Behavior Analysis in Practice (2023) 16:1138–1150
https://doi.org/10.1007/s40617-023-00793-1

RESEARCH ARTICLE

# A Comparison of Stimulus Set Sizes: Systematic Replication with Operant Analysis Acquisition Criteria

Hui Zhi[1] · Daniel M. Fienup[1] · R. Douglas Greer[1] · Sydney S. Henderson[1]

Accepted: 30 March 2023 / Published online: 24 April 2023

## Abstract

We conducted a systematic replication of Kodak et al.'s *Journal of Applied Behavior Analysis, 53*(1), 265–283 (2020) and Vladescu et al.'s *Behavior Analysis in Practice, 14*(1), 193–197 (2021) experiments on the effects of stimulus set sizes on skill acquisition. The researchers manipulated the stimulus set sizes by teaching 3, 6, and 12 sight words simultaneously during learn unit instruction. Researchers taught participants until the participant's responding reached the acquisition criterion for 12 different sight words per set size condition. The acquisition criterion was set for an individual operant, whereby when accuracy met criterion for a single sight word, that sight word was replaced in the following session. The results showed that the set-size-3 was more efficient in producing criterion-level responding during acquisition than the set-size-6, and -12, which was consistent with Vladescu et al.'s findings. However, the set-size-12 reliably produced the highest maintenance levels for all participants. The definition of "effectiveness" based on acquisition or maintenance was discussed.

**Keywords** Learn unit instruction · Stimulus set sizes · Operant analysis acquisition criteria · Skill acquisition

Children with learning disabilities require explicit and intensive academic interventions (National Autism Center, 2015). In special education settings, it is common for teachers to break down long-term objectives into several short-term objectives that drive what single sessions of teaching entail. For example, a long-term objective may be for a child to learn 12 color names at 90% accuracy or higher. From this long-term objective, the teacher may decide on four short-term objectives of learning three specific color names at 90% accuracy or higher. Stimulus set size refers to the number of stimuli (e.g., three color names) that teachers and therapists need to teach simultaneously in an instructional/therapy session (Kodak et al., 2020). Past research on skill acquisition has varied widely regarding the selection of stimulus set sizes, ranging from one to 15 stimuli per session, without providing the rationale for the selection of a stimulus set (Haq et al., 2015; Kodak et al., 2020; Leaf & McEachin, 1999; Lovaas, 2003; Maurice et al., 2001; Yaw et al., 2014). Several curricular manuals for children with disabilities recommended including at least three targets in a stimulus set when learning listener and speaker responses to decrease the probability of the establishment of faulty stimulus control (Greer & Ross, 2008; Greer et al., 2020; Grow & LeBlanc, 2013; LaMarca & LaMarca, 2018; MacDonald & Langer, 2018). However, few researchers have compared the effects of stimulus set size on skill acquisition (Kodak et al., 2020; Vladescu et al., 2021) and research on this variable is important as this variable is included in all skill acquisition studies, regardless of whether one attends to this fact.

Kodak et al. (2020) conducted the first empirical comparison of different stimulus set sizes in the discrete trial instruction (DTI). DTI involves small units of three-term contingencies, including the discriminative stimuli, prompts (as necessary), student responses, contingent consequences, and intertrial intervals (Smith, 2001). Kodak et al. taught 12 tacts in each of four stimulus set size conditions: teaching 3, 4, 6, and 12 stimuli simultaneously until 12 total tacts met the acquisition criterion. Participants included three children (3- to 6-year-old males) and one adolescent (15-year-old male) diagnosed with ASD. Each instructional session consisted of 12 trials, including four presentations per stimulus in the set-size-3 condition, three presentations the set-size-4 condition, two presentations the set-size-6 condition, and one presentation the set-size-12 condition. The acquisition

✉ Daniel M. Fienup
fienup@tc.columbia.edu

[1] Department of Health and Behavior Studies, Teachers College, Columbia University, 525 West 120th Street, Box 223, New York, NY 10027, USA

criterion was set at 100% of trials with independent correct responses across two consecutive sessions. With set sizes of three, four, and six stimuli, meeting criterion with a set signaled moving on to the next set until 12 total tacts met the acquisition criterion. The results showed that all participants required fewer trials and shorter durations to meet the criterion for the tacts assigned to the larger stimulus set sizes (6- and 12- stimuli) compared to the smaller set sizes (3- and 4- stimuli).

Kodak et al. (2020) demonstrated that this overlooked variable affected the speed of acquisition. Nevertheless, the results may have been influenced by the implementation of the same acquisition criterion across different set sizes. For example, participants needed to emit eight consecutive independent correct responses per stimulus to meet criterion in the set-size-three condition (to respond with 100% accuracy across two consecutive sessions), whereas they needed to emit two consecutive independent correct responses per stimulus in the set-size-12 condition. Thus, the number of independent correct responses required to meet the criterion decreased as the stimulus set size increased (Vladescu et al., 2021). In addition, because the researchers did not control the number of presentations per stimulus in each session, the results might be due to the unnecessary trials for the smaller set sizes and in turn failed to provide an accurate measurement on the effectiveness of different stimulus sizes.

