Journal of
Applied Behavior Analysis

Journal of Applied Behavior Analysis 2022, **9999,** 1–15 NUMBER 9999 ()

Units of analysis in acquisition-performance criteria for "mastery": A systematic replication

Kristina K. Wong and Daniel M. Fienup

Department of Health and Behavior Studies, Teachers College Columbia University

This study compared 2 units of analysis for assessing acquisition mastery during sight word instruction for 3 participants. The unit of analysis refers to the specific performances that criteria are applied to, either sets of stimuli or individual operants. In the Set Analysis condition, we applied the acquisition-performance criterion to the aggregated accuracy of a set of 4 target operants. In the Operant Analysis (OA) condition, we assessed the criterion for individual operants and replaced targets as they met the acquisition criterion. All participants acquired novel textual responses to sight words faster under the OA condition and response maintenance was similar between conditions. This study extended previous research by showing enhanced response maintenance in the OA condition by increasing the performance criterion from 1 observation of 100% accuracy to 2. This study also suggests a unique contribution of OA to quickening learning.

Key words: acquisition-performance criterion, mastery criterion, operant analysis, response maintenance, set analysis, unit of analysis

Applied Behavior Analysis (ABA) emphasizes the study of the behavior of individual organisms. Radical behaviorism was a dramatic departure from conventional schools of psychological thought because it rejected conclusions based on statistical means and groups of organisms. Adolphe Quetelet, a prominent mathematician in the 19th century, influenced data interpretation in the social sciences by introducing the concept of the "average man" (Donnelly, 2015). However, behavior analysts criticize psychologists' treatment of data that use normal distribution curves to identify an "average." Behavior analysts argue that this approach neglects individuals because, quite often, aggregated "averages" do not represent the behavior of any particular individual. Instead, behavior analysts prioritize the study of an individual's behavior and the controlling variables of that individual's behavior (Chiesa, 1994).

Correspondingly, ABA pedagogies, such as Discrete Trial Instruction (DTI), learn unit instruction, and personalized systems of instruction (Albers & Greer, 1991; Keller, 1968; Lovaas, 1987), are applied at the level of the individual student wherein the individual's behavior determines consequences, modifications to programming, and the progression from one learning goal to the next. With respect to the latter, instructors establish an acquisition-performance criterion, which is often referred to as the mastery criterion[1] (Richling

The authors declare no conflict of interest. Data are available by contacting the corresponding author.

This study was completed by the first author in partial completion of a Ph.D. in Applied Behavior Analysis at Teachers College Columbia University under the mentorship of the second author. We thank Dr. R. Douglas Greer, Dr. Jessica Dudek, Dr. Matthew Zajic, and Dr. Sarah Richling for helpful comments on an earlier version of this manuscript. We thank Dr. Terry Falcomata for comments on a late draft of this manuscript. We thank Kyla Mackay, Carli Heiman, Maren Jacobson, Regina Spilotras, and Dr. Jessica Horton for assistance with data collection.

Address correspondence to: Daniel M. Fienup, Department of Health and Behavior Studies, Teachers College Columbia University, 525 W. 120 Street, Box 223, New York, NY 10027. Email: fienup@tc.columbia.edu

doi: 10.1002/jaba.915

[1]We believe the term "mastery criterion" is misleading because it is typically applied to the acquisition phase of learning and one can be said to have "mastered" some objectives without having demonstrated important outcomes such as maintenance or generalization (for a more in-depth discussion, see Richling et al., in press). In this paper, we refer to acquisition-performance criteria to specify the application of the performance criterion to the acquisition phase (Fienup & Carr, 2021).

et al., in press). When a student performs at the predetermined acquisition-performance criterion, the instructor concludes that the response or set of responses are acquired (or "mastered") and that the current phase of intervention may discontinue (Fienup & Carr, 2021; Fuller & Fienup, 2018). While ABA instructors apply performance criteria at the level of the individual, those performances are typically *a set of multiple independent operants*, or skills. That is, measured performances are *aggregated* across multiple independent responses of the individual student.

Wong et al. (2021) shed light on this apparent discrepancy within our field: the rejection of aggregating performances across individuals but apparent acceptance of aggregating performances within an individual. Wong et al. noted two potential problems with aggregating independent responses into a single measure of performance: (1) aggregating performance may mask one or more responses that are not accurate (e.g., "if an instructor teaches a set of four operants, each with five response opportunities per session and a mastery criterion of 90% correct, a student can be declared to have mastered the set of operants despite responding with only 60% accuracy [3/5 correct] to one of the operants," p. 3), and (2) aggregated performances may be most representative of the slowest to learn responses. Wong et al. studied this phenomenon by manipulating the unit of behavior at which the acquisition-performance criterion was applied to children with disabilities. In one condition, participants learned four sight words at a time and instruction continued until the aggregated performance was 100% accuracy in a single session—called the set analysis (SA) condition. In another condition, participants learned four sight words at a time and instruction continued until performance was 100% accuracy in one session for a single sight word—called the operant analysis (OA) condition. At that point in the OA condition, the sight word that achieved the acquisition-performance criterion was replaced with a new word in the next session and instruction continued. Wong et al. found that participants learned sight words quicker in the OA condition and maintained a higher number of sight words. The results provided preliminary evidence of unnecessary overtraining in the SA condition because aggregated performance measures, indeed, were not representative of individual performances.

