first trial only (for discontinuous recording). Training on a particular skill continued until the participant exhibited correct responding on 100% of trials (either all 10 trials or the first trial only) for two consecutive sessions. Follow-up (i.e., maintenance) was examined by evaluating performance on each skill 3 weeks after mastery. Results showed that about 50% of the skills met the acquisition criterion more quickly when the therapist used discontinuous measurement rather than continuous measurement. Discontinuous recording also resulted in a mean of 108 fewer trials per skill, or 57 fewer minutes in training, per participant. However, when the experimenters examined the follow-up data for the skills that the children acquired more quickly under discontinuous recording, they found that the children showed poorer maintenance on 45% of these skills when compared to skills measured using continuous recording. The authors concluded that continuous recording led to more conservative decisions about skill mastery and, hence, better

long-term maintenance. After noting that continuous recording also permits within-session analysis of progress, the authors recommended that instructors use continuous recording unless brief session duration is a priority.

Najdowski et al. (2009) replicated and extended Cummings and Carr (2009). Participants were 11 children with autism. Therapists taught four skills to each child, with two skills randomly assigned to continuous measurement and two skills randomly assigned to discontinuous measurement. Procedures were similar to those used by Cummings and Carr. However, teaching continued until the participant exhibited correct responding on 80% of trials (for continuous recording) or exhibited a correct response on the first trial (for discontinuous recording) across three consecutive sessions. Like Cummings and Carr, the experimenters compared the number of sessions required to meet the acquisition criterion for the two groups of skills. Results were inconsistent with Cummings and Carr because participants required the same number of sessions to meet the mastery criterion, regardless of the recording method. The experimenters also conducted the same analysis for the skills associated with continuous recording by comparing the first-trial data from these records to the all-trial data. Contrary to the conclusions of Cummings and Carr, results of this within-skills analysis showed that the participants met the mastery criterion more quickly under *continuous* recording. This suggested that discontinuous recording would lead to more conservative decisions about skill mastery than continuous recording. Comparisons across the two studies are difficult, however, because the experimenters used different mastery criteria and methods for programming and assessing skill maintenance.

The purpose of the current study was to replicate and extend this previous work by examining a broader range of recording methods and by examining multiple mastery criteria. In both Cummings and Carr (2009) and Najdowski et al. (2009), different targets were randomly assigned to the different measurement conditions so that the experimenters could evaluate the impact of therapist behaviors on the acquisition and maintenance of skills. For example, acquisition may be delayed if more effortful measurement systems compromise the integrity and efficiency of teaching (although these two therapist-related variables were not directly measured in the previous studies). In addition, as demonstrated by Cummings and Carr, skills may not maintain if a particular recording method leads to premature decisions about skill mastery.

However, this type of between-target comparison also may be limited in several respects. First, the targets assigned to one particular recording method could be more difficult for the participants than those assigned to other conditions, creating variations in the participants' performance. Second, the sensitivity of the measurement systems per se cannot be evaluated separately from possible differences in the child's performance as a result of differences in the therapist's behavior (e.g., integrity or efficiency of the teaching). Third, a between-targets analysis limits the number of different comparisons that would be practical to conduct.

An alternative to a between-target analysis is a within-target analysis, which was also conducted by Najdowski et al. (2009). For example, therapists could collect data on all trials (continuous recording) and include the prompt level needed (specific recording). These data records then could be re-analyzed to determine how the outcomes would have appeared if data had been collected on a subset of these trials or with less specificity. Prior experimenters have adopted this approach to evaluate the sensitivity of different data collection (e.g., Hanley, Cammilleri, Tiger, & Ingvarsson, 2007; Meany-Daboul, Roscoe, Bourret, & Ahearn, 2007).

Thus, we re-analyzed data collected via continuous, specific recording to address the following questions: (a) How closely does a subset of trials predict performance across all trials? (b) What is the sensitivity of continuous versus discontinuous measurement to changes in the learner's performance during initial acquisition? and (c) What is the sensitivity of specific recording (i.e., documenting the prompt level) versus nonspecific recording to changes in the learner's performance during initial acquisition?