Vladescu et al. (2021) conducted a systematic replication of Kodak et al. (2020) with two adolescents with ASD and extended the literature by controlling the number of independent correct responses required to meet criterion across the set-size-3, -6, and -12 conditions during tact training. In other words, the acquisition criterion was condition-specific according to the difference in the number of presentations per stimulus in an instructional session across the three conditions. In particular, the acquisition criteria were set at 100% independent correct responses for one session in the set-size-three condition, for two consecutive sessions in the set-size-6 condition, and for four consecutive sessions in the set-size-12 condition. The results showed that the smaller stimulus set sizes (3- and 6- stimuli) were more efficient than the large set size (12-stimuli) on tact acquisition for both participants, with set size of 3 stimuli produced the fastest tact acquisition for one participant and set size of 6 stimuli for the other. The results contrasted with the findings of Kodak et al.'s regarding the efficiency of the small (3-stimuli) and large (12-stimuli) stimulus set sizes in skill acquisition. Although Vladescu et al. controlled the number of independent correct responses required to mastery across different conditions by using condition-specific acquisition criteria, the design did not control the number of presentations per stimulus within a session, such that the number of presentations per stimulus within a session decreased as the set size increased. For example, each session included four presentations per stimulus in the set-size-3 condition but only one presentation per stimulus in the set-size-12 condition. Thus, the higher number of presentations per stimulus per session might lead to the faster acquisition rate in the set-size-3 and -6 conditions than the -12 condition.

Because the studies of Kodak et al. (2020) and Vladescu et al. (2021) demonstrated different results, the purpose of this research was to conduct a systematic replication of Kodak et al.'s and Vladescu et al.'s with further experimental control on the number of presentations per stimulus in each session on skill acquisition. The researchers manipulated the stimulus set-size-3, -6, and -12 during learn unit instruction,[1] a type of DTI. Each of the three conditions included learning a total of 12 sight words while simultaneously teaching 3 stimuli per session, 6 stimuli per session, and 12 stimuli per session, respectively. Each stimulus was presented three times in each instructional session for session lengths of 9, 18, and 36 trials, respectively.

In addition, we implemented an operant analysis (OA) acquisition criterion (Cordeiro et al., 2022; Kim et al., 2023; Wong et al., 2022; Wong & Fienup, 2022) that targeted the acquisition mastery of individual operants instead of set-based acquisition criteria as used by Kodak et al. (2020) and Vladescu et al. (2021). Most studies focusing on skill acquisition implemented set-based analysis in which researchers randomized blocks of trials that consisted of multiple operants and presentations of each antecedent (Wong et al., 2022) and evaluate acquisition criteria based on the percentage of trials with independent correct responses across all responses emitted in a session (Grow et al., 2011, 2014; Grow & Van der Hijde, 2017). Thus, the instruction efficiency might be affected if a student acquired an operant much slower than the remaining operants in a set (for an explanation, see Wong et al., 2022). For example, in teaching a total of 12 tacts, researchers needed to deliver four stimulus sets in the set-size-3 condition in comparison with one stimulus set in the set-size-12 condition. Because the total number of stimulus sets increased as the set size decreased, the influence of set-based acquisition criterion on the instruction efficiency was more evident for the smaller set sizes than the bigger set size. The use of OA addressed this problem by setting acquisition criterion per operant, independent of the learning rates of other operants in the same set. If the student learned one operant much slower than the other operants in a set, we would substitute novel targets for the mastered operants while maintaining the nonmastered

[1] Learn unit instruction interlocks three-term contingencies for both student and teacher. The student's observing responses are discriminative stimuli (SDs) for a teacher to provide antecedents and the student's responses are SDs for a teacher to deliver contingent consequences (Albers & Greer, 1991).

operant in an instruction session, without interfering the instruction efficiency.

The goal of this research was to further examine the effects of stimulus set size on the acquisition and maintenance of speaker responses, with the use of operant analysis (OA) acquisition criterion (Wong et al., 2022; Wong & Fienup, 2022). In OA, when responding to a single operant meets or exceeds the criterion, that operant is considered "mastered" and is replaced in the following session with a new operant. Because previous studies solely involved participants with ASD, we extended past findings for preschoolers with and without disabilities. The results will help teachers and therapists effectively facilitate students' learning and making scientifically based decisions regarding the learning goals.

## Method

### Participants

Three preschoolers, ranging in age from 3- to 4-years old participated in the study. Two participants were educationally classified as a preschooler with a developmental delay; one participant had no diagnoses. The three participants attended a suburban full-time publicly funded preschool for students with and without disabilities. The classroom staff implemented the Comprehensive Application of Behavior Analysis to Schooling (CABAS) model of instruction. Indie was a 4-year-old Black female with developmental delays, who was able to follow two-step vocal directions with 100% accuracy, echo phrases of up to five words with 100% accuracy, emit independent mands and tacts using full sentences with 100% accuracy, and verbally respond to questions using full sentences with 100% accuracy. She textually responded to the lower case and uppercase letter names and sounds with 70% accuracy at the onset of the study. Indie received behavioral analytic and speech services for the past 2 years. Kevin was a 4-year-old white male with developmental delays. He was able to follow two-step vocal directions with 100% accuracy, echo phrases of up to five words with 100% accuracy, emit independent mands and tacts using full sentences with 100% accuracy, and verbally respond to questions using full sentences with 80% accuracy. Prior to the onset of the study, Kevin textually responded to the lower case and uppercase letter names and sounds with 70% accuracy. He received behavioral analytic services for the past 2 years. Nathan was a 3-year-old white typically developing male, who was able to follow two-step vocal directions, echo phrases of up to five words, emit independent mands and tacts using full sentences, all with 100% accuracy. He verbally responded to questions using full sentences with 70% accuracy. Nathan textually responded to the lower case and uppercase letter names and sounds with 80% accuracy prior to the onset of the study.

