EVALUATING PROGRESS IN BEHAVIORAL PROGRAMS FOR CHILDREN WITH AUTISM SPECTRUM DISORDERS VIA CONTINUOUS AND DISCONTINUOUS MEASUREMENT

ANNE R. CUMMINGS AND JAMES E. CARR

WESTERN MICHIGAN UNIVERSITY

We evaluated the influence of two different frequencies of data collection on skill acquisition and maintenance within behavioral treatment programs for children with autism spectrum disorders. Six children were taught multiple skills in up to four different behavioral programs. Half of the skills were measured continuously (i.e., trial by trial), and the other half were measured discontinuously (i.e., first trial only). When differences were detected, quicker acquisition was typically associated with discontinuous measurement, and stronger maintenance was typically associated with continuous measurement.

DESCRIPTORS: autism spectrum disorders, continuous measurement, discontinuous measurement, skill acquisition

The intensive delivery of behavior-analytic treatment has been shown to produce substantial improvements in the repertoires of children with autism spectrum disorders (Cohen, Amerine-Dickens, & Smith, 2006; Howard, Sparkman, Cohen, Green, & Stanislaw, 2005; Sallows & Graupner, 2005; Smith, 1999). Such treatment is often delivered during one-to-one instruction for up to 40 hr per week (e.g., Lovaas, 1987). Curricula are usually developmentally sequenced, with numerous skill programs (e.g., matching to sample, imitation) taught concurrently (Maurice, Green, & Luce, 1996). A common teaching technique employed in the behavioral treatment of autism is discrete-trial teaching (Ghezzi, 2007). This approach is essentially a restricted-operant arrangement in which, during each trial, a clear instructional antecedent is followed by an opportunity to respond, which is then followed by error correction or programmed reinforcement (Smith, 2001). Due to the intensity of many behavioral treatment programs and the concurrent implementation of multiple curricular programs, it is not unusual for thousands of trials to be presented during a week. Thus, the therapist needs to simultaneously balance a rather intensive acquisition technology with measurement of learner progress. This is often a significant challenge.

Two disparate approaches for measuring restricted-operant performance have emerged in the autism treatment community.[1] In *continuous measurement* systems, data regarding learner responding and prompt level are recorded for every trial (Lovaas, 2003). This approach allows a comprehensive, ongoing account of the learner's performance across all programmed learning opportunities. However, the effort required to measure behavior continuously has led some to recommend an alterative

This article is based on a dissertation submitted by the first author, under the supervision of the second author, in partial fulfillment of the requirements for the PhD degree in psychology at Western Michigan University. We thank committee members Jack Michael, Linda LeBlanc, Scott Gaynor, and Nickola Nelson for helpful comments on the dissertation. We also thank Kendra Priest and Stacy Osredkar for their assistance with data collection.

Anne R. Cummings is now affiliated with Central East Autism Program.

Address correspondence to James E. Carr, who is now at the Department of Psychology, Auburn University, 226 Thach Hall, Auburn, Alabama 36849 (e-mail: carr@auburn.edu).

doi: 10.1901/jaba.2009.42-57

[1] The issue of how much behavior should be measured during skill-acquisition sessions is relevant to multiple areas of application. However, the present study is framed in the context of the behavioral treatment of autism because measurement frequency is a widely discussed professional concern in this area.

approach, *discontinuous measurement* (Dollins & Carbone, 2003; Sundberg & Hale, 2003). During discontinuous measurement, data are recorded for a subset of learning opportunities (e.g., the first trial of a session, the first and last trials). Although continuous measurement might produce a better overall behavioral sample, frequent recording could increase the duration of sessions and interfere with important aspects of teaching (e.g., immediate reinforcer delivery, short intertrial intervals). Alternatively, discontinuous measurement produces an incomplete performance record. This could be detrimental to acquisition programming if, for example, a skill was falsely considered to be mastered because of an insufficient behavioral sample. Unfortunately, relatively little research exists to inform this debate.

Bijou, Peterson, Harris, Allen, and Johnston (1969) examined the effects of varying the frequency of observations on subsequent data analysis. The researchers recorded the frequency of a 4-year-old boy's verbalizations to other children during a play period each day in a preschool environment. The researchers then displayed frequency of verbalizations in three separate graphs. The first graph depicted the data continuously (i.e., every day). The second graph included data from every other day beginning with the 1st day. The third graph included data from every other day beginning with the 2nd day. The researchers found that teachers interpreted the three graphs similarly, regardless of the continuous or discontinuous nature of the graphed data. These findings appear to support the use of both continuous and discontinuous measurement approaches.

In a related investigation, Munger, Snell, and Loyd (1989) presented special education teachers with four graphs of student acquisition data from intervention programs across 60 days of continuous measurement (5 days per week). The first graph depicted the data continuously. The second graph included data from only 3 days per week (Mondays, Wednesdays, and Fridays). The third graph included data from only 2 days per week (Tuesdays and Thursdays). The final version included data from only 1 day per week (Wednesdays). Munger et al. found that teachers' judgments across graphs were similar when the continuous data graph depicted an ascending trend. However, when the continuous data graph depicted descending, stable, or variable trends, teachers' judgments differed across graphs. These findings appear to caution against the use of discontinuous measurement systems, given that it is difficult to predict the nature of graphed data a priori.

Although the Bijou et al. (1969) and Munger et al. (1989) investigations are relevant to the issue of measurement frequency, their procedures limit extrapolation of their findings to contemporary autism treatment programs. Judgments in both studies were based on data collected over an extended time frame and did not represent performance within a single acquisition program. In autism treatment programs, data from a single program are recorded and reported as responses within a rather restricted time frame (e.g., 10 min). Furthermore, neither of the aforementioned studies evaluated the impact of discontinuous measurement on treatment decisions. In other words, it is unknown how learner behavior would have changed had the results of visual inspection altered programming or determined whether an acquisition criterion had been met.

The literature on continuous and discontinuous measurement of free-operant behavior appears to be somewhat more relevant to the present problem. Continuous measurement methods (e.g., event recording, duration recording) are usually preferred for producing complete records of behavior during an observation (Johnston & Pennypacker, 1993). However, discontinuous measurement methods (e.g., interval recording, momentary time sampling) are chosen when certain restrictions are placed on the observer (e.g., the need to measure the behavior of multiple participants). A number of studies have compared various discontinuous measurement methods (usually partial-interval recording and momentary time sampling) with continuous measurement benchmarks (e.g., Harrop, Daniels, & Foulkes, 1990; Meany-Daboul, Roscoe, Bourret, & Ahearn, 2007; Rapp et al., 2007). Studies from this area have been able to determine (a) the relative utility of specific discontinuous measurement methods for estimating behavior and (b) how these methods affect clinical decisions (Meany-Daboul et al.; Rapp et al.). Thus, although the findings from these studies cannot be directly applied to restricted-operant procedures, they do serve as a methodological model for studying the impact of measurement within these procedures.