Wong et al.'s (2021) study joined an emerging body of literature evaluating acquisition-performance criteria with children with disabilities.[2] Other studies in this area have demonstrated the importance of requiring high levels of accuracy during acquisition and the relation between acquisition performances and response maintenance. Specifically, researchers have demonstrated that a minimum of 90% accuracy (Fuller & Fienup, 2018; Longino et al., 2021; Pitts & Hoerger, 2021) or 100% accuracy (Richling et al., 2019) is needed during acquisition to produce high levels of response maintenance 3 to 4 weeks after instruction is terminated. Wong et al. added to this literature by demonstrating that the unit of analysis for acquisition-performance criteria also affected the acquisition and maintenance of responses. However, Wong et al.'s study had two notable limitations. First, Wong et al. included a decision-making protocol for modifying instruction when inadequate progress was observed. This resulted in modifications to instruction in the OA condition and calls into question whether the application of performance criteria to individual operants, the decision-making protocol, or both explained acquisition differences between conditions. Second, Wong et al. reported two measures of maintenance. The OA condition produced a

[2]Researchers have evaluated the effects of performance criterion with college students during personalized systems of instruction (Johnston & O'Neill, 1973; Semb, 1974) and equivalence-based instruction (Fienup & Brodsky, 2017) and also found high levels of performance during acquisition promote maintenance and the emergence of derived relations.

larger number of maintained operants than the SA condition because the OA condition resulted in learning more operants in a fixed amount of time. However, when examining the percentage of operants maintained, the OA condition resulted in less reliable production of responses that maintained over 3 to 4 weeks for a subset of participants.

The purpose of this study was to systematically replicate the procedures of Wong et al. (2021) and extend the previous findings by addressing key limitations. Specifically, we sought to determine the effects of an operant-based acquisition-performance criterion and set-based acquisition-performance criterion on acquisition rate and maintenance of sight word reading skills. In this study, we taught sight words to three participants by applying acquisition-performance criterion to individual sight words (OA condition) or sets of sight words (SA condition). The conditions were arranged in the same manner as in Wong et al. To address the systematic confound of the decision-making protocol, we eliminated this from the current study. To address the lower reliability of maintenance, we increased the acquisition-performance criterion from 100% accuracy in one observation to 100% accuracy across two consecutive observations/sessions (Schneider et al., 2021) for both OA and SA conditions.

Method

Participants

Three elementary students participated in the study. The students attended a public elementary school in a self-contained special education classroom. The classroom teachers implemented the Comprehensive Application of Behavior Analysis to Schooling (CABAS®) model (Greer, 2002). This educational model implements a combination of technologies based on research in applied behavior analysis in order to teach academics, self-management, and verbal behavior. Eligibility inclusion criteria included the following: (a) attention to instructors and instructional tasks for 10 consecutive minutes with minimal prompts for redirection, (b) emission of three- to five-word mand and tact utterances, (c) emission of echoics for one or more syllable words, and (d) their community of reinforcers included social praise. The experimenters included these criteria to ensure participants could engage in the respective academic task. The experimenter assessed the aforementioned inclusion criteria by conducting baseline observation sessions as a part of the *Early Learner Curricula and Achievement Record* (ELCAR; Greer et al., 2019) with all participants prior to the start of the study. Additionally, the participants' Individualized Education Plan (IEP) had academic goals that were directly related to learning textual responses to sight words. Participants completed standard sight-word instruction as a part of their daily academic programming. Thus, the intervention procedures did not interfere with the necessary instruction they would have received on a daily basis, regardless of their participation in the study.

Patrick was a 6-year-old male in first grade, educationally classified with a Speech and Language Impairment (SLI), and received behavior-analytic services in a CABAS® classroom for two years. Patrick had a large verbal repertoire that included beginning reading and writing repertoires. His educational level at the onset of the study included reading Level D stories proficiently from the Reading A-Z curriculum. Patrick accurately identified over 200 words from the Fry Sight Word List (Fry, 2004) as well. The second participant, Katie, was a 5-year-old female in kindergarten. Katie was educationally classified with Autism Spectrum Disorder (ASD) and this was her first year receiving behavior-analytic services in a CABAS® classroom. Katie had a large verbal repertoire that included beginning reading and writing repertoires. Her education level at the

onset of the study included reading Level E stories from the Reading A-Z curriculum. She accurately identified over 200 words from the Fry Sight Word list. William was also a 5-year-old male student in kindergarten educationally classified with SLI, and this was his first year receiving behavior-analytic services in a CABAS® classroom. William had a verbal repertoire that included some mands and tacts. At the onset of the study, William was working on reading Level AA storybooks from the Reading A-Z curriculum, and he accurately identified fewer than 10 words in repertoire from the Fry Sight Word list. All participants had a history of instruction that closely mimicked the SA performance-criterion condition.

We report data from a fourth participant, Zara (see online Supplemental A). Zara was an 8-year-old female student in third grade, and she received behavior-analytic services in a CABAS® classroom for 2 years. Due to several issues that arose during her analysis, we report her data separately.

Setting

The experimenter conducted each in-person session of the study within the participants' self-contained kindergarten through second grade classroom of a public elementary school. The participants attended a classroom with eight students, one teacher, and two teacher aides. Each session took place at a student desk that was positioned in one of the corners of the classroom or in the front of the room with minimal visual distractors. The experimenter sat at the desk beside the participant during all sessions of the study. All sessions were conducted in person for Patrick and Katie.

For William, due to hybrid in-person/remote learning models during his school year, 70% of sessions were conducted in person in the same setting as Patrick and Katie and 30% of the intervention and postintervention sessions took place over Zoom® video calls. During remote sessions, William sat at his kitchen table next to his mom, with a laptop on the table.

Materials

The experimenter used a PowerPoint® slideshow presented on a 34.29 cm MacBook laptop to deliver sight word instruction for each condition of the study; this was held constant across remote and in-person instruction for William as well. During instruction, the sight words were presented in black font with four font variations, including Times New Roman, Comic Sans MS, Century Gothic, and Calibri. Each word was positioned in the center of the slide with a white background and size 100 pt. Additional data collection materials included a black-inked pen, data sheets, and treatment fidelity data sheets.