Method

Participants and Settings

Eleven children and their therapists participated. The children, ages 4 years to 15 years, were receiving one-on-one teaching sessions with trained therapists at three-day programs that specialized in behavior-analytic therapy. The children were diagnosed with moderate to severe developmental disabilities or autism. They had been receiving therapy at the day programs from 4 weeks to 6 years at the time of the study. A total of 10 therapists participated. The therapists had from 1 year to 6 years of relevant experience at the time of the study. Seven of the therapists had bachelor's degrees and two had master's degrees and were Board Certified Behavior Analysts.

Data Collection

A total of 24 targeted skills were included in the analysis (i.e., 2 or 3 targets for each participant). The following criteria were used to select targets to include in the analysis: (a) The participant exhibited no correct responses across a minimum of three baseline probe trials, (b) the target involved a nonvocal response, and (c) assessments conducted by staff at the day program indicated that the participant had the appropriate prerequisite skills. Board Certified Behavior Analysts and therapists at each day program selected the targets as part of their routine supervision of the children's behavior programming. The acquisition targets included gross motor imitation, receptive identification of objects, and instruction following.

The therapists used a data sheet to record the occurrence of a correct unprompted response or the prompt level needed to obtain a correct prompted response on each trial for each targeted skill. The therapists scored the outcome of each trial at the conclusion of the trial and before the start of the next trial (i.e., during the inter-trial interval). A correct unprompted

response was scored if the participant exhibited the targeted response within 3 s of the initial instruction, delivered in the absence of a prompt. A prompt level was recorded if the participant exhibited the targeted response within 3 s of a verbal, gesture, model, or physical prompt.

A second observer collected data independently during at least 25% of the sessions for each target. The data records were compared on a trial-by-trial basis for reliability purposes. An agreement was scored if both observers recorded the occurrence of a correct unprompted response or if both observers recorded the same prompt level needed to obtain a correct prompted response. Interobserver agreement was calculated by dividing the total number of agreements by the total number of agreements plus disagreements and multiplying by 100 to obtain the percentage of agreement. Mean agreement across targets was 99.5%, with a range of 87.5% to 100% across sessions.

Procedure

Each session consisted of 8 or 9 teaching trials, interspersed with trials that assessed previously mastered targets. It should be noted the number of trials in each session varied across participants, not across sessions. Data on previously mastered targets were not included in the analysis. The therapists taught more than one acquisition target during an instructional session if this format was consistent with the participant's regular therapy sessions. In these cases, the participant received 8 or 9 teaching trials of *each* acquisition target. Otherwise, the therapists taught just one acquisition target during each session. Eight of the participants received one training session with the targets per day, 1 to 3 days per week. Three of the participants received two training sessions per day, with a minimum of 2 hrs between each session, 4 days per week.

Baseline probe trials. The purpose of the baseline probe trials was to identify targets with zero levels of correct responding prior to training. The therapist presented the initial instruction at the start of each trial, waited 3 s for a response, and delivered no prompts or consequences for correct or incorrect responses. The therapist conducted at least three probe trials for each target, with no more than one probe trial conducted in a single day.

Teaching. The therapist presented the initial instruction at the start of each trial and waited 3 s for a response. If the participant responded incorrectly or did not respond, the therapist delivered a prompt. The therapist faded the prompts by starting with the most intrusive prompt possible in the first session and gradually transitioning to less intrusive prompts across trials and sessions. Specifically, during the first teaching session, the therapist delivered a full physical prompt on the first trial that a prompt was needed for a particular acquisition target. On the next trial with that target, the therapist delivered a less intrusive prompt (e.g., model) if a correct response did not occur within 3 s of the initial instruction. In this manner, prompts were faded after each trial with a correct response. If a child did not respond correctly to a prompt, the therapist delivered increasingly more intrusive prompts within the trial