Given the importance of measurement to behavioral acquisition programs, the effort it requires of therapists, and the current disagreement over necessary measurement frequency, further research in this area is clearly warranted. Thus, the purpose of the current study was to compare performance under a continuous and a discontinuous measurement system across a number of curriculum areas in behavioral treatment programs for children with autism spectrum disorders. It was essential to compare the influence of these measurement systems across different curriculum areas to ensure that any similarities or differences were consistent across different curricular domains rather than specific to a certain type of learning program. The current study evaluated the two most extreme types of these measurement systems: trial-by-trial continuous measurement and first-trial-only discontinuous measurement.

METHOD

Participants and Setting

Six children with autism spectrum disorders participated in this study. All participants had at least 1 month of prior exposure to the discrete-trial teaching format, but did not receive additional behavioral intervention during the present study. Erin was 8 years old and had been diagnosed with autism. Her vocal behavior consisted of single-sound utterances and echolalia. Jeff was 5 years old and had been diagnosed with pervasive developmental disorder not otherwise specified (PDD-NOS). He had good vocal skills and was able to comment, answer questions, and request using full sentences. Patrick was 7 years old and had been diagnosed with autism. He had good vocal skills and was able to comment, answer questions, and request using full sentences. Peter was 5 years old and had been diagnosed with PDD-NOS. He had good vocal skills and was able to comment, answer questions, and request using two-word phrases. Mary was 4 years old and had been diagnosed with autism. Her vocal behavior consisted of single-sound utterances and echolalia. Allison was 6 years old and had been diagnosed with autism. She had limited vocal skills and was primarily echolalic, making single-sound utterances and some single words and approximations.

Each participant was assessed using the Behavioral Language Assessment Form (BLAF; Sundberg & Partington, 1998) to document his or her prestudy verbal repertoire. The BLAF is a rapid informant assessment that assesses 12 basic language-related skill areas (e.g., cooperation, motor imitation, conversation). Mean BLAF scores for Erin, Jeff, Patrick, Peter, Mary, and Allison were 1.9 (*SD* = 0.9), 4.3 (*SD* = 1.1), 4.4 (*SD* = 1.1), 4.1 (*SD* = 1.2), 2.5 (*SD* = 1.0), and 3.4 (*SD* = 1.5), respectively. Scores in each BLAF skill area are available from the corresponding author.

Sessions were conducted in a quiet area of the participant's home (Erin, Jeff, Peter, Mary, and Allison) or school (Patrick) and in a university research room (Peter and Mary). Sessions were conducted two to four times per day, 3 to 7 days per week by the first author. At the time of the study, the first author was a doctoral student in behavior analysis who had approximately 10 years of experience implementing behavioral treatment programs for children with autism.

Table 1
Description of Curricular Programs

Program	Description	Participant					
		E	J	Pa	Pe	Ma	Al
Nonvocal imitation	The experimenter instructed the participant to "do this" as she simultaneously modeled a motor action (e.g., blowing a kiss). A correct response was defined as the participant imitating the motor action.					A	M
Echoics (vocal imitation)	The experimenter instructed the participant to say a specific word (e.g., "say hot"). A correct response was defined as the participant imitating the experimenter's vocal behavior.	M					A
Receptive instruction	The experimenter instructed the participant to perform a simple action (e.g., "stand up," "look to the right"). A correct response was defined as the participant performing the action specified by the experimenter.	A	A		A		
Receptive discrimination	The experimenter presented three pictures (photographs or drawings) to the participant. The experimenter then instructed the participant to "show me —." A correct response was defined as the participant touching the picture specified by the experimenter.	A		A	A	M	A
Receptive by feature, function, and class	This program was identical to the receptive discrimination program except that the participant was instructed to select the picture depicting a particular feature (e.g., "show me the one that is green"), function (e.g., "show me the one that jumps"), or class (e.g., "show me the animal").		A	A	A		
Tacts	The experimenter held up a photograph of an object or person and said, "What [who] is it?" A correct response was defined as the participant naming the object or person.		A	A		A	
Requests	The experimenter held up a photograph of an object and asked the participant, "What do you want?" A correct response was defined as the participant saying or signing, "I want [object]."					A	A
Intraverbals (social questions)	The experimenter asked the participant a question about personal information (e.g., "When is your birthday?" "What is your brother's name?"). A correct response was defined as the participant appropriately answering the question.				M		
Intraverbals (occupations)	The experimenter asked the participant a question about the materials used in various occupations (e.g., "Who would use a stethoscope, needle, and tongue depressor?"). A correct response was defined as the participant naming the occupation (e.g., a doctor).		A	A	A		
Drawing	The experimenter instructed the participant to draw or spell a specific object or word (e.g., "draw a D"). A correct response was defined as the participant appropriately drawing the requested stimulus.		M	M			

Note. A = acquisition program and M = maintenance program; E = Erin, J = Jeff, Pa = Patrick, Pe = Peter, Ma = Mary, and Al = Allison.

One or two additional data collectors were present during each session. Data collectors were advanced undergraduate students who had received proficiency-based training on the study's methods and completed a course in behavior-analytic research methods.

Curricular Programs

Participants were exposed to two types of curricular program: programs that were new to them (acquisition programs) and programs in which they had previously mastered skills (maintenance programs). Descriptions of the acquisition and maintenance programs implemented during the study are presented in Table 1. Programs were selected for each participant based on parental interview and a direct-observation curriculum assessment.

Data Collection

During the course of the study, each participant received discrete-trial teaching in at least two acquisition programs. Each skill was taught in a 20-trial interspersed training format until the participant reached the acquisition criterion. The acquisition program trials comprised 10 of the trials and were scored using either continuous or discontinuous measurement. A trial from the maintenance program was presented after every acquisition program trial (a 1:1 ratio).

Each session consisted of 20 trials (10 acquisition and 10 maintenance). The dependent measure for each session was the percentage of correct responses in the acquisition program. Data on maintenance trials were not recorded. A correct response was defined as the participant responding correctly and independently (i.e., without prompts) to the instruction within 5 s. Data were collected continuously or discontinuously depending on the experimental condition in effect. Correct responses were recorded immediately after reinforcer delivery, and incorrect responses were recorded immediately after error correction. The mastery criterion for each skill was two consecutive sessions at 100%.

Two primary dependent variables were assessed for every skill taught under each measurement condition. The first was the number of sessions required to reach the mastery criterion (i.e., two consecutive sessions at 100%). The second was the percentage of correct responses produced during each follow-up probe (described below). In addition, the cumulative number of minutes spent in training before reaching the mastery criterion was also recorded.