The experimenter curated a list of 40 four-syllable words to teach Patrick and Katie by conducting a Google internet search. Words for these participants were selected in this manner to increase the probability that Patrick and Katie would not have encountered the words in their past instructional history and would not encounter the words in their daily lives throughout the duration of the study (See online supplemental information). We did not use any words from the Fry's Word List, as in Wong et al. (2021) because Patrick and Katie accurately identified many more sight words compared to the other participants. We curated a list of 24 one-syllable words from the Fry's Word List (Fry, 2004) to teach William (See online supplemental information). The Fry's Word List was a part of William's daily curriculum, but the specific words taught in the study were more advanced than those presented in William's current programming, which ensured that he did not contact the target stimuli during daily instruction outside of the study. (Specific word assignments to conditions can be viewed in the online supplemental information).

Measurement

The dependent variable was a participant's accurate textual responses to the presentation of

the sight words (Wong et al., 2021). Accuracy was reported in two primary contexts, including the cumulative number of novel sight words acquired during the instruction phase and response maintenance four weeks following the end of instruction. We defined an *accurate textual response* as the participant's vocal production of a word with point-to-point correspondence to the target sight word that was presented on the computer screen. The participant was expected to emit a response within 5 s of the presentation of the sight word in order for the response to be considered correct. An *incorrect response* was defined as any response from the participant that did not have point-to-point correspondence with the target sight word or the absence of a vocal response within 5 s of the presentation of the sight word. The experimenter calculated the percentage of accurate responses after each instructional session and 4 weeks following the initial acquisition of the sight word. The independent variable was the manipulation of acquisition-performance criteria (OA, 100% correct for an individual operant across two observations; SA, 100% correct aggregate responding for a set of operants across two observations), and also measured the cumulative number of acquired operants in both conditions in the study.

Experimental Design and Procedure

The experimenter used an adapted alternating treatments design (Sindelar et al., 1985) to evaluate the effects the operant- and the set-based application of acquisition-performance criteria. After the target identification, target assignment, and three sessions of formal baseline, the intervention phase began. The experimenter taught sight words under both conditions in an alternating and counterbalanced fashion and conducted one session of each condition daily until all sight words were acquired. Response maintenance assessments were conducted 28 days (4 weeks) following the acquisition of each target sight word. The experimenter ensured that the participants did not come into contact with the acquired words during daily instruction during the period between initial acquisition and the maintenance assessments.

The experimenters implemented a number of counterbalancing measures. Experimenters counterbalanced teaching sessions across participants and time of day. For example, the SA condition occurred in the morning and the OA condition occurred in the afternoon for Patrick. During that same day, the OA condition occurred in the morning and the SA condition occurred in the afternoon for Katie. The participant who received the first teaching session alternated each day. Experimenters also counterbalanced the order of sight words in a session, such that across the 20 trials in a session, each block of four trials (1-4, 5-8, etc.) contained one instance of each sight word and a single sight word was not presented on two consecutive trials. Experimenters quasirandomized the distribution of font types across trials in a session.

Target Identification and Baseline

Prior to the onset of the study, the experimenter selected 40 novel sight words to teach the Patrick and Katie, and 24 sight words to teach William. To control for any unintended contact with the target words outside of the study, the experimenter selected words that were at least three levels above the current curriculum for William. Patrick and Katie had already mastered the majority of the Fry Words. Thus, the experimenter conducted an internet search to find words of greater difficulty. The assignment of each word was done in a quasirandomized fashion that was identical to the target identification process used in Wong et al. (2021) and was based on the best practices of equating targets reported by Cariveau et al. (2021). The inclusion criteria for the target sight words included a) four-syllable words (for Patrick and Katie) and one-syllable

words (for William), b) each four-syllable word contained 9-12 letters and each one-syllable word contained four letters, c) no two words that were phonetically or visually similar were presented in the same instructional session, and d) no two words with the same initial letter were presented in the same instructional session. During the quasirandomized assignment of the 40 four-syllable words, we numbered each word from one to 40. The experimenter assigned the odd numbered words into one condition and the even numbered words into the other condition. The experimenter ensured that across both conditions, there was the same or similar number of nine-letter words, 10-letter words, 11-letter words, and 12-letter words. During the quasirandomized assignment of the 24 one-syllable words, the experimenter also numbered each word from one to 24 and assigned the odd numbered words into one condition and the even numbered words into the other condition.

The baseline procedure involved the first author presenting all sight words individually on the PowerPoint® slideshow and collecting data on correct and incorrect responses. During the baseline assessments, the first author sat next to the participant at the desk and opened up the slideshow. The first author presented the sight word on the screen and allowed the participant 5 s to emit a response. After the participant emitted a response or 5 s passed without any response, the first author recorded a correct or incorrect response, continued to the next sight word presentation, and provided no consequences for correct or incorrect responses during the assessment. The order in which the sight words were presented varied across baseline assessment sessions. In order for the sight word to be included in the study, the data from the baseline assessment had to indicate zero correct responses across three consecutive sessions of the sight word presentations. If the participant emitted a correct response at any point during the baseline sessions, the experimenter substituted the known word for another target sight word that met the inclusion criteria. Furthermore, the experimenter ensured that the sight words taught in the study were not incorporated into the daily academic programming the participants received. On average, each baseline session lasted about 2-3 min. After three baseline sessions, the experimenter assigned 20 words into the OA condition and 20 words into the SA condition for Patrick and Katie. The words in each set were counterbalanced across participants. That is, the 10 odd numbered words were assigned to the OA condition for Patrick and the SA condition for Katie. The 10 even numbered words were assigned to the SA condition for Katie and the OA condition for Patrick. The experimenter assigned 12 words to the OA condition and 12 words to the SA condition for William and Zara, which were also counterbalanced (see online supplemental materials).