Figure 1. Mean number of sessions to meet the two-session or three-session mastery criterion for the 24 targets when examining data from the first trial, first three trials, or all trials. The narrow bars show the standard deviation for each.

until a correct response occurred; in addition, the initial prompt delivered on the next trial with that target (if needed) was more intrusive than the initial prompt delivered on the previous trial with that target. The therapist delivered reinforcement contingent on correct responses. The type (e.g., food items, leisure materials, tokens) and schedule of reinforcement (e.g., continuous, fixed ratio 2) were identical to those used during the child's regular behavior programming. Training terminated following a minimum of three consecutive sessions with (a) correct unprompted responses at or above 88% of the trials (continuous data) and (b) a correct unprompted response on the first trial.

Data Analyses and Results

Question #1: How Closely Does a Subset of Trials Predict Performance Across All Trials?

If data collected on the first trial or first three trials predicts performance across all trials, then practitioners, teachers, and parents could confidently use a less effortful measurement system. This might increase the practicality and accuracy of data collection, particularly in settings with competing demands on the instructor's time (e.g., busy classrooms) or for instructors with limited experience collecting data. We conducted three different analyses to examine the correspondence between data collected on the first trial (discontinuous data), the first three trials (discontinuous data), and all trials (continuous data).

Prior to conducting these analyses, we calculated the percentage of all trials with correct unprompted responses in each session for each acquisition target by dividing the total number of trials with a correct unprompted response by the total number of trials and multiplying by 100. This provided our continuous data. To generate our discontinuous data (i.e., first-trial and three-trial data) from these records, we calculated the percentage of the first three trials with correct unprompted responses in each session for each acquisition target by dividing the total number of the first three trials with a correct unprompted response by 3 and multiplying by 100%. We also examined whether or not a correct unprompted response occurred on the first trial of each session. The outcome of the first trial was either 0% (prompt needed) or 100% (occurrence of a correct unprompted response).

For the first analysis, we replicated the comparisons conducted by Cummings and Carr (2009) and Najdowski et al. (2009) by determining the number of sessions required to meet a mastery criterion based on the continuous versus discontinuous data. One criterion was based on performance across two consecutive sessions (Cummings & Carr, 2009), and the other criterion was based on performance across three consecutive sessions (Najdowski et al., 2009). The specific criterion selected provided a middle ground between the 100% of correct trials required by Cummings and Carr and the 80% of correct trials required by Najdowski et al. Our mastery criterion was responding at or above 88% of all trials, with a correct response

Figure 2. Probability that responding was correct on more than 50% of all trials when a correct response occurred on the first trial (left bar) and when an incorrect response occurred on the first trial (right bar) for the 24 targets. The narrow bars show the standard deviation for each.

required on the first trial of the session. Essentially, this criterion permitted the child to exhibit one incorrect response in each session, as long as the prompted response did not occur on the first trial of the session. We felt that this initial acquisition criterion was similar to that commonly used in practice.

Figure 1 shows the mean number of sessions (along with the standard deviation) to reach the mastery criterion when performance was based on two consecutive sessions (bars on the left) and when performance was based on three consecutive sessions (bars on the right). The majority of targets (83%) met the two-session mastery criterion in fewer sessions when performance was examined for the first trial only ($M = 13.1$ sessions; range, 4 to 25) compared to performance on the first three trials ($M = 16.8$ sessions; range, 6 to 41) or on all trials ($M = 17.1$ sessions; range, 6 to 41). This difference occurred for less than half of the targets (46%) when using the three-session mastery criterion ($M = 16.9$; range, 5 to 25 for the first trial; $M = 19.1$; range 7 to 42 for the first 3 trials; and $M = 19$; range, 7 to 42 for all trials). We found no differences when comparing the three-trial and continuous data. These findings indicate that the first-trial data would have produced premature determinations of skill mastery for most of the targets unless we based the criterion on performance during three or more consecutive sessions.