Interobserver Agreement

Interobserver agreement was assessed for every session for each participant. An agreement was defined as two independent observers agreeing on whether a participant's behavior during a trial was correct or incorrect. Agreement was calculated by dividing the number of agreements by the number of agreements disagreements, and this ratio was converted to a percentage. Peter's mean agreement score was 98% (range, 90% to 100%). Agreement scores were 100% for Erin, Jeff, Patrick, Mary, and Allison. Interobserver agreement was not assessed for the cumulative number of minutes spent in training.

Procedure

Stimulus preference assessment. At the beginning of the study, a reinforcer identification interview was administered to each participant's caregivers to identify preferred foods and toys (Fisher, Piazza, Bowman, & Amari, 1996). These stimuli were incorporated into brief multiple-stimulus (without replacement) preference assessments (Carr, Nicolson, & Higbee, 2000) conducted at the beginning of each day's sessions. The top three ranked stimuli were used as programmed consequences during the day's subsequent sessions. Consequent stimuli included food (e.g., cookies, cereal), beverages (e.g., grape juice, soy milk), toys (e.g., bubbles, spinning light), and social interactions (e.g., singing, high fives).

Training structure. The framework for each scored acquisition training trial was as follows. First, the experimenter delivered an instruction to the participant, who was given 5 s to respond. If the participant responded correctly, he or she received immediate praise and brief access to one of the preferred stimuli. If the participant did not respond or responded incorrectly, the experimenter immediately provided corrective feedback (e.g., "try again") and then repeated the instruction. If the participant responded correctly, praise was provided immediately. If not, the experimenter used a verbal, gestural, or mild physical prompt (depending on the participant and skill) to occasion the correct response.

The framework for unscored maintenance-program trials was as follows. First, the experimenter delivered an instruction to the participant, who was given 5 s to respond. Immediately following the participant's response (correct or incorrect), the experimenter delivered a neutral statement (e.g., "okay"). If the participant did not respond within 5 s, the experimenter immediately presented the next acquisition trial.

The term *training set* will be used from this point forward to refer to two separate skills from the same acquisition program that were taught concurrently, one per measurement condition. Thus, sessions were partitioned into pairs that contained one session of a skill taught under continuous measurement and one session of a different skill taught under discontinuous measurement. Before a training set was introduced, a coin toss determined which skill was assigned to each condition. Although skills were not formally matched on response difficulty (e.g., number of syllables), they were chosen from the same curricular area (e.g., nonvocal imitation). Furthermore, when acquisition programs required the participant to respond to visual stimuli (e.g., receptive discriminations), the stimuli were matched in complexity (e.g., number of visual elements in a photograph).

Continuous measurement. During this condition, the experimenter recorded the participant's response during every acquisition program trial. Thus, each session score represented the percentage of correct responses in the 10 acquisition program trials.

Discontinuous measurement. During this condition, the experimenter recorded the participant's response for only the first of the 10 acquisition trials. Thus, each session score was either 0% or 100% correct. Because no data were recorded after the first trial, we were concerned that the experimenter might lose track of the number of trials left to be administered. In order for the experimenter to administer exactly 10 acquisition trials in a session, she came to the session with 20 different teaching stimuli: 10 stimuli for the acquisition program and 10 stimuli for the maintenance program. This helped to ensure that the experimenter conducted the appropriate number of trials, which is often a function served by continuous measurement.

Follow-up assessment. To assess the study's second primary dependent measure, probes were conducted for each skill 3 weeks after it had been mastered. Each probe was implemented using the general training format described above except that it was administered in a massed five-trial format with no programmed consequences (i.e., no preferred stimuli or error correction was provided). Data were recorded for every trial during follow-up probe sessions.

Experimental Design

An adapted alternating treatments design (Sindelar, Rosenberg, & Wilson, 1985) was used to evaluate the number of sessions to mastery and the percentage correct during follow-up probes across both measurement conditions. The adapted alternating treatments design is different from the conventional version because each condition is applied to a separate response rather than evaluating the effects of multiple conditions on the same response.

Procedural Integrity

The experimenter's behavior was scored during each trial to determine whether (a) stimuli were presented correctly, (b) the instruction was presented correctly, (c) the correct programmed consequence was delivered immediately after the participant's response, and (d) the experimenter recorded data on the trial (these data were examined for each session to verify the presence of continuous or discontinuous measurement). The procedural integrity score was calculated as the percentage of correct experimenter responses during a session. This procedural integrity measure was calculated for 100% of the sessions and was 100% for each participant. Interobserver agreement was assessed for 100% of these sessions using the point-by-point agreement formula. Agreement for the procedural integrity measure was 100%.

RESULTS

The primary findings are depicted in Figures 1 through 6. The panels in the left columns of each figure depict the number of sessions before the mastery criterion was met for skills under each condition. Of the 100 training sets (skill pairs) taught to participants, 49 reached the acquisition criterion more quickly under discontinuous measurement, 16 reached the acquisition criterion more quickly under