General Teaching Procedure

The experimenter and a trained instructor delivered teaching trials through learn unit instruction (Albers & Greer, 1991) throughout the teaching phase. The trained instructor was a master's level student-teacher studying ABA. The experimenter replicated the general instructional procedure used to teach sight words described by Wong et al. (2021). This included presenting a sight word as an antecedent, praise contingent on a correct response on an FR1 schedule, and error correction contingent on an incorrect response (modeling correct response, re-presenting the antecedent to allow for an independent opportunity to respond, and withholding praise for corrected-correct responses). The experimenter or the trained instructor presented the correction procedure up to three times before moving on to the next target sight word. There were no prompts utilized during teaching. However, if a participant did not attend to the screen, the instructor

provided a vocal and gestural prompt to facilitate the participants' attending response prior to the delivery of the discriminative stimulus. In both conditions, the experimenter taught four target sight words in each 20-trial session, with five teaching trials per target sight word. Each instructional session lasted from 5 to 15 min. The acquisition-performance criterion was 100% accuracy across two consecutive observations/sessions in both conditions. This criterion was applied at the level of the individual sight word (OA) or at the level of the set of sight words (SA), depending on the condition. The experimenter conducted one session of each condition per day, 3 to 5 days per week.

Set Analysis. The experimenter taught static sets of four target stimuli until the participant responded with 100% accuracy across two consecutive observations. When the participant's responding met the established acquisition-performance criterion, the experimenter introduced four novel stimuli to teach in the set. The teaching process continued until the participant acquired all words in the condition.

Operant Analysis. The experimenter taught dynamic sets of four target stimuli per session. These sets were dynamic because the acquisition-performance criterion was applied to each individual operant. Once a participant's responding met the acquisition-performance criterion for a single sight word, the experimenter replaced that single sight word with a novel word in the next session. There was a constant cycle of new operants being taught to the participant until the participant acquired all words in the condition.

In the OA condition, when there were fewer than four target operants remaining at the end of the comparison, the experimenter presented distractors or previously acquired words that were being assessed under the response maintenance condition. Thus, there were always 20-trial sessions until there were fewer than four operants to learn, at which point sessions contained 15-20 trials.[3]

Response Maintenance

The experimenter measured the accuracy of textual responses to the acquired sight words during response maintenance probe sessions for 4 weeks after each specific sight word was acquired under both OA and SA conditions. The experimenter conducted the response maintenance sessions in the exact manner as Wong et al. (2021). The experimenter noted the specific day that each sight word was acquired under the OA condition in order to assess for maintenance at exactly 28 days postacquisition. Likewise, the experimenter noted the specific day that each set of four sight words was acquired under the SA condition to assess for maintenance at exactly 28 days postacquisition. Like Wong et al., the maintenance sessions were embedded into skill acquisition sessions. More specifically, the words under maintenance were presented within the same session as sight words that had yet to achieve acquisition criterion. Within the embedded sessions, the experimenter presented each word a total of five times. The maintenance-performance criterion was 100% correct responding during the session for each word. That is, response maintenance was calculated on a per-operant basis for both conditions to highlight the maintenance of individual responses and not obfuscate individual behavior by aggregating accuracy across responses and response opportunities. If the participant textually responded to every presentation of the word during the session correctly, the operant was considered maintained. The experimenter did not provide consequences for correct or incorrect responses during the response maintenance sessions.

[3]Note, this did not affect our main analysis for Patrick and Katie given the modification noted.

Modifications

Due to the COVID-19 pandemic and school shut-downs, modifications were made to this study. The original plan was to teach Patrick and Katie 20 total sight words per condition (40 words total) and collect 4-week response maintenance data. However, the school was shut down before Patrick and Katie completed the study. Before the school closed, the experimenter taught both Patrick and Katie 12 sight words per condition that also had 4-week maintenance data. Additional sight words had met the acquisition-performance criterion for each participant; however, to directly compare acquisition and maintenance data of equal sizes, the reporting/analysis is focused on acquisition and maintenance of the first 12 sight words acquired, with all acquisition data reported, nonetheless. This modification affected William's analysis in that the goal was changed from teaching 20 sight words per condition to 12 sight words per condition on an a priori basis.

Interobserver Agreement and Treatment Fidelity

A trained independent observer collected trial-by-trial interobserver agreement data. The experimenter calculated trial-by-trial agreement by dividing the number of agreed trials by the total number of trials (i.e., 20) and multiplying that number by 100 to get a percentage of agreement. The experimenter collected secondary data for 33% of the baseline sessions for the three participants, 47% of the intervention and maintenance sessions for Patrick, 37% of the intervention and maintenance sessions for Katie, and 63% of the intervention and maintenance sessions for William. Interobserver agreement was 100% during all observations.

The process of measuring treatment fidelity data involved a trained independent observer completing a *Teacher Performance of Rate and Accuracy* (TPRA, Ingham & Greer, 1992) form for the implementation of sight word instruction. On the TPRA form, an independent observer assessed the accuracy of each antecedent and consequence delivered by the instructor for each learning trial. The experimenter collected treatment fidelity data for 33% of the baseline sessions and 47%, 37%, and 63% of intervention and maintenance sessions for Patrick, Katie, and William, respectively. Treatment fidelity was calculated by dividing the total number of correct-response deliveries by the total number of responses recorded and multiplying that number by 100 to get a percentage of fidelity. Treatment fidelity was 100% for the participants across all phases of the study.