For the second analysis, we calculated the probability that the mean level of correct responding across all trials was above 50% given (a) a correct response on the first trial, and (b) an

incorrect response on the first trial. We assumed that performance above the 50% level would reflect learning, whereas performance at or below this level could indicate learning but might also result from correct guessing. It should be noted, however, that the 50% accuracy level was not considered an indicator of chance responding, which only would be the case for targets with two response options. Instead, it was selected because it represents the midpoint of our percentage scale. We conducted this particular analysis to determine the extent to which performance on the first trial was a reliable indicator of learning or lack thereof. This analysis also might indicate whether correct guesses might inflate first-trial performance and whether instruction and prompting early in the session might inflate all-trial performance. The former would be suggested if a correct response on the first trial was *not* a strong predictor of above-50% responding on all trials. The latter would be suggested if an incorrect response on the first trial *was* a strong predictor of above-50% responding on all trials. Finally, for the last analysis, we compared the mean levels of responding across all sessions for each target when data were based on performance during the first trial, first three trials, and all trials.

Figure 2 shows the probability that mean correct responding on all trials exceeded 50% given an unprompted (i.e., correct) response on the first trial or a prompted (i.e., incorrect) response on the first trial. The probability that correct responding exceeded 50% of trials when a correct response occurred on

Figure 3. Percentage of trials with correct responding across all sessions when examining first-trial, three-trial, and all-trial data for each target.

the first trial was 0.92 (range, .66 to 1.0). It should be noted that all but two targets exceeded a probability of .80 and 17 of the 24 targets showed a perfect correlation (probability of 1.0) between a correct response on the first trial and above-50% responding on all trials. The probability that correct responding exceeded 50% of trials when an incorrect response occurred on the first trial was just 0.26 (range, 0 to .71), with 16 of the 24 targets falling below the mean. This indicates that performance on the first trial was a fairly good predictor of whether performance across all trials exceeded or fell below 50% correct. This also suggests that correct guesses did not substantially influence the first-trial data because all-trial data were typically above 50% correct when the participant responded correctly on the first trial. Conversely, prior instruction and prompting did not substantially influence the all-trial data because these data were typically below 50% when the participant responded incorrectly on the first trial. Nonetheless, some exceptions did occur, as indicated by the ranges and individual results.

Figure 3 shows the mean percentage of correct responses for each target when data were calculated using the first trial, the first three trials, and all trials of the instructional sessions. For visual inspection purposes, the data points are ordered based on level of responding across all trials. For all but four targets, the mean percentages of correct responses for the continuous data were the same as or greater than those for the discontinuous data. On average, the mean level of performance across sessions was 40.5% correct for the first-trial data, 44.1% correct for the three-trial data, and 51% correct for the all-trial data.

These findings indicate that the discontinuous data tended to underestimate all-trial performance.

Question #2: What Is the Sensitivity of Continuous Versus Discontinuous Measurement to Changes in the Learner's Performance During Initial Acquisition?

The analyses conducted to address Question #1 did not provide information about the potential *sensitivity* of the continuous versus discontinuous data to *changes* in the learner's performance during the initial stages of acquisition. That is, it is also important to determine how readily each recording method might indicate learning or lack thereof during the instruction of a new skill. When ongoing data suggest little or no improvement in performance, instructors can modify their teaching methods to promote more rapid learning. Data indicative of learning then inform instructors regarding the adequacy of their teaching. The benefits of using less effortful measurement systems may not outweigh the costs if insufficient sampling of behavior (via discontinuous recording) reduces the sensitivity of measurement to changes in performance.

To examine differences in the sensitivity of continuous versus discontinuous data during initial acquisition, we again examined our session-by-session data on the percentage of correct, unprompted responses generated to address Question #1. For each target, we compared the number of sessions the therapist conducted prior to the occurrence of the first nonzero point when data were calculated from the first trial, the first three trials, and all trials. Recall that, for this study, we only

Figure 4. Mean number of sessions before the first nonzero point when examining the first-trial, three-trial, and all-trial data for the 24 targets. The narrow bars show the standard deviation for each.

included targets associated with no correct responses during the baseline. This permitted us to use a relatively straightforward and objective criterion for “improvement in performance.”