The participants were selected according to the following inclusion criteria: (1) was able to sit appropriately in the chair (i.e., bottom in the chair and feet on the ground) without emitting any problem behaviors and attend to the instruction for at least 5 min; (2) had an instructional history of learning listener (e.g., pointing) and speaker responses (e.g., tacts) involving set sizes of three, four, and five stimuli in the learn unit instruction; (3) had mastered prerequisite repertoires for learning sight words such as imitating gross motor actions, orienting to and observing two dimensional stimuli, and echoing with point-to-point accuracy; (4) no sight word instructional history; (5) sight words as learning goals according to the Early Learner Curriculum and Achievement Record: A CABAS Developmental Inventory (Greer et al., 2020); and (6) teachers' praise functioned as a conditioned reinforcer. The participants' repertoires were assessed using the Early Learner Curriculum and Achievement Record: A CABAS Developmental Inventory (Greer et al., 2020) within 14 days at the onset of the study. Detailed information regarding the participants' verbal behavior development is outlined in Table 1.

### Interventionists

The primary researcher was the classroom head teacher who was a second year PhD student in applied behavior analysis, held certification as a board certified behavior analyst, and dual certifications in special and general education. In addition, a first-year master's student in applied behavior analysis who was a teaching assistant in the respective classroom served as an independent observer to collect data for interobserver agreement.

### Setting and Materials

Data were collected in the participant's respective classroom. The classroom consisted of seven students, including two typically developing students and five students with developmental delays, one teacher, and one teaching assistant. The teacher or teaching assistant delivered instruction to non-participant students in 1:1 or small group settings while the researchers conducted the study. We measured all instructional responses during the daily instruction (including this experiment) with frequent interobserver agreement checks and rating of teacher accuracy of implementing instruction using the Teacher Performance Rate and Accuracy (Ingham & Greer, 1992) assessment tool.

The classroom contained a play area, four small group stations, and a large rectangular communal table. The researcher and the participant sat across the head of the communal table during the instruction sessions and the probe

**Table 1** Detailed information of relevant verbal behavior cusps for all participants

| Cusps and Capabilities | Description | Participants | | |
|---|---|---|---|---|
| | | Indie | Kevin | Nathan |
| Bidirectional Naming (BiN) | BiN allows the learning of word-object relations incidentally without the need of direct instruction. BiN has three levels: Pre-UniN, Unidirectional Naming (UniN), Bidirectional Naming (BiN). | Pre-UniN (23% accuracy) | Pre-UniN (20% accuracy) | Pre-UniN (30% accuracy) |
| Generalized Imitation (GI) | GI enables the "see-do" correspondence without the need of direct instruction. | 100% accuarcy for 2-step imitations | 100% accuarcy for 2-step imitations | 100% accuarcy for 2-step imitations |
| Auditory Matching (AM) | AM allows auditory matching to sample (MTS) for speech. | 100% accuracy | 95% accuracy | 97% accuracy |
| Listener Literacy (Phonemic stimulus control) | Listener literacy enables learning from spoken instructions with or without the presence of visual distractors. | 100% accuracy for 2-step vocal directions | 100% accuracy for 2-step vocal directions | 100% accuracy for 2-step vocal directions |
| Independent Mand (IM) | IM allows mediating the environment under the condition of deprivation with verbal operants. | 100% accuracy using full sentences | 100% accuracy using full sentences | 100% accuracy using full sentences |
| Independent Tact (IT) | IT allows recruiting adult social attention under non-verbal stimulus control with verbal operants. | 100% accuracy using full sentences | 100% accuracy using full sentences | 100% accuracy using full sentences |
| Intraverbal | Intraverbal enables verbally responding to questions in academic and social interactions. | 80% accuracy using full sentences | 80% accuracy using full sentences | 70% accuracy using full sentences |
| Autoclitics | Autoclitics enables one to describe, qualify, quantify, manipulate, and relate to tacts and mands without direct instruction. | 100% accuracy using 1 autoclitic | 100% accuracy using 1 autoclitic | 100% accuracy using 1 autoclitic |

Verbal behavior development includes four stages: preverbal, listener, speaker, and joining of listener and speaker (Greer et al., 2017). The level of verbal behavior for all participants is listener and speaker

**Table 2** List of sight words used in the experiment

| Number | Set 1 | Set 2 | Set 3 |
|---|---|---|---|
| 1 | am | egg | my |
| 2 | eat | more | all |
| 3 | find | no | this |
| 4 | they | time | now |
| 5 | us | are | to |
| 6 | toy | first | get |
| 7 | go | one | of |
| 8 | box | use | big |
| 9 | on | book | part |
| 10 | pig | if | sit |
| 11 | said | were | what |
| 12 | when | she | car |

The words were selected randomly from Think Tank Sight Words Flash Cards (Pre-K)

**Table 3** The assignment of the sets of sight words to conditions of set-size-3, -6, and -12 for each participant

| Conditions | Indie | Kevin | Nathan |
|---|---|---|---|
| Set-size-3 | Set 1 | Set 3 | Set 2 |
| Set-size-6 | Set 2 | Set 1 | Set 3 |
| Set-size-12 | Set 3 | Set 2 | Set 1 |

Set-size-three represents the condition with three target words per session, Set-size-6 represents 6 target words per session, and Set-size-12 represents 12 target words per session

sessions for maintenance skills. The researcher used PowerPoint to present the sight words during the pre- and postinstruction probes and during instruction. Each sight word was displayed in three different fonts (Abadi, Calibri Light, Courier) randomized with three different colors (black, blue, red) in 130-point font. A laptop, data collection sheets, a timer, and pens with black ink were used in the procedure. The sets of sight words used in the learning process and probe sessions for each participant are listed in Table 2. The details regarding the assignment of the sets of stimuli to each condition are listed in Table 3.