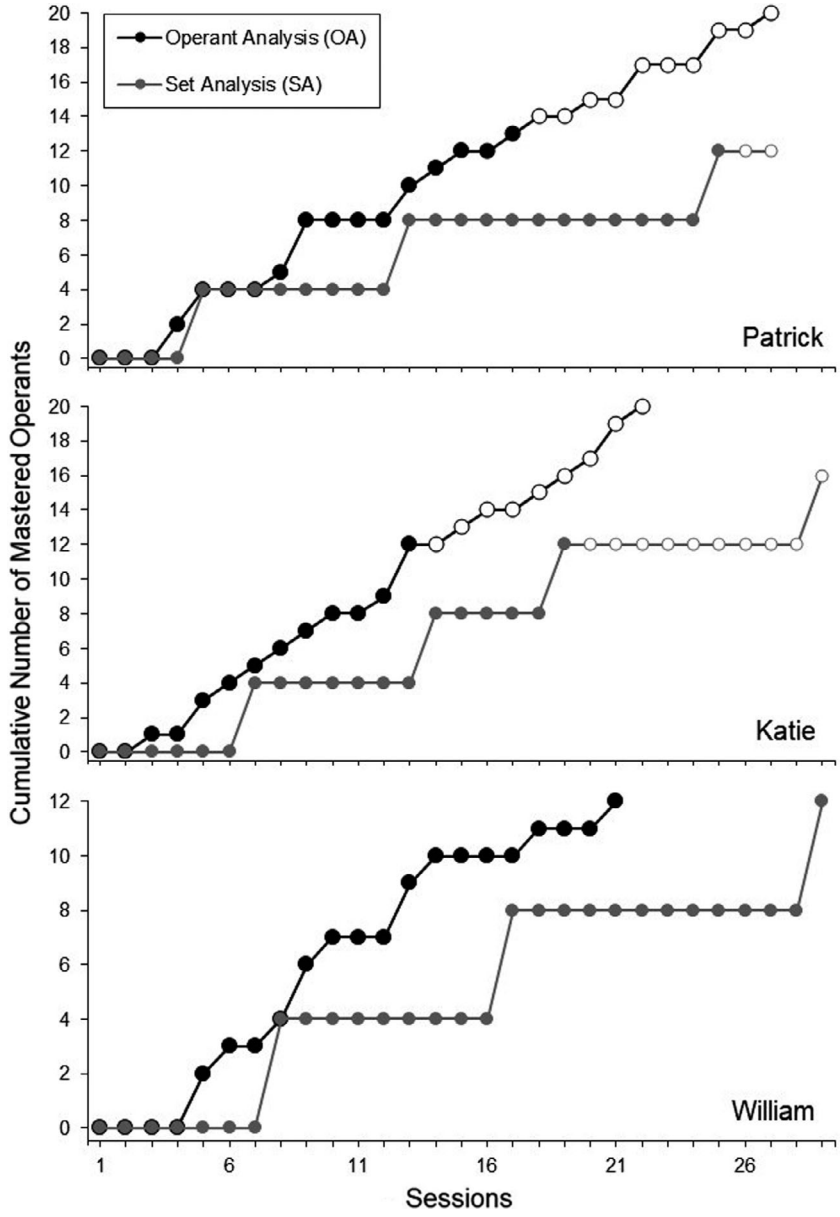
Results

Patrick, Katie, and William emitted zero correct responses to three consecutive sessions of the sight word presentations prior to the start of the intervention.

Figure 1 displays the cumulative number of sight words that Patrick, Katie, and William acquired. Patrick (top panel) achieved the acquisition-performance criterion for 20 sight words in the OA condition and 12 in the SA condition in 27 sessions. When comparing only 12 acquired operants for both conditions, Patrick required potentially 10 additional sessions to acquire 12 operants in the SA condition, or 40% fewer sessions to acquire 12 operants in the OA condition. Katie (middle panel) achieved the acquisition-performance criterion for 20 sight words in the OA condition in 22 sessions and 16 sight words under the SA condition in 29 sessions. When comparing only 12 acquired operants for both conditions, Katie required potentially six additional sessions to acquire 12 operants under the SA condition, or 32% fewer sessions to acquire 12 operants in the OA condition. William (bottom panel) acquired all 12 sight words in 21 sessions in the OA condition and in 29 sessions for the SA condition, or 28% fewer sessions to acquire 12 operants in the OA condition.

Figure 1

The Cumulative Number of Operants Acquired Under Each Condition



Note. The graphs display the cumulative number of operants (sight words) acquired under operant analysis (black circles) and set analysis (gray circles) conditions. The open circles represent the operants beyond the 12 sight words contained in our maintenance analysis.

Supplemental Figures B (Patrick), C (Katie), and D (William) display all acquisition data for the participants per set (SA) or sight word (OA).

Table 1 provides data comparing the acquisition of 12 words in both SA and OA for Patrick, Katie, and William. The experimenter reported the total number of teaching trials to

acquire 12 sight words to conduct an equal comparison between both conditions where there were also available 4-week maintenance data. The experimenter calculated the mean teaching trials to meet the acquisition-performance criterion for the 12 operants in each condition. The table shows that all participants required potentially many more teaching trials to acquire the sight words under the SA condition. For Patrick, there was an 82% increase in the number of teaching trials potentially required to acquire all operants under SA as compared to OA. For Katie, there was a 49% increase in the number of teaching trials required potentially to acquire all operants under SA as compared to OA. For William, there was a 76% increase in the number of teaching trials potentially required to acquire all operants under SA as compared to OA.

Response Maintenance

The experimenter examined the percentage of operants maintained at 4 weeks following the acquisition of 12 sight words per condition for each participant. The maintenance-performance criterion was 100% accuracy across all five presentations of the sight word during a maintenance assessment session. Figure 2 displays the percentage of operants that met the maintenance criterion 4 weeks following the initial acquisition of each sight word. Graphically, we focused on data for the terminal maintenance assessment period because it provides the most conservative representation of operants that were maintained. Patrick (top panel) maintained all 12 (100%) of the operants acquired under both OA and SA conditions. Katie (bottom panel) maintained all 12 (100%) of the operants acquired under the SA condition and 11 out of 12 operants (92%) acquired under the OA condition. William's overall response maintenance data were lower than Patrick and Katie. William maintained eight (67%) of the operants acquired under the OA condition and seven (58%) of the operants acquired under the SA condition. Overall, maintenance data were similar for all three participants across both conditions.