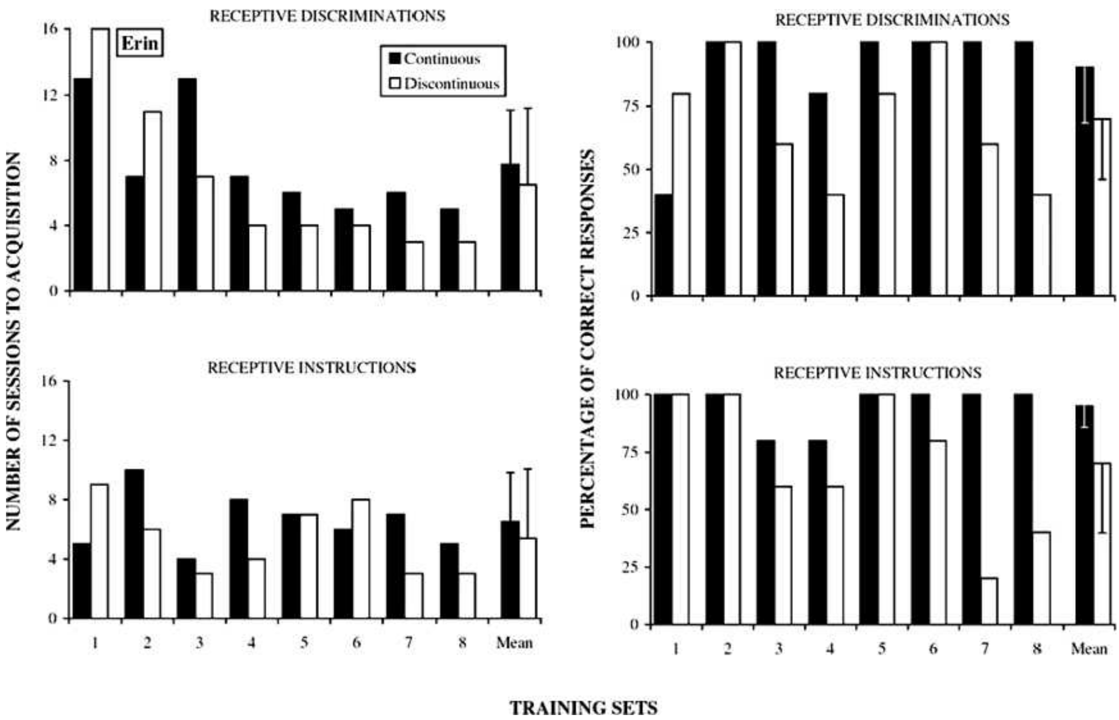

Figure 1. Number of sessions to skill mastery under continuous (every trial) and discontinuous (first trial) measurement across program areas for Erin (left). Percentage of correct responses during follow-up assessments of skills previously mastered under continuous and discontinuous measurement across program areas (right). Error bars represent standard deviations.

continuous measurement, and there was no difference for 35 training sets. Furthermore, training in the discontinuous measurement condition resulted in 192, 48, 144, 32, 108, and 126 fewer trials per skill for Erin, Jeff, Patrick, Peter, Mary, and Allison, respectively. These trial savings translated to 100, 33, 68, 24, 55, and 65 fewer minutes in the discontinuous measurement condition than in the continuous measurement condition for Erin, Jeff, Patrick, Peter, Mary, and Allison, respectively.

The panels in the right columns of Figures 1 through 6 depict the performance of skills (under extinction) 3 weeks after the mastery criterion had been met. Of the 100 training sets taught across participants, 33 were maintained better at follow-up under continuous measurement, 66 showed equal maintenance across conditions, and 1 was maintained better at follow-up under discontinuous measurement. Of the 49 training sets that were acquired more quickly in the discontinuous measurement condition, 22 were maintained more poorly at follow-up than their counterparts that were acquired in the continuous measurement condition.

DISCUSSION

We provided the first analysis of how continuous (i.e., trial by trial) and discontinuous (i.e., first trial only) measurement systems might influence the behavioral treatment of children with autism spectrum disorders. Comparisons were made across 6 children and eight different curricular programs for a total of 100 comparisons. Specific exemplars ranged from simple nonvocal imitation (e.g., clapping hands) to answering questions (e.g., "Who would use a bandage?"). In general, the findings indicate that when performance was discontin-

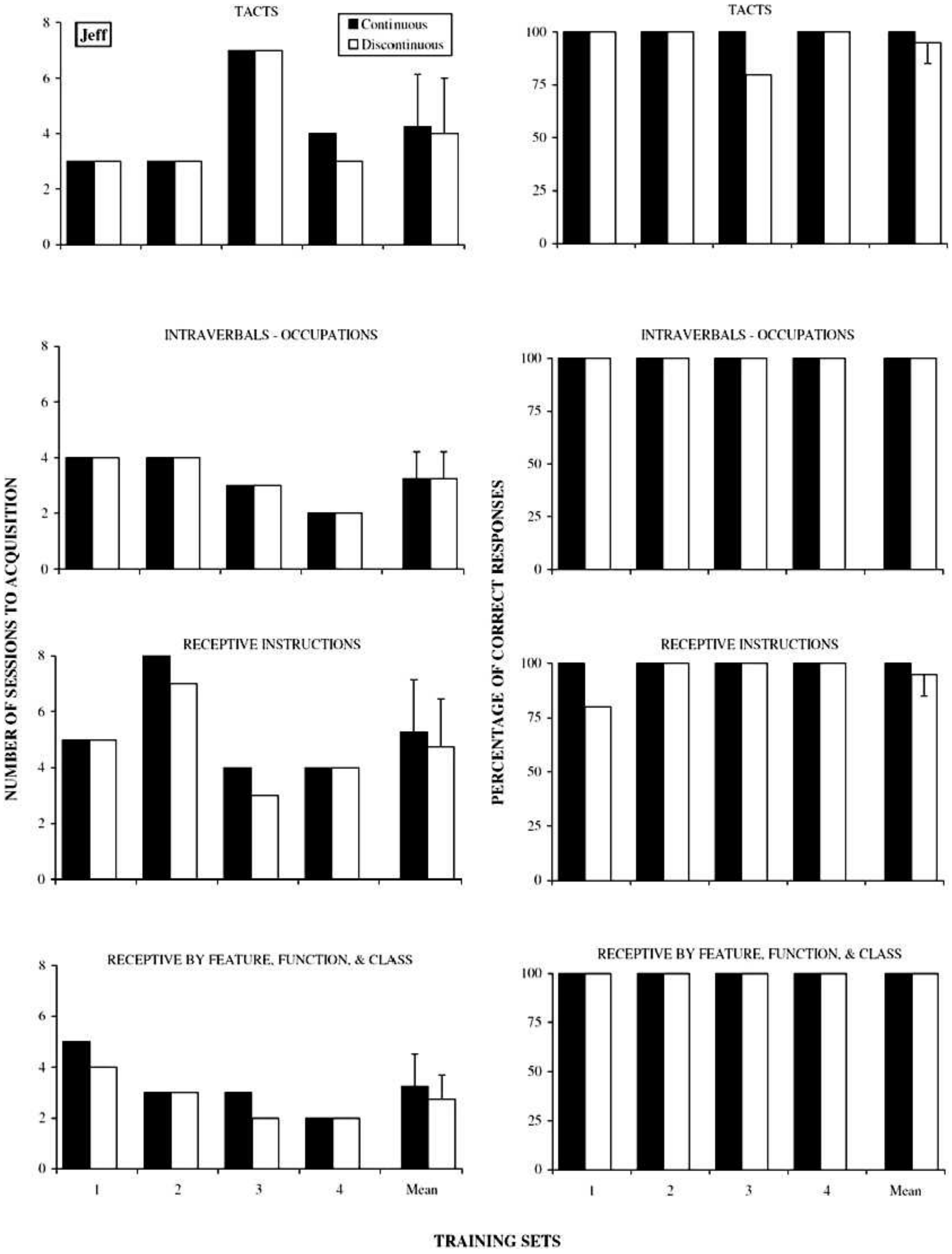

Figure 2. Number of sessions to skill mastery under continuous (every trial) and discontinuous (first trial) measurement across program areas for Jeff (left). Percentage of correct responses during follow-up assessments of skills previously mastered under continuous and discontinuous measurement across program areas (right). Error bars represent standard deviations.