## Measurement

We measured four dependent variables in the learning of novel sight words for each participant. A correct response in a trial was defined as the participant's tacting the word accurately within 5 s of the presentation of the antecedent visual stimulus. An incorrect response in a trial was defined as the participant's emitting the word inaccurately or not responding within 5 s of the presentation of the word. Considering that the acquisition mastery criterion was set for an individual word (Wong et al., 2022; Wong & Fienup, 2022) and the number of instructional trials per session was different across conditions, we evaluated the participants' performance by calculating the cumulative number of targets mastered with respect to the cumulative trials in an instructional session. The cumulative words mastered in a session was calculated by adding the number of words mastered in the current session with the total number of words mastered in the previous sessions. The cumulative trials delivered in a session was calculated by multiplying the total number of sessions delivered (including the current session) by the respective number of trials per session in each condition (i.e., 9 trials in the set-size-3 condition, 18 trials in the set-size-6 condition, and 36 trials in the set-size-12 condition).

The second dependent variable was the total number of adjusted and unadjusted instructional trials required by the participant to meet criteria for all the 12 words in each condition. When there was no new word to replace the mastered words in a condition, the words mastered in the previous sessions were delivered with nonmastered words in an instructional session. This would happen throughout the instruction in the set-size-12 condition and after the participants had mastered 10 words in the set-size-3 condition and seven words in the set-size-6 condition. Thus, we adjusted the total number of trials required to master 12 words across conditions by excluding the trials delivered to the words that were mastered in the previous sessions.

We also measured the adjusted and unadjusted total duration required by the participant to meet the mastery criterion for the 12 sight words in each of the three conditions. The duration of a session was measured as the time elapsed from the presentation of the first word until the end of the consequence of the last learning opportunity (i.e., the delivery of praise following the correct response and correction following the incorrect response). The unadjusted total duration was calculated by adding the duration of all sessions until the mastery of 12 words. We adjusted the total duration in proportion to the adjusted total trials, such that we divided the adjusted total number of trials by the unadjusted total number of trials and multiplied by the unadjusted total duration. In addition, maintenance probes were conducted to measure the number of correct responses in each condition every 10 days for up to 30 days following the mastery of a word. During maintenance, the correct and incorrect responses were recorded in the same manner as during acquisition.

## Independent Variable

We compared the participants' skill acquisition and maintenance with different stimulus set sizes in the learn unit instruction. The researchers praised correct responses and implemented a correction procedure contingent on incorrect responses following the antecedent instructions. The correction procedure involved the researcher modeling the correct response and re-delivering the antecedent and allowing the participant the opportunity to independently respond two times. Each of the three conditions included a total of 12 sight words with simultaneous teaching of 3 stimuli per session (set-size-3), 6 stimuli per session (set-size-6), and 12 stimuli per session (set-size-12). Each stimulus was presented three times in each instructional session, resulting in 9, 18, and 36 trial sessions per condition, respectively.

## Experimental Design

We used an adapted alternating treatment design (Sindelar et al., 1985) to investigate the effects of stimulus set sizes on skill acquisition and maintenance. The researchers counterbalanced the assignment of stimulus sets to the three conditions across the participants, such that all participants had different stimuli sets in each condition (see Table 3). In addition, the auditory and visual characteristics of the words were controlled in each set, such that none of the words in a set shared the same first syllable, rhyming with each other, or looked similar (Cariveau et al., 2022). We selected 36 novel one-syllable sight words and randomly assigned them to three stimuli sets. We counterbalanced the daily order of instruction conditions across the three participants and within each participant to decrease the possibilities of sequence and carry-over effects through this procedure for individual participants. The participants engaged in a different activity (e.g., free play, coloring, and writing letters) for at least 10 min between any two instructional sessions. One to three instructional sessions for each set size condition was conducted for all participants per day, all before lunch.

## Procedure

The procedure was presented in the following order: (1) Probe the number of correct responses for the three sets of sight words; (2) start the instruction across the three conditions; and (3) probe the number of correct responses to the mastered words in the maintenance assessment every 10 days for up to 30 days.

### Preexperimental Probes to Identify Target Stimuli

We randomly selected 36 words from the Think Tank Sight Words Flash Cards (Pre-K). The Think Tank Sight Words (Pre-K) included 104 flash cards of most common words used in books for preschoolers. Prior to the experiment, we randomly assigned the words to three stimulus sets: if a word shared the same first syllable, rhyming with, or looked similar with another word in a set, we took the flash card of the word out and randomly pick another flash card in the pool until no word in a set shared similar auditory and visual characteristics (see Experimental Design section above). Then we conducted probes to identify the sight words not in the participant's repertoire. The researcher sat across the participant at a rectangular table and presented the words in PowerPoint on a laptop. Each probe session included one presentation of each word for a total of 12 trials. We presented each stimulus to the participant and waited up to 5 s for the participant's response and moved on to next stimulus. No consequences were delivered following correct or incorrect responses. A word was included if no correct response was emitted during this probe. If the participant emitted a correct response to a word, the word was replaced with another word and the replacement word was then tested.