Within-SA Condition Analysis of Overtraining

It was clear that all participants acquired operants quicker in the OA condition, requiring fewer teaching sessions (Figure 1) and trials (Table 1) to acquire textual responses. The response maintenance data also showed that OA was effective in producing similarly accurate responses 4 weeks after the termination of instruction. The experimenter further analyzed the data in the SA condition to examine if there was potentially unnecessary instruction (overtraining trials). Following the conclusion of the study, the experimenters disaggregated data from the SA condition (data displayed in top right graph of Supplemental Figures B, C, and D) by applying the OA acquisition-performance criterion to the first 12 sight

Table 1

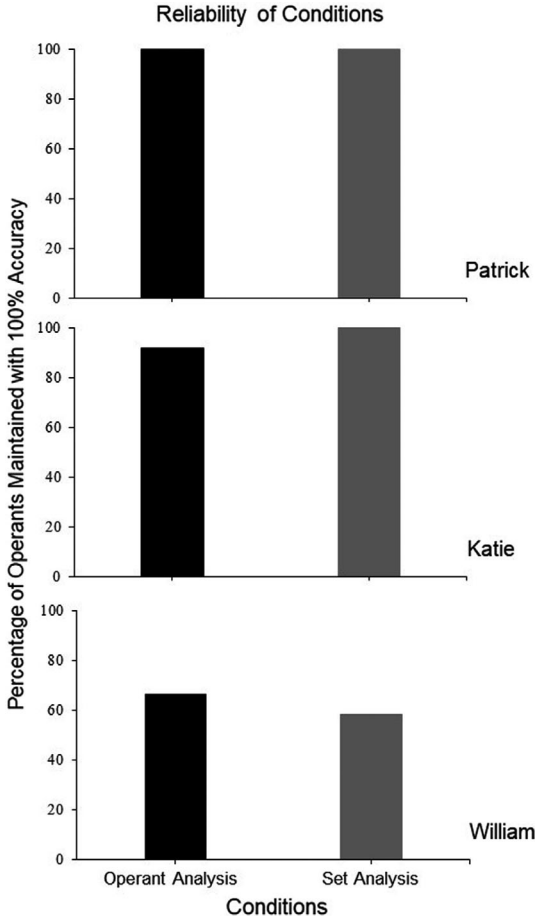
The Total Number of Teaching Trials Potentially Required per Condition

	Patrick		Katie		William	
	OA	SA	OA	SA	OA	SA
Number of Trials	275	500	255	380	330	580
Mean Trials to Criterion per Operant	23	42	21	32	28	48

Note. The total number of teaching trials reported were to acquire 12 target operants for each experimental condition. OA represents Operant Analysis and SA represents Set Analysis.

Figure 2

The Percentage of Operants Maintained at Four Weeks



Note. The graphs display response maintenance results four weeks following the acquisition of sight word operants for each participant. The maintenance-performance criterion was 100% accuracy, thus, the graphs display the percentage of 12 operants that met the maintenance criterion.

words acquired in the SA condition and adding all the teaching trials required potentially to achieve the OA performance criterion for each participant. The experimenters analyzed the SA sight words in this manner to investigate how many trials it would have taken the participants to meet the acquisition-performance criterion for all 12 words, had the sight words been assessed using OA. This process allowed us to calculate how many potentially unnecessary teaching trials were delivered under the SA condition. An example of this analysis can be found in Wong et al. (2021).

Table 2 displays the number of overtraining trials for Patrick, Katie, and William in their respective SA conditions for the first 12 words that met the acquisition-performance criterion. Because sets of four words were taught at one time, there were three overall instructional phases in the design. The experimenters calculated the average number of overtraining trials per operant by dividing the total number of trials by the total number of sight words acquired in the SA condition. To clarify the meaning of each number, if the average was 0, there was no overtraining. Any number higher than 0 indicates the presence of potentially unnecessary teaching. On average, 28% of the trials delivered to Patrick were potentially unnecessary, 30% of the trials delivered to Katie were potentially unnecessary, and 35% of the trials delivered to William were potentially unnecessary during the SA condition. The experimenter delivered an average of 13, 10, and 18 extra trials per operant to Patrick, Katie, and William, respectively.

Discussion

The results replicated and extended the findings of Wong et al. (2021). The outcomes of this study provide further evidence of the benefits of a dynamic, individual-operant application of acquisition-performance criteria while participants "master" new responses. All participants acquired sight words faster in the OA condition. More importantly, when the 100% acquisition-performance criterion was applied across two consecutive sessions (as opposed to one session in Wong et al., 2021), accuracy during maintenance probe sessions was comparable under the OA and SA conditions. This study isolated the effects of units of analysis of acquisition-performance criterion by eliminating the

Table 2

Sessions of Potentially Unnecessary Overtraining Trials during the SA Condition

	Patrick	Katie	William
Phase 1 Overtraining Trials (%)	10 (10.0%)	45 (32.1%)	45 (28.1%)
Phase 2 Overtraining Trials (%)	50 (31.3%)	45 (32.1%)	50 (27.8%)
Phase 3 Overtraining Trials (%)	100 (41.7%)	25 (25.0%)	115 (47.9%)
Total Overtraining Trials (%)	160 (27.7%)	115 (29.7%)	210 (34.6%)
Number of Operants Acquired	12	12	12
Average Overtraining Trials Per Operant	13	10	18

Note. The percentages represent the percent of the total number of trials in each phase that were potentially unnecessary. To calculate average overtraining trials per operant, we divided the total overtraining trials by the number of operants acquired. Each of the three phases represented the opportunity to achieve acquisition-performance criterion for four sight words.

decision protocol included in a previous study and found robust effects of applying performance criteria to individual operants.