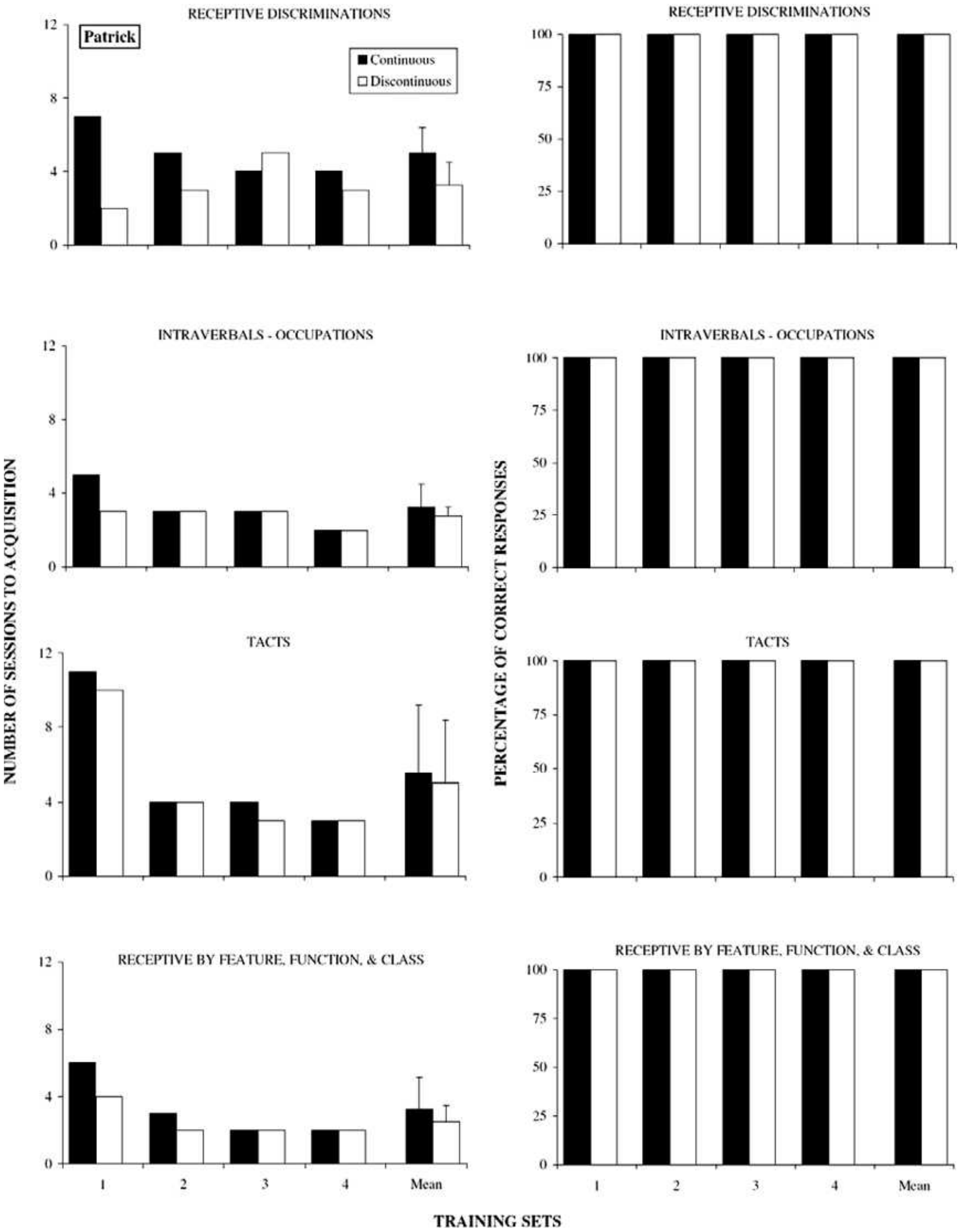

Figure 3. Number of sessions to skill mastery under continuous (every trial) and discontinuous (first trial) measurement across program areas for Patrick (left). Percentage of correct responses during follow-up assessments of skills previously mastered under continuous and discontinuous measurement across program areas (right). Error bars represent standard deviations.