### Baseline

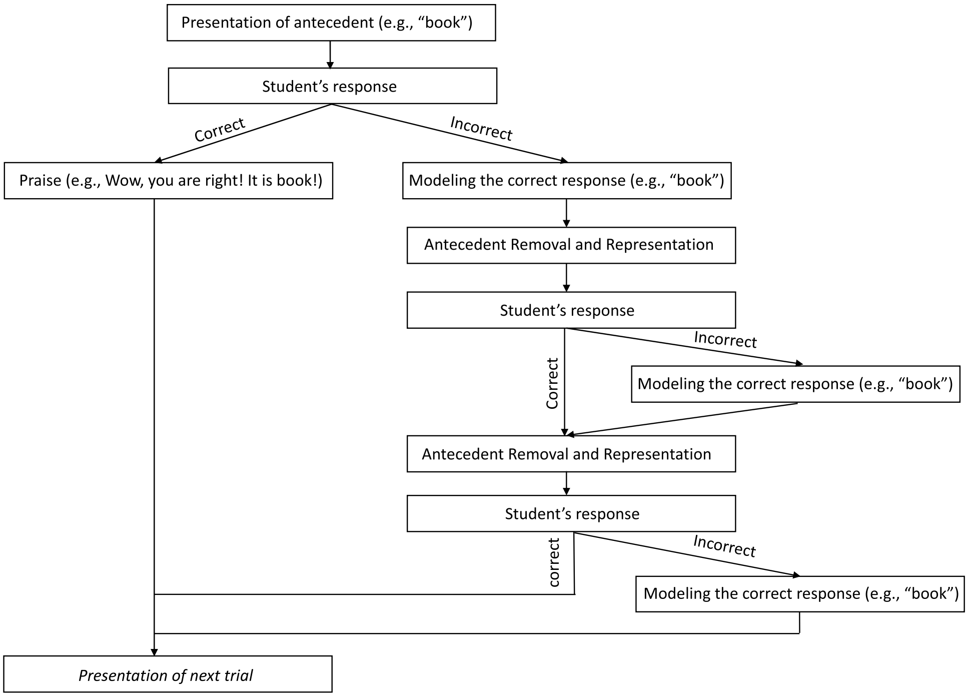
After identifying stimuli that were not in a participant's repertoire and assigning words to sets of stimuli, we conducted three sessions for each stimulus set. The procedure was the same as that of preinstruction probes.

### General Instructional Procedures

The researcher sat across the participant at a rectangular table and presented the sight words on a laptop. We implemented learn unit instruction during acquisition, which consists of three components: presentation of antecedent while the participant is attending, an opportunity for the participant to respond, and the delivery of contingent consequences. The researcher recorded the occurrences of correct and incorrect responses as well as session duration. In each session, three variations of the words that corresponded to the antecedent stimulus (different fonts and colors of a word) were rotated in a random order. The researcher presented antecedent stimuli (e.g., sight word "book"), vocally praised correct responses, and implemented a correction procedure for incorrect responses. For the correction procedure: (1) the researcher modeled the correct response; (2) removed and represented the antecedent; (3) and the participant had an opportunity to independently respond. The correction procedure was presented two times regardless the participant's response. The correct responses (prompted and independent correct responses) emitted by the participant during the correction procedure were unconsequated. The acquisition criterion was set at 100% accuracy for a single word in one session. The detailed instruction procedure is outlined in Fig. 1.

**Set-size-3** Each session consisted of nine trials, including three presentations for each of the three words. If the participant met the acquisition criterion for a word in a session, the researchers replaced this word with a novel word in the

**Fig. 1** The general procedure of the learn unit instruction


Presentation of antecedent (e.g., "book")
Student's response
Correct
Incorrect
Praise (e.g., Wow, you are right! It is book!)
Modeling the correct response (e.g., "book")
Antecedent Removal and Representation
Student's response
Incorrect
Correct
Modeling the correct response (e.g., "book")
Antecedent Removal and Representation
Student's response
Incorrect
correct
Modeling the correct response (e.g., "book")
Presentation of next trial

corresponding word set. We continued this process until the participants mastered the assigned 12 words.

**Set-size-6** This condition was identical to the set-size-3 condition except that each session consisted of 18 trials, including three presentations for each of the six words.

**Set-size-12** This condition was the same as set-size-3 except that each session consisted of 36 trials, including three presentations for each of the 12 words. Because each session simultaneously taught all words in the set, there was no replacement of word that met the acquisition criterion. Rather, we continued teaching all 12 words simultaneously until accuracy on all 12 words met the acquisition criterion.

### Maintenance Probes

The researchers conducted maintenance probes every 10 days for up to 30 days following the mastery of a word—relative to the specific date when a word was mastered. Because we applied acquisition criteria to individual operants, this meant that "sets" of words were not presented as was during baseline. Rather, as an operant met the acquisition criterion, that operant was assessed for maintenance 10 days later and 30 days later from the exact date. The rest of the procedure was identical to baseline. The researchers controlled the exposures to the mastered words for all participants throughout the maintenance probes in the school settings, such that none of the mastered words were presented in the instructional and noninstructional sessions outside of the maintenance probes.