Unlike in Wong et al. (2021), response maintenance outcomes were undifferentiated in this study when the acquisition-performance criterion was changed from 100% in one session to 100% across two sessions. Undifferentiated response maintenance suggests that both procedures for learning operants are effective and give rise to examining whether there are efficiency differences between the two techniques. Indeed, participants required potentially many fewer sessions and teaching trials to learn operants in the OA condition. Wong et al. (2021) did not produce reliable response maintenance following a performance criterion of 100% across one session for individual operants. The simple decision to increase the criterion-level frequency (i.e., number of consecutive observations of criterion-level responding, Schneider et al., 2021) from one to two observations appears to have resulted in durable behavior change under the OA condition (reliable, durable behavior change was observed in the SA condition in Wong et al., 2021). ABA practitioners, instructors, and researchers should consider the effects of different performance-criterion frequencies (across one, two, or three observations) at a particular criterion level (e.g., 90% or 100% accuracy) on long-lasting behavior change. To date, only one published study has directly manipulated this component of acquisition-performance criteria (Schneider et al., 2021) and there is a need to better understand this component and its effects on acquisition and maintenance in an OA context.

During the systematic comparison of OA and SA, we did not implement the decision-making protocol in either condition; this was done to isolate the effects of the acquisition-performance criterion unit of analysis manipulation. Additionally, for purposes of research, we did not implement the decision-making protocol to promote equated conditions. Decision-making protocols are a set of rules for selecting, continuing, discontinuing, and modifying treatment based on performance data. Keohane and Greer (2005) studied the implementation of such a decision-making protocol and found that when teachers implemented the protocol, students learned more quickly. Wong et al. (2021) implemented that protocol and, as a result, there were unintended differences in the number of decisions made in each condition. More instructional decisions were made in the OA condition when the protocol was applied based on individual operant data and thus this systematic confound called into question whether the decision-making protocol, OA condition, or some combination of both accounted for differences in the efficiency of learning. Foregoing the decision protocol in

this study helped to isolate the effects of the OA and SA conditions on learning. However, perhaps the number of decisions made should be a secondary dependent variable to be studied in the future. Practically speaking, if there are more decisions made in the OA condition (e.g., Wong et al., 2021), that may serve as an added benefit of the OA procedure because decision points allow instructors to continually assess the potential need to modify teaching tactics if the student is not learning.

There are some limitations that are worthy of discussion. In this study, we compared very stringent performance criteria: 100% performance criterion levels across two sessions (at the individual operant level and the set level). The performance criterion of 100% accuracy across two sessions is not indicative of the most widely used performance criterion across ABA practitioners or ABA researchers (McDougale et al., 2020; Richling et al., 2019); however, at least three empirical studies support such stringent criteria and their positive effects on response maintenance (Longino et al., 2021; Pitts & Hoerger, 2021; Richling et al., 2019). For the participants in this study, the performance criterion used during regular instruction outside of this study was 100% across one session or 90% accuracy across two sessions. Future research should evaluate 100% accuracy across two sessions for individual operants compared to more commonly used set-based performance criteria (e.g., 90% across two sessions) because the evidence thus far seems to support higher levels of accuracy combined with the OA application. Nevertheless, we chose the 100% across two sessions criterion for both conditions in order conduct an equal comparison.

The original focus of this study was to measure and report the results of 40 sight words for Patrick and Katie. The analysis was abruptly stopped, however, due to the COVID-19 pandemic that shut down in-person instruction for the participants, thus creating another limitation of this study. William completed a minority of his sessions in a remote setting, causing some inconsistency with the settings in which the intervention and assessments took place as well. The pandemic caused several issues in Zara's analysis, which led us to report her data separately (see online supplemental materials).

We believe the OA condition—as opposed to the SA condition—better aligns with the core tenets of radical behaviorism's focus on the behavior of the individual (Chiesa, 1994) and questions what constitutes a relevant unit of behavior for analysis. The outcomes suggest that focusing on the separate behaviors of an individual provides a teacher or interventionist a more precise evaluation of those behaviors. A similar argument has been made within the problem behavior literature—that functions of topographies should be analyzed separately (Beavers & Iwata, 2011; Derby et al., 2000). During academic skill acquisition, instructors are often not dealing with response classes. Rather, instructors are trying to bring responding under the control of particular antecedents (e.g., see the word 'cat', say the word "cat"). In this case, the outcomes of this study suggest the appropriate unit of analysis is each antecedent-behavior relation and that aggregating behavior across antecedents leads to less precise information that can extend teaching with no apparent benefit to the student (see outcomes of SA condition). Additional research is needed to examine the conditions under which aggregating an individual's behavior is appropriate. Analyzing behavior at the level of the individual operant provides important additional information to teachers and instructors that allows them to better serve the needs of the student, as was demonstrated in this study by enhanced instructional efficiency of applying acquisition-performance criteria to individual behaviors (OA condition).

A future replication of this study should examine the effects of OA with students of different learning levels. It is possible that students with more advanced verbal repertoires may not

experience as much of a discrepancy in acquisition differences because they may learn quickly under any condition. Students with less sophisticated verbal repertoires may experience even greater differences. It would be interesting to evaluate the potential impact of OA for learners of all levels to examine for whom OA accelerates learning. Another consideration for a future study could be to change the number of target stimuli taught in each set. It is possible that larger stimulus set sizes combined with a strict acquisition-performance criterion could potentially increase the efficiency of the OA condition (Kodak et al., 2020).

The results of this study demonstrate the need to continually examine well-established acquisition-performance criterion procedures in our field and to question whether there are small changes we can make to our instruction that can accelerate learning (Richling et al., 2019). The rules instructors set should be based on scientific evidence rather than traditions passed down from prior practices (see survey outcomes of Richling et al., 2019). Moreover, the rules instructors set during instruction potentially have great effects on learning and should continue to undergo rigorous empirical evaluations.

REFERENCES

Albers, A. E., & Greer, R. D. (1991). Is the three-term contingency trial a predictor of effective instruction? *Journal of Behavioral Education*, *1*(3), 337-354. https://doi.org/10.1007/BF00947188

Beavers, G. A., & Iwata, B. A. (2011). Prevalence of multiply controlled problem behavior. *Journal of Applied Behavior Analysis*, *44*(3), 593-597. https://doi.org/10.1901/jaba.2011.44-593

Cariveau, T., Helvey, C. I., Moseley, T. K., & Hester, J. (2021). Equating and assigning targets in the adapted alternating treatments design: Review of special education journals. *Remedial and Special Education*, *43*(1), 58-71. https://doi.org/10.1177/0741932521996071

Chiesa, M. (1994). *Radical behaviorism: The philosophy and the science*. Authors Cooperative.