Figure 4 shows the mean number of sessions (along with the standard deviation) that occurred before the data revealed any change in performance (i.e., the first nonzero point) during initial acquisition when examining data from the first trial, the first three trials, and all trials. The mean number of sessions was 8.6 (range, 2 to 21 sessions) for the first-trial data, 6.8 (range, 2 to 17 sessions) for the three-trial data, and 4.7 (range, 1 to 14 sessions) for the continuous data. This specific pattern (i.e., a negative relation between the frequency of data collection and the mean number of sessions across the three data methods) was observed for 58% of targets. Moreover, the first-trial data required a greater number of sessions to reveal a change in performance compared to the continuous data for nearly all of the targets (96%), and this difference was somewhat substantial. The instructors had conducted more than twice the number of sessions, on average, before the first-trial data showed any change in responding than before the all-trial data showed any change in responding. These findings indicate that the continuous data were more sensitive to change than the discontinuous data, particularly the first-trial data.

Question #3: What Is the Sensitivity of Specific Versus Nonspecific Recording to Changes in the Learner’s Performance During Initial Acquisition?

For our final question, we wondered if recording the prompt level needed by the learner for trials with incorrect (unprompted) responses would improve the sensitivity of either the continuous, first-trial, or three-trial data. As noted previously,

data collection methods that readily identify improvements in performance (or lack thereof) provide instructors with much-needed feedback about the adequacy of their teaching methods. To address this question, we calculated the mean prompt level required for each target in each session by assigning numerical scores for each prompt level based on the number and types of prompts used for the target (e.g., physical prompt = 4; model prompt = 3; partial model prompt = 2; gesture prompt = 1; no prompt = 0). If the therapist did not use a particular prompt level for a target (e.g., model prompts for motor imitation targets; gesture prompts for certain targets), the numerical scores reflected the number and types of prompts that the therapists actually used for that target. We then totaled these scores across all trials and divided this sum by the total number of trials with the target in the session (continuous, specific data). We calculated the mean prompt level required for each target across the first three trials of each session (discontinuous, specific data) by assigning numerical scores for each prompt level, totaling these scores, and dividing this sum by three. Finally, we examined the prompt level needed on the first trial of each session and assigned the appropriate numerical score for our first-trial only data (discontinuous, specific data).

After our data had been prepared as described above, we were interested in answering two questions. First, we were interested in determining if data on prompt levels would enhance the sensitivity of our continuous data. To do so, we calculated the number of sessions that the therapist conducted before the occurrence of the first data point that showed a decrease in the mean prompt level (relative to the first teaching session). This number was compared to the number of sessions that the therapist conducted before the data records for the nonspecific

Figure 5. Mean number of sessions before the first session with a decrease in prompt level when examining data from all trials, the first three trials, and the first trial (top panel). Mean number of sessions before the first nonzero point when examining the occurrence of unprompted correct responses on all trials, the first three trials, and the first trial (bottom panel). The narrow bars show the standard deviation for each.

data (i.e., percentage of trials with unprompted responses) revealed the first nonzero point. These nonspecific data records had already been generated to address Questions #1 and #2. Second, we were interested in determining if data on prompt levels would enhance the sensitivity of the discontinuous data. We were specifically interested in the discontinuous data records that showed less sensitivity than the continuous data in our analyses for Question #2. This subset of the discontinuous data was subjected to the same analysis as that just described for the

continuous data to determine if the use of specific recording would improve the sensitivity of these data.