## Interobserver Agreement and Procedural Fidelity

An observer simultaneously and independently collected data on students' responses in a session while the researcher conducted the experiment. The trial-by-trial correspondence was compared at the end of each session. Interobserver agreement was calculated by comparing each individual trial to determine if each trial was scored the same. The researcher divided the number of agreed trials by the total number of trials then multiplied by 100%. The independent observer collected data for 100% of the target identification and baseline sessions, 36.9% of the instructional sessions (35.6% sessions for Indie, 37.7% sessions for Kevin, and 37.8% for Nathan), and 84.9% of maintenance sessions (69% sessions for Indie, 86% sessions for Kevin, and 100% sessions for Nathan), all with a arithmetic mean agreement of 100%.

Procedural integrity was collected and measured using the Teacher Performance Rate and Accuracy form (Ingham & Greer, 1992). A supervisor observed the sessions and collected data on the extent of the researcher's adherence to the experiment procedure in each trial of a session. The supervisor measured the accuracy of three components in each trial: (1) the researcher's presenting antecedents; (2) recording data; and (3) delivering contingent consequences following the participant's responses. A trial was scored as incorrect if the researcher incorrectly implemented one or more components

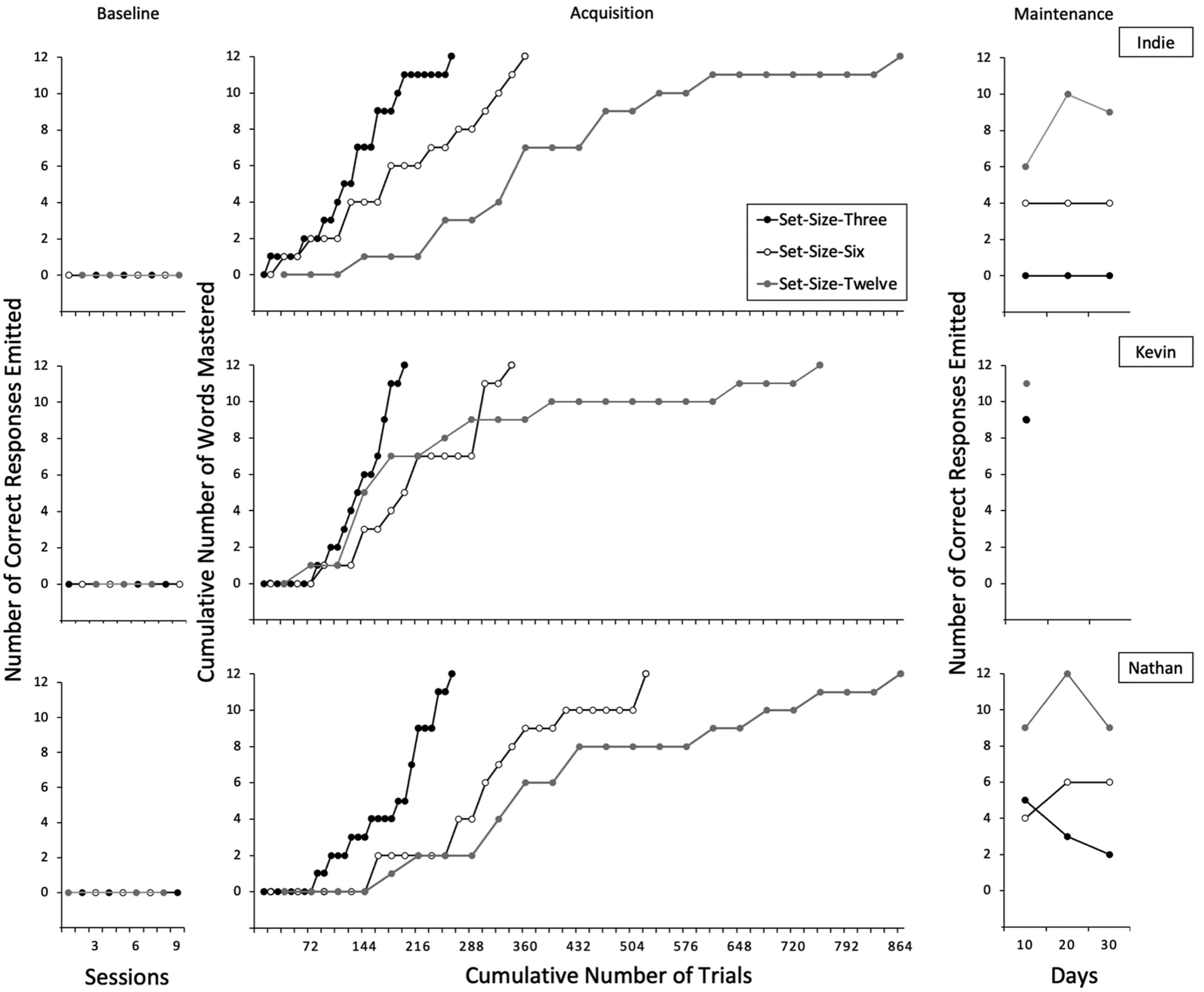


**Fig. 2** The cumulative sight words mastered during the baseline, acquisition, and maintenance for all participants. *Note.* Each interval during the phase of acquisition on the x-axis represents 9 trials. Each dot represents the cumulative number of words mastered per number of trials delivered

in the trial. The researcher calculated procedural integrity by dividing the number of correct trials by the total number of trials and multiplied by 100%. The supervisor collected data for 100% of baseline, 34.1% of instructional sessions (32.9% sessions for Indie, 33.9% sessions for Kevin, and 35.4% for Nathan), and 74.6% of maintenance sessions (45% sessions for Indie, 86% sessions for Kevin, and 100% sessions for Nathan), all with procedural integrity of 100%.