Derby, K. M., Hagopian, L., Fisher, W. W., Richman, D., Augustine, M., Fahs, A., & Thompson, R. (2000). Functional analysis of aberrant behavior through measurement of separate response topographies. *Journal of Applied Behavior Analysis*, *33*(1), 113-117. https://doi.org/10.1901/jaba.2000.33-113

Donnelly, K. (2015). *Adolphe Quetelet, social physics and the average men of science, 1796-1874*. Routledge.

Fienup, D. M., & Brodsky, J. (2017). Effects of mastery criterion on the emergence of derived equivalence relations. *Journal of Applied Behavior Analysis*, *50*(4), 843-848. https://doi.org/10.1002/jaba.416

Fienup, D. M., & Carr, J. E. (2021). The use of performance criteria for determining "mastery" in discrete-trial instruction: A call for research. *Behavioral Interventions*, *36*(4), 756-763. https://doi.org/10.1002/bin.1827

Fry, E. (2004). *1000 instant words: The most common words for teaching reading, writing and spelling*. Jossey-Bass.

Fuller, J. L., & Fienup, D. M. (2018). A preliminary analysis of mastery criterion levels: Effects on response maintenance. *Behavior Analysis in Practice*, *11*(4), 1–8. https://doi.org/10.1007/s40617-017-0201-0.

Greer, R. D. (2002). Designing teaching strategies: An applied behavior analysis systems approach. *Elsevier.*

Greer, R. D., Speckman, J., Dudek, J., Cahill, C., Weber, J., Du, L., & Longano, J. (2019). Early learner curriculum and achievement record (ELCAR): A CABAS® developmental inventory. https://www.scienceofteaching.org/product-page/elcar-early-learner-curriculum-and-achievement-record

Ingham, P., & Greer, R. D. (1992). Changes in student and teacher responses in observed and generalized settings as a function of supervisor observations. *Journal of Applied Behavior Analysis*, *25*(1), 153-164. https://doi.org/10.1901/jaba.1992.25-153

Johnston, J. M., & O'Neill, G. (1973). The analysis of performance criteria defining course grades as a determinant of college student academic performance. *Journal of Applied Behavior Analysis*, *6*(2), 261–268. https://doi.org/10.1901/jaba.1973.6-261

Keller, F. S. (1968). "Good-bye, teacher…" *Journal of Applied Behavior Analysis*, *1*(1), 79-89. https://doi.org/10.1901/jaba.1968.1-79

Keohane, D. D., & Greer, R. D. (2005). Teachers' use of a verbally governed algorithm and student learning. *International Journal of Behavioral Consultation and Therapy*, *1*(3), 252-271. https://doi.org/10.1037/h0100749

Kodak, T., Halbur, M., Bergmann, S., Costello, D. R., Benitez, B., Olsen, M., Gorgan, E., & Cliett, T. (2020). A comparison of stimulus set size on tact training for children with autism spectrum disorder. *Journal of Applied Behavior Analysis*, *53*(1), 265-283. https://doi.org/10.1002/jaba.553

Longino, E. B., McDougale, C. M., Richling, S. M., & Palmier, J. (2021). The effects of mastery criteria on maintenance: A replication with most-to-least prompting. *Behavior Analysis in Practice*. Advance online publication. https://doi.org/10.1007/s40617-021-00562-y

Lovaas, O. I. (1987). Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Consulting and Clinical Psychology*, *55*(1), 3-9. https://doi.org/10.1037/0022-006X.55.1.3

McDougale, C., Richling, S. M., Longino, E. B., & O'Rourke, S. A. (2020). Mastery criteria and maintenance: A descriptive analysis of applied research procedures. *Behavior Analysis in Practice*, *13*(2), 402-410. https://doi.org/10.1007/s40617-019-00365-2

Pitts, L., & Hoerger, M. L. (2021). Mastery criteria and the maintenance of skills in children with developmental disabilities. *Behavioral Interventions*, *36*(2), 522-531. https://doi.org/10.1002/bin.1778

Richling, S. M., Fienup, D. M., & Wong, K. (in press). Establishing performance criteria for mastery. In J. L. Matson (Ed.), *Applied behavior analysis: A comprehensive handbook*. Springer Nature.

Richling, S, M., Williams, W. L., & Carr, J. E. (2019). The effects of different mastery criteria on the skill maintenance of children with developmental disabilities. *Journal of Applied Behavior Analysis*, *52*(3), 701-717. https://doi.org/10.1002/jaba.580

Schneider, A., Fienup, D. M., Gussin, R., & Moss, P. (2021). *A preliminary comparison of mastery criterion frequency values: Effects on acquisition and maintenance*. Advance online publication. Behavioral Interventions. https://doi.org/10.1002/bin.1834

Semb, G. (1974). The effects of mastery criteria and assignment length on college-student test performance. *Journal of Applied Behavior Analysis*, *1*(1), 61–69. https://doi.org/10.1901/jaba.1974.7-61

Sindelar, P. T., Rosenberg, M. S., & Wilson, R. J. (1985). An adapted alternating treatments design for instructional research. *Education and Treatment of Children*, *8*(1), 67–76. http://www.jstor.org/stable/42898888

Wong, K. K., Bajwa, T., & Fienup, D. M. (2021). The application of mastery criterion to individual operants and the effects on acquisition and maintenance of responses. *Journal of Behavioral Education*. Advance online publication. https://doi.org/10.1007/s10864-020-09420-3

Received June 8, 2021
Final acceptance February 12, 2022
Action Editor, Jeanne Donaldson

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's website.