Figure 5 shows the mean number of sessions (along with the standard deviation) that occurred before the continuous and discontinuous data revealed any change in performance during initial acquisition when data were collected on the prompt level needed to obtain a correct response (specific recording; top panel) versus when data were collected on unprompted correct responses (nonspecific recording; bottom panel). For

the continuous (all-trial) data, the mean number of sessions was 3 (range, 1 to 14 sessions) when examining changes in the prompt level and 4.7 (range, 1 to 14 sessions) when examining changes in the percentage of unprompted responses. These results suggest that data on the learner's required prompt level did not increase the sensitivity of the continuous data. In fact, the majority of targets (58%) showed no difference in the number of sessions. For the first-trial data, the mean number of sessions required to reveal a change in performance was 3.7 (range, 2 to 15) when examining changes in the prompt level and 9.8 (range, 4 to 21) when examining changes in the percentage of unprompted responses. This analysis was conducted for 22 of the 24 total targets, and all showed an increase in sensitivity with specific recording. For the three-trial data, the mean number of sessions required to reveal a change in performance was 3.2 (range, 2 to 15) for specific recording and 6.7 (range, 2 to 17) for nonspecific recording. Specific recording showed greater sensitivity to behavior change for 79% of the 14 targets included in this analysis. These findings suggest that data on the learners' required prompt level increased the sensitivity of the discontinuous data and that the amount of change in sensitivity was noteworthy.

Conclusions

Together, results of the present analyses suggest that first-trial recording could have led to premature determinations about skill mastery, particularly if the criterion was based on performance across two sessions. This finding is consistent with that obtained by Cummings and Carr (2009), who used a two-session mastery criterion. On the other hand, the instructors would have considered the majority of targets mastered in approximately the same amount of time under continuous and discontinuous recording if they had collected data on a larger subset of trials (i.e., three-trial recording) or if they had based the criterion on performance across three consecutive sessions. This latter finding is consistent with that obtained by Najdowski et al. (2009), who used a three-session mastery criterion.

Although the first-trial-only data were limited when determining skill mastery, results indicated that performance on the first trial was a good predictor of whether performance on all trials would either exceed or fall below 50% of correct trials. Thus, this method of discontinuous data collection may be useful for obtaining a rough estimate of performance under continuous recording. It should be noted, however, that both types of discontinuous data tended to underestimate all-trial performance, as indicated by our evaluation of overall mean levels of correct responding. We were concerned that correct guesses might artificially inflate first-trial data. These findings appear to suggest otherwise for the majority of targets. Furthermore, instruction and prompts that occurred early in the teaching session did not appear to inflate performance on the remainder of the trials because correct responding on all trials was typically below 50% when the participant responded incorrectly on the first trial.

Given that one key goal of data collection is to monitor ongoing progress, our analysis of sensitivity to changes in responding has important clinical implications. We found that both continuous data and prompt level data were more sensitive to changes in responding than either the discontinuous or nonspecific data. This was particularly striking when comparing the first-trial data to the all-trial data. Therapists conducted twice as many sessions, on average, before the first-trial data showed any improvement in responding when compared to the all-trial data. On the other hand, monitoring the prompt level needed to obtain a correct response did *not* increase the sensitivity of continuous recording. Nonetheless, it was beneficial for the discontinuous data, particularly the first-trial data.

Together, these findings suggest that data collected on a subset of trials (i.e., 3 out of 8 or 9 trials) would have adequate correspondence with continuous data and reveal similar changes in performance over time. Recording data on just the first trial would give a rough estimate of overall performance in the session, but this approach may lead to premature determinations of skill mastery. Instructors are likely to reduce or discontinue teaching sessions for skills that appear to be mastered. If the learner has not yet fully acquired these skills, his or her overall progress may be further delayed. First-trial data also appear to be relatively insensitive to initial changes in performance. Thus, the instructor must conduct more training sessions to determine if the learner is making adequate progress. This may postpone the implementation of important refinements to instruction.

It should be emphasized, however, that these findings may be particular to our analysis and teaching method. Regarding our analysis, we directly drew the first-trial and three-trial data from data that therapists had collected using continuous specific recording. This strategy ensured that (a) task difficulty would be equivalent across different recording methods, (b) any differences in the performance data would be due to differences in the recording method per se, and (c) we could conduct multiple comparisons of the methods. Nonetheless, extracting all of the data from the same records may have artificially increased the correspondence between the first-trial, three-trial, and all-trial data. Furthermore, the children's performance was not impacted by differences in therapist behavior (e.g., a reduction in procedural integrity; less conservative programming decisions; longer inter-trial intervals) that may have resulted from using different recording methods.