## Results

Prior to teaching tacts to sight words, the researcher conducted the baseline probe for each set of stimuli with each participant. The baseline data remained stable at 0 correct responses for all participants across three sessions (left panels, Fig. 2). The middle panels of Fig. 2 display the cumulative number of sight words mastered by the participants with respect to the number of trials administered during acquisition. The right side of Fig. 2 displays the number of correct responses during maintenance probes. All participants successfully mastered the 12 sight words assigned to the conditions of set-size-3, -6, and -12. During acquisition, Indie met the acquisition criterion (100% accuracy per operant in one session) for all set-size-3 words in 261 trials (top panel, Fig. 2). She required a total of 360 trials (37.8% more trials) in the set-size-6 condition and 804 trials (231.0% more trials) in the set-size-12 condition. Indie responded with 0/12 accuracy in the set-size-3 condition, 4/12 in the set-size-6 condition, and 9/12 in the set-size-12 condition, at 30 days

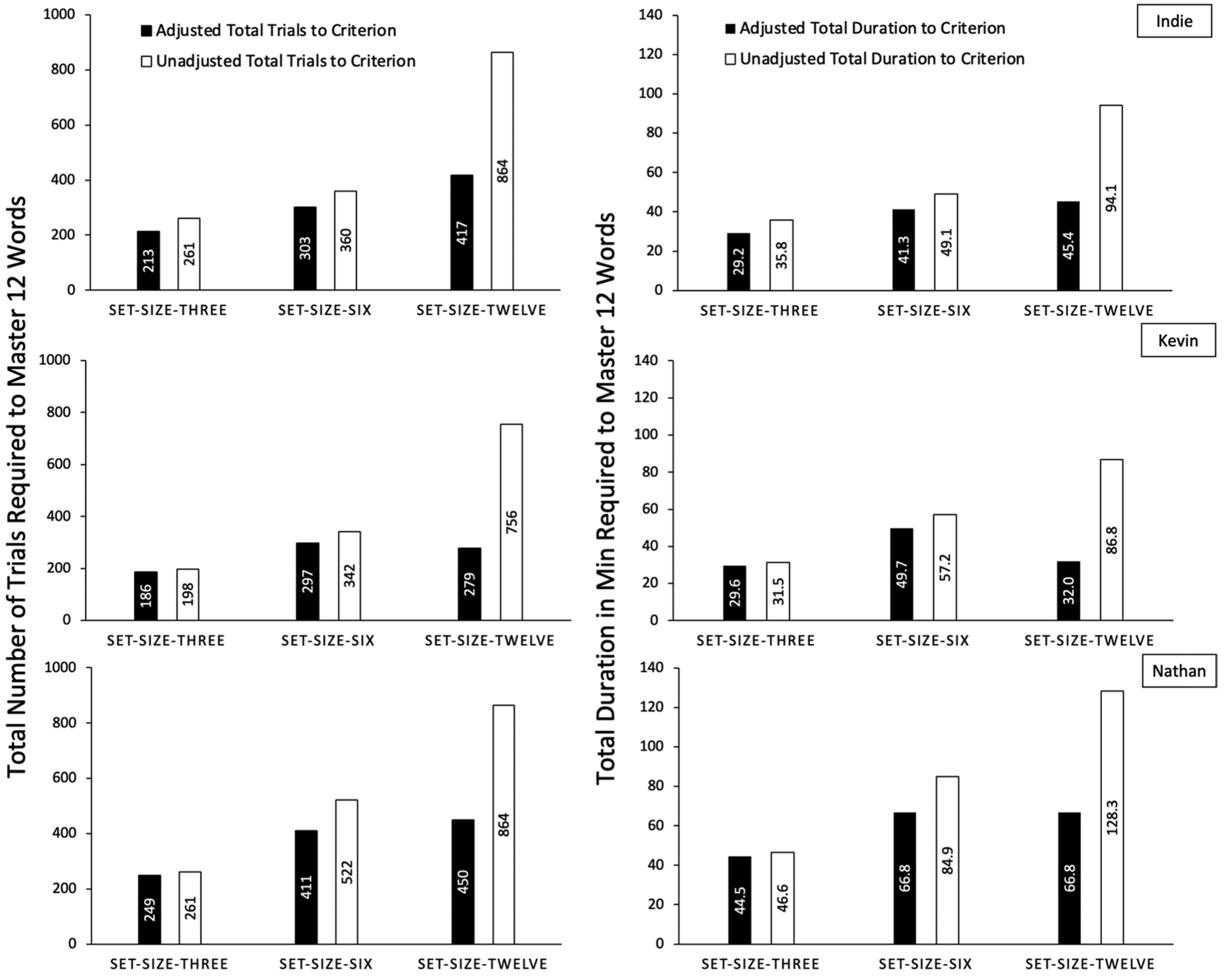


**Fig. 3** Adjusted and Unadjusted Total Number of Trials and Total Duration Required to Meet the Acquisition Criterion. *Note.* The adjusted total number of trials was calculated by excluding the trials delivered to the words that were mastered in the previous sessions. The adjusted total duration was calculated by dividing the adjusted total number of trials by the unadjusted total and multiplying by the unadjusted total duration