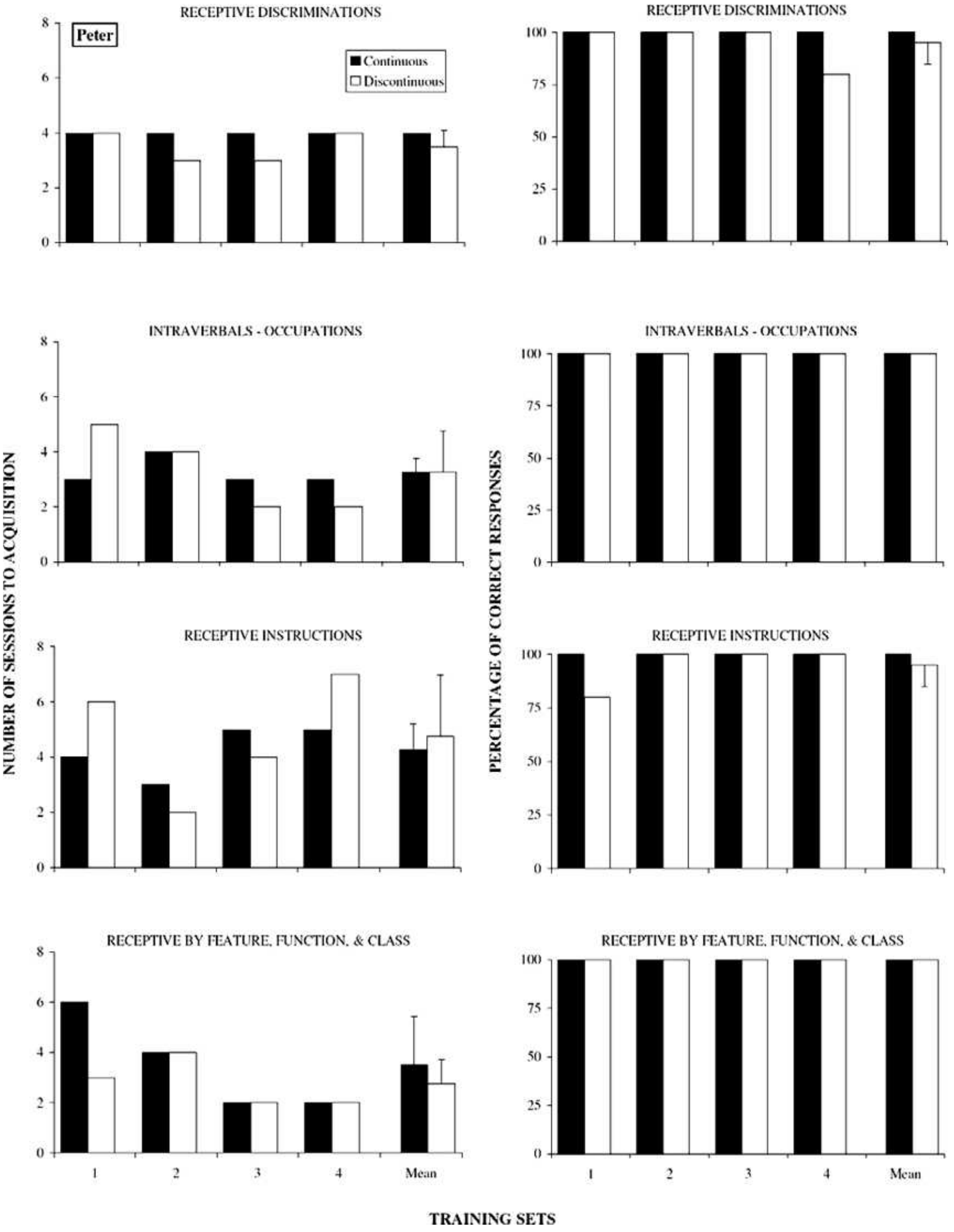

Figure 4. Number of sessions to skill mastery under continuous (every trial) and discontinuous (first trial) measurement across program areas for Peter (left). Percentage of correct responses during follow-up assessments of skills previously mastered under continuous and discontinuous measurement across program areas (right). Error bars represent standard deviations.

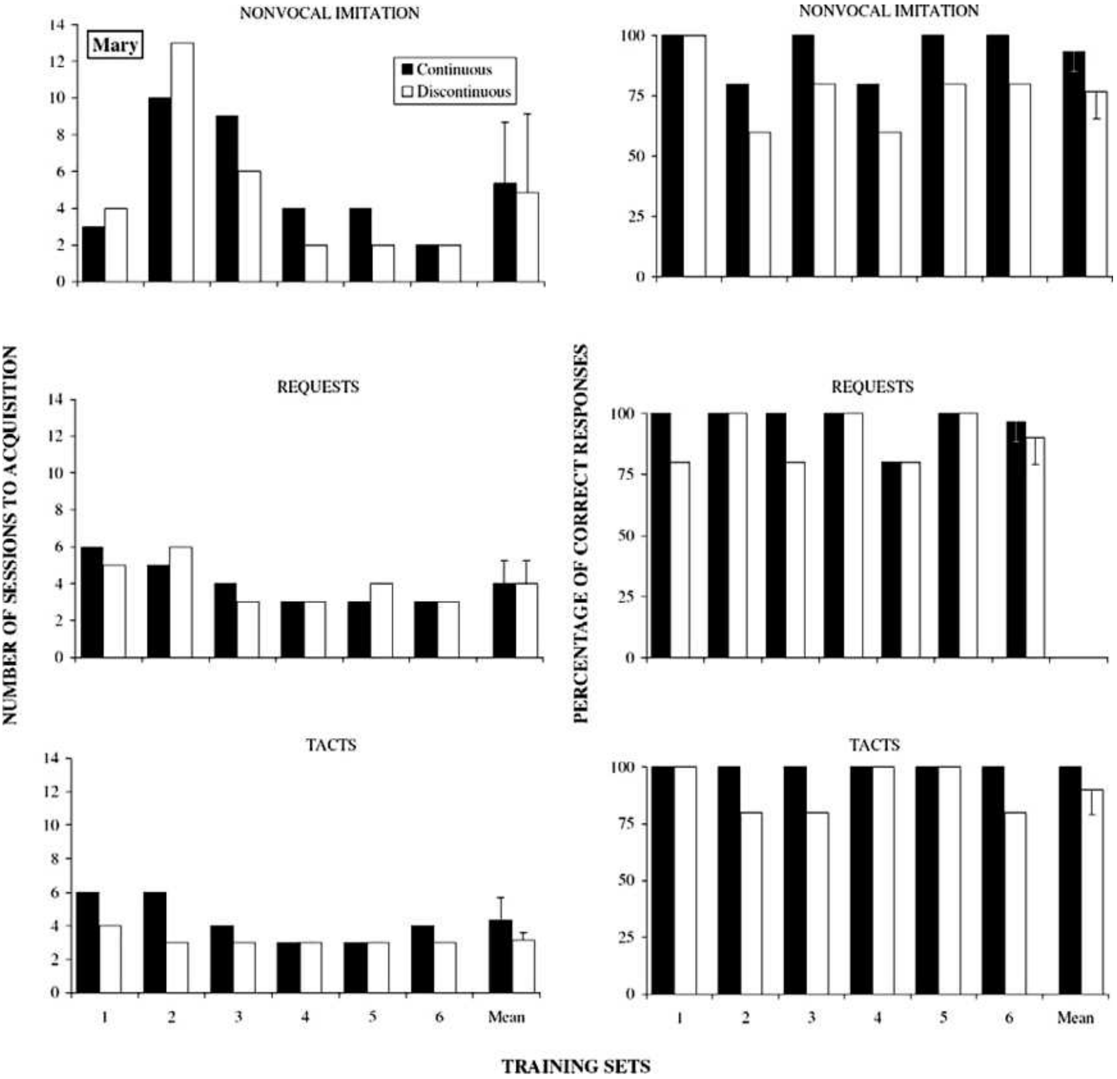

Figure 5. Number of sessions to skill mastery under continuous (every trial) and discontinuous (first trial) measurement across program areas for Mary (left). Percentage of correct responses during follow-up assessments of skills previously mastered under continuous and discontinuous measurement across program areas (right). Error bars represent standard deviations.

uously measured, skills were usually acquired in fewer sessions (less time) than when performance was continuously measured. However, skills sometimes were maintained slightly better at a 3-week follow-up assessment when performance was measured continuously. It is possible that some of the variance in the data might have been a result of not formally controlling skill difficulty between conditions, but because skills were randomly assigned to measurement conditions, it is likely that these effects were a function of the measurement conditions.