Regarding our teaching methods, targets were presented multiple times in each teaching session rather than distributed over longer periods of time. Although this structure is common in many discrete-trial teaching programs, therapists often use other instructional arrangements when teaching skills to individuals with autism. For example, embedding learning trials into daily routines or naturalistic situations typically results in more distributed practice of targets. Discontinuous data may be less predictive of continuous data when trials are spaced in time. Other elements of instruction that might impact the outcomes of our analyses and, hence, limit the generality of our

findings include the number of targets taught in each session, the number of trials conducted for each target, the number of sessions conducted per day, and the inclusion of mastered targets. Some of these elements varied across participants in the current study (i.e., number of different targets in each session, number of sessions per day), but these differences did not appear to be associated with different outcomes.

Recommendations for Best Practice

Our results, combined with those of Cummings and Carr (2009) and Najdowski et al. (2009), provide some guidelines for best practice. Like Cummings and Carr, we recommend the use of continuous recording if ease and efficiency of data collection are not a top concern. Continuous measurement is associated with a number of advantages compared to discontinuous measurement. Continuous recording provides more information about progress and, relative to discontinuous recording, has been shown to improve response maintenance when it leads to more stringent mastery criteria (Cummings & Carr). Continuous recording also may help monitor therapist behavior (i.e., document the number of trials delivered) during instructional sessions. Although we did not find any benefits for recording data on prompt level when using continuous recording, information about prompt level may be useful for other reasons (e.g., to track fading steps).

On the other hand, continuous recording has a number of disadvantages. Relative to discontinuous recording, it may result in lengthier sessions, delays to reinforcement (if the therapist records data prior to delivering the reinforcer), and poorer data reliability. It should be noted, however, that neither our study nor the prior two studies (Cummings & Carr, 2009; Najdowski et al., 2007) directly examined these possible effects of continuous recording on therapist behavior. Nonetheless,

our results and those of Najdowski indicate that discontinuous recording is a viable alternative to continuous recording when ease and efficiency are a concern. In these cases, we recommend recording performance on a small subset of trials (rather than the first trial only), including the prompt level needed on these trials, and basing the mastery criterion on performance across more extended sessions (e.g., a minimum of three sessions).

References

- Cummings, A. R., & Carr, J. E. (2009). Evaluating progress in behavioral programs for children with autism spectrum disorders via continuous and discontinuous measurement. *Journal of Applied Behavior Analysis, 42*, 57–71.
- 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.
- Love, J. R., Carr, J. E., Almason, S. M., & Petursdottir, A. I. (2009). Early and intensive behavioral intervention for autism: A survey of clinical practices. *Research in Autism Spectrum Disorders, 3*, 421–428.
- Meany-Daboul, M. G., Roscoe, E. M., Bourret, J. C., & Ahearn, W. H. (2007). A comparison of momentary time sampling and partial-interval recording for evaluating functional relations. *Journal of Applied Behavior Analysis, 40*, 501–514.
- Najdowski, A. C., Chilingaryan, V., Berstrom, R., Granpeesheh, D., Balasanyan, S., Aguilar, B., & Tarbox, J. (2009). Comparison of data collection methods in a behavioral intervention program for children with pervasive developmental disorders: A replication. *Journal of Applied Behavior Analysis, 42*, 827–832.

Action Editor: James Carr

A | B | A | I
Association for Behavior Analysis International

New ABAI Organizations

ABAI welcomes new affiliated chapters in 2011:

Albania ABA
Arizona ABA
Switzerland ABA

ABAI welcomes new special interest groups in 2011:

History of Behavior Analysis SIG
Multicultural SIG: Multicultural Alliance of Behavior Analysts

Learn more about ABAI's affiliated chapters and special interest groups on our website:

www.abainternational.org/chapters.asp
www.abainternational.org/Special_Interests/si.asp