An interesting phenomenon exists in the data that is worth noting because of its relevance to methodology in this area. Some of the participants (e.g., Patrick) acquired each new skill exemplar at a relatively rapid pace (e.g., within two to three sessions). By using a

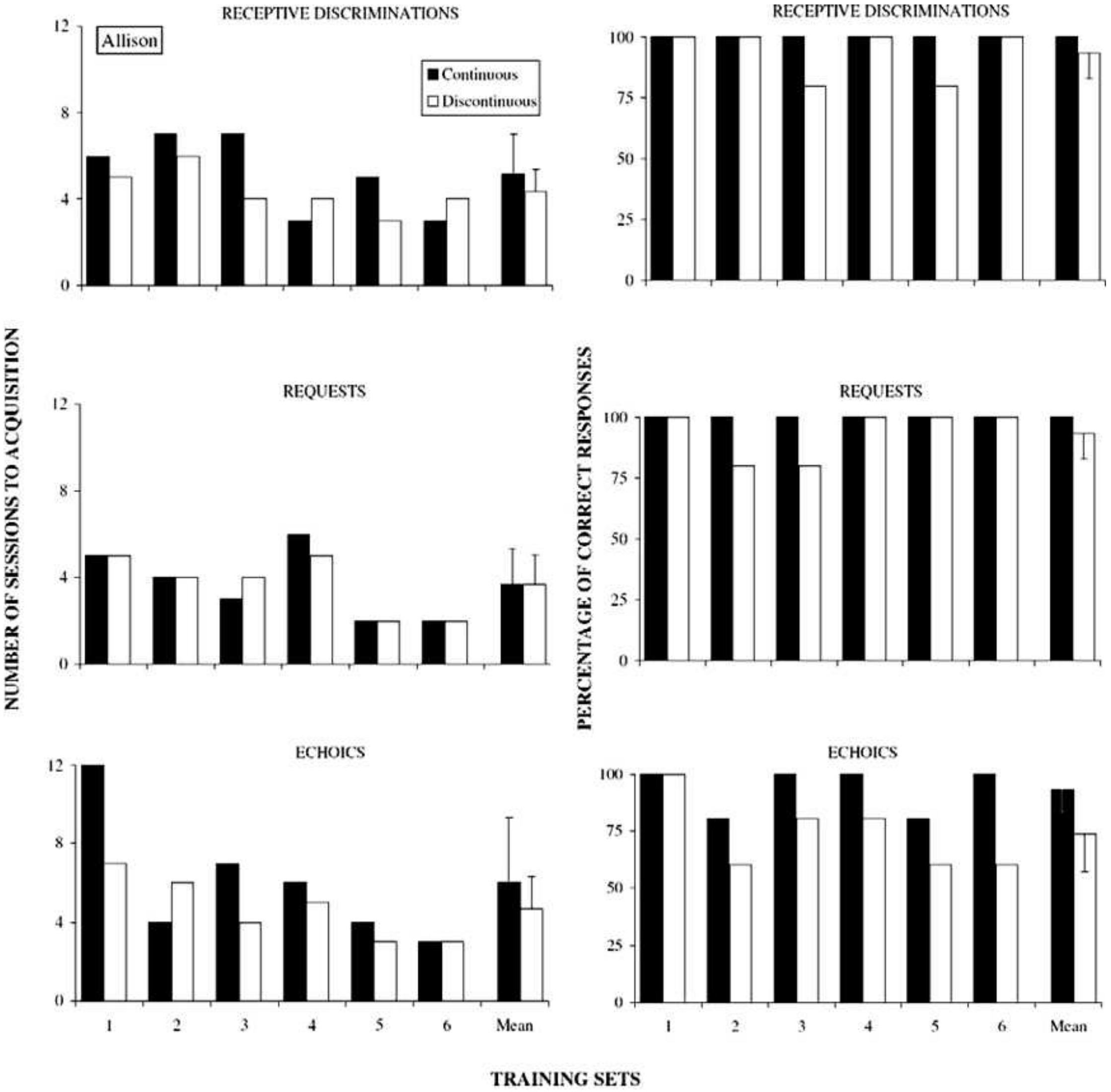

Figure 6. Number of sessions to skill mastery under continuous (every trial) and discontinuous (first trial) measurement across program areas for Allison (left). Percentage of correct responses during follow-up assessments of skills previously mastered under continuous and discontinuous measurement across program areas (right). Error bars represent standard deviations.

research strategy in which exemplars within a curricular area are randomly assigned to different experimental conditions (e.g., an adapted alternating treatments design), it is necessary for exemplars to be acquired at a moderate-to-slow pace such that acquisition differences can be detected. In the present investigation, Erin, Mary, and Allison appear to best meet this profile and thus might be considered the purest assessment of the study's research question. For these 3 participants, performance at follow-up was stronger in the continuous condition for 29 of 52 comparisons and was stronger in the discontinuous condition for only two of these comparisons. Follow-up performance was equivalent in 21 of the comparisons.

The most parsimonious explanation for the present findings is that skills were prematurely considered to be mastered due to the incomplete performance record produced by the discontinuous measurement system. For example, although a skill might have been correct on the first trial of two consecutive sessions (satisfying the acquisition criterion), performance on subsequent trials might have been mixed. However, because performance on these trials was not recorded, the skill's strength was not detected. Thus, 3 weeks later, skills measured discontinuously may be weaker than the ones that met an acquisition criterion under continuous measurement.

The current study provides preliminary evidence against which some of the claims made regarding the influence of continuous and discontinuous measurement can be evaluated. Dollins and Carbone (2003) have suggested that teaching and continuously recording data would result in lengthier sessions. The present data support this claim, although the social importance of the difference was not determined. A second claim is that recording data continuously might interfere with teaching by, for example, delaying reinforcer delivery while the therapist records whether the skill was correctly performed. The current study does not directly inform this issue because delays between responses and reinforcers were not recorded, and the experimenter in the present study recorded responses only after reinforcers were delivered. Delays in reinforcer delivery presumably would have resulted in slower acquisition under the continuous measurement condition. Because data were collected following reinforcer delivery in the present study, slower acquisition during the continuous measurement condition was more likely a function of longer intertrial intervals (perhaps due to data collection) or the stringent mastery criterion (100% correct).

In addition to being associated with slightly better skill maintenance at follow-up, continuous measurement has at least one additional benefit over discontinuous measurement systems. If data were recorded not only on the learner's response (i.e., correct vs. incorrect) but also on the type of prompt required to occasion the response (e.g., gestural vs. physical), a clinician could determine whether a learner was making progress by requiring progressively less intrusive prompts, even though the percentage correct score might be low. Although these kinds of data could still be recorded during discontinuous measurement, such patterns are more likely to be evident at the within-session level rather than the between-sessions level. Thus, continuous measurement (or a less discontinuous system than the one evaluated in the present study) might better enable a clinician to evaluate progress across prompt levels. Ultimately, this is an empirical question that could be answered in future studies.

The findings of the present study should be evaluated in light of five variables that might affect their generality. First, discontinuous measurement has primarily been recommended in the context of language programs based on Skinner's analysis of verbal behavior (see Sundberg & Partington, 1998). It is common in these programs for different curricular programs (e.g., mands, tacts, intraverbals) to be taught in an interspersal or mixed format such that a learner might be presented with, for example, four different programs across four trials (e.g., Arntzen & Almås, 2002). In the current study, sessions included only two programs (acquisition and maintenance), and data were recorded for only one of them. It is possible that discontinuous and continuous measurement might differentially affect learning with a more varied interspersal format. For example, if every trial constituted a different program, continuous measurement might indeed interfere with teaching.

Second, the present study was conducted with a therapist (the first author) with 10 years of experience collecting data using continuous methods. Thus, it would be important to examine the same research question with less experienced therapists who might not balance teaching and measurement demands as effectively.

Third, after a skill met the mastery criterion in the present study, it was never assessed again until the 3-week follow-up assessment. However, it is more common in clinical practice for previously acquired skills to be periodically reassessed and reinforced. Such maintenance programming might mitigate the present findings.

Fourth, our modest acquisition criterion (two consecutive sessions at 100% correct) might have influenced the findings. Had a more stringent criterion (e.g., four consecutive sessions at 100% correct) been employed, it is possible that the probability of falsely concluding that a skill was mastered under discontinuous measurement would have been reduced. In this way, an acquisition criterion might be conceptualized as an independent variable. Thus, we recommend that future researchers consider the specific impact of their acquisition criterion.

Finally, although errorless procedures are often recommended for skill acquisition (MacDuff, Krantz, & McClannahan, 2001), we did not employ them in the present study so that there would be sufficient time to detect differences between conditions. As mentioned earlier, differential treatment effects can be difficult to detect when learning occurs too rapidly. Nonetheless, such an evaluation is warranted to determine whether the present effects can be replicated with different prompting hierarchies.

The present study suggests that both continuous and discontinuous measurement tactics interact with behavioral teaching methods to affect learning, but in different ways. Continuous measurement appears to increase session duration but might lead to more conservative decisions regarding skill mastery, such that long-term maintenance is enhanced. Furthermore, continuous measurement permits the within-session analysis of response and prompt patterns. Future research is necessary to determine (a) whether the present outcomes can be replicated, (b) whether the same effects can be reproduced with less experienced therapists, and (c) whether the potentially detrimental long-term effects associated with discontinuous measurement can be mitigated with interim maintenance training. Until additional research is conducted, our preliminary recommendation is that teachers use continuous measurement methods unless session duration takes priority over other performance outcomes.

REFERENCES

Arntzen, E., & Almås, I. K. (2002). Effects of mand-tact versus tact-only training on the acquisition of tacts. *Journal of Applied Behavior Analysis*, *35*, 419–422.

Bijou, S. W., Peterson, R. F., Harris, F. R., Allen, K. E., & Johnston, M. C. (1969). Methodology for experimental studies of young children in natural settings. *The Psychological Record*, *19*, 177–210.

Carr, J. E., Nicolson, A. C., & Higbee, T. S. (2000). Evaluation of a brief multiple-stimulus assessment in a naturalistic context. *Journal of Applied Behavior Analysis*, *33*, 353–357.

Cohen, H., Amerine-Dickens, M., & Smith, T. (2006). Early intensive behavioral treatment: Replication of the UCLA model in a community setting. *Journal of Developmental and Behavioral Pediatrics*, *27*, S145–S155.

Dollins, P., & Carbone, V. J. (2003, May). Using probe data recording methods to assess learner acquisition of skills. In V. J. Carbone (Chair), *Research related to Skinner's analysis of verbal behavior with children with autism*. Symposium conducted at the 29th annual convention of the Association for Behavior Analysis, San Francisco.

Fisher, W. W., Piazza, C. C., Bowman, L. G., & Amari, A. (1996). Integrating caregiver report with a systematic choice assessment to enhance reinforcer identification. *American Journal on Mental Retardation*, *101*, 15–25.

Ghezzi, P. M. (2007). Discrete trials teaching. *Psychology in the Schools*, *44*, 667–679.

Harrop, A., Daniels, M., & Foulkes, C. (1990). The use of momentary time sampling and partial interval recording in behavioural research. *Behavioural Psychotherapy*, *18*, 121–127.

Howard, J. S., Sparkman, C. R., Cohen, H. G., Green, G., & Stanislaw, H. (2005). A comparison of intensive behavior analytic and eclectic treatments for young children with autism. *Research in Developmental Disabilities*, *26*, 359–383.

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

Lovaas, O. I. (1987). Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Consulting and Clinical Psychology*, *55*, 3–9.

Lovaas, O. I. (2003). *Teaching individuals with developmental delays: Basic intervention techniques*. Austin, TX: Pro-Ed.

MacDuff, G. S., Krantz, P. J., & McClannahan, L. E. (2001). Prompts and prompt-fading strategies for people with autism. In C. Maurice, G. Green, & R. M. Foxx (Eds.), *Making a difference: Behavioral intervention for autism* (pp. 37–50). Austin, TX: Pro-Ed.

Maurice, C., Green, G., & Luce, S. C. (Eds.). (1996). *Behavioral intervention for young children with autism: A manual for parents and professionals*. Austin, TX: Pro-Ed.

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.

Munger, G. F., Snell, M. E., & Loyd, B. H. (1989). A study of the effects of frequency of probe data collection and graph characteristics on teachers' visual analysis. *Research in Developmental Disabilities*, *10*, 109–127.

Rapp, J. T., Colby, A. M., Vollmer, T. R., Roane, H. S., Lomas, J., & Britton, L. N. (2007). Interval recording for duration events: A re-evaluation. *Behavioral Interventions*, *22*, 319–345.

Sallows, G. O., & Graupner, T. D. (2005). Intensive behavioral treatment for children with autism: Four-year outcome and predictors. *American Journal on Mental Retardation*, *110*, 417–438.

Sindelar, P. T., Rosenberg, M. S., & Wilson, R. J. (1985). An adapted alternating treatments design for instructional research. *Education and Treatment of Children*, *8*, 67–76.

Smith, T. (1999). Outcome of early intervention for children with autism. *Clinical Psychology: Science and Practice*, *6*, 33–49.

Smith, T. (2001). Discrete trial training in the treatment of autism. *Focus on Autism and Other Developmental Disabilities*, *16*, 86–92.

Sundberg, M. L., & Hale, L. (2003, May). Using textual stimuli to teach vocal-intraverbal behaviors. In A. I. Petursdottir (Chair), *Methods for teaching intraverbal behavior to children*. Symposium conducted at the 29th annual convention of the Association for Behavior Analysis, San Francisco.

Sundberg, M. L., & Partington, J. W. (1998). *Teaching language to children with autism or other developmental disabilities*. Pleasant Hill, CA: Behavior Analysts, Inc.

Received October 19, 2006
Final acceptance July 8, 2008
Action Editor, Gregory Hanley