following mastery. Kevin met the acquisition criterion for all set-size-3 words in 198 trials. He required a total of 342 trials (72.7% more trials) in the set-size-6 condition and 756 trials (281.8% more trials) in the set-size-12 condition. Due to the impact of school closings due to the COVID-19 pandemic, we only collected maintenance data for up to 10 days for Kevin. He responded with 9/12 accuracy in the set-size-3 and -6 conditions and 12/12 accuracy in the set-size-12 condition. Nathan met the acquisition criterion for all set-size-3 words in 261 trials. He required a total of 522 trials (100.0% more trials) in the set-size-6 condition and 864 trials (231.0% more trials) in the set-size-12 condition. During maintenance, Nathan responded with 2/12 accuracy in the set-size-3 condition, 6/12 in the set-size-6 condition, and 9/12 in the set-size-12 condition, at 30 days following mastery. Overall, all participants required fewer trials to master all words in the set-size-3 condition than in the set-size-6 and -12 conditions. However, the highest maintenance outcomes were reliably observed during the set size 12 condition for all participants.

Figure 3 shows the adjusted and unadjusted total number of instructional trials and total minutes required until mastery during acquisition for all participants. As mentioned above, adjustments were made that removed trials and time allocated to previously mastered sight words during the acquisition phase. Prior to adjustment, all participants required the fewest trials and minutes to master the sight words assigned to the set-size-3 condition followed by the set-size-6 and -12 conditions. After adjustment, the total number of trials and total duration required to master the

12 sight words decreased across all conditions for all participants, especially for the set-size-12 condition followed by the set-size-6 and -3 conditions. Nevertheless, the set-size-3 condition consistently produced the fastest acquisition of 12 sight words for all participants, prior to and post adjustment. Except for Kevin, all participants required fewer trials and the same or fewer minutes until mastery in the set-size-6 condition than the set-size-12 condition after adjusting the total trials and duration. Post adjustment, Indie met the acquisition criterion for 12 sight words with 213 trials and 29.2 min in the set-size-3 condition (top panels, Fig. 3). She required 90 additional trials (42.3% more trials) and 12.1 more min (41.4% more min) in the set-size-6 condition and 204 additional trials (95.8% more trials) and 16.2 more min (55.5% more min) in the set-size-12 condition (top left panel, Fig. 3). Kevin met the acquisition criterion for 12 sight words with 186 trials and 29.6 min in the set-size-3 condition (middle panels, Fig. 3). He required 111 additional trials (59.7% more trials) and 20.1 more min (67.9% more min) in the set-size-6 condition and 93 additional trials (50.0% more trials) and 2.4 more min (8.1% more min) in the set-size-12 condition. Unlike in Fig. 2, after adjustment Kevin met the acquisition criterion with 18 fewer trials (6.1% fewer trials) and 17.7 fewer min (35.6% fewer min) in the set-size-12 condition than in the set-size-6 condition. Nathan met the acquisition criterion for 12 sight words with 249 trials and 44.5 min in the set-size-3 condition (bottom panels, Fig. 3). He required 162 additional trials (65.1% more trials) in the set-size-6 condition and 201 additional trials (80.7% more trials) in the set-size-12 condition. Nathan required 22.3 more min (50.1% more min) in the set-size-6 and -12 conditions. Overall, after adjustment all participants learned the targets with an average of 121 more trials (56.0% more trials) and 18.2 more min (52.8% more min) in the set-size-6 condition and with 166 more trials (76.9% more trials) and 13.6 more min (39.6% more min) in the set-size-12 condition than in the set-size-3 condition (*M* total trials = 216, *M* total duration = 34.4 min).

## Discussion

This study systematically replicated the set size manipulations conducted by Kodak et al. (2020) and Vladescu et al. (2021) research with operant analysis (OA) acquisition criterion (Cordeiro et al., 2022; Wong et al., 2022; Wong & Fienup, 2022). The results showed that the set-size-3 condition was consistently more efficient than the set-size-12 in terms of acquiring new operants prior to and post the adjustment of the instructional trials and duration, which is consistent with Vladescu et al.'s findings. Based on the acquisition data alone, our results contrasted with Kodak et al.'s findings that the larger set-size-twelve was more efficient than the small set-size-three in skill acquisition. Furthermore, whereas Kodak et al. found no differences on the number of independent correct responses across different set sizes at 2 weeks following the mastery, our results showed that the set-size-12 condition reliably produced the best maintenance levels throughout the maintenance probes for all participants. Indie and Nathan responded with an average of 75% accuracy in the set-size-12 condition at 30 days following the instruction, followed by an average of 42% and 8% accuracy in the set-size-6 and -3 conditions, respectively. Kevin responded with 92% accuracy in the set-size-12 condition and 75% accuracy in the set-size-6 and -3 conditions at 10 days following mastery. The opposing acquisition and maintenance results for small set-size (e.g., three stimuli) and large set-size (e.g., 12 stimuli) in current study suggest one might choose set sizes based on different instruction goals. Although further verification needed, the results of this study suggest that if the goal is on the rate of skill acquisition, the small set size (e.g., 3 stimuli) might be more efficacious (Ghaemmaghami et al., 2021); if the goal is on the level of skill maintenance, the larger set size (e.g., 12 stimuli) might be more effective.

Several curricular manuals on skill acquisition for children with disabilities recommended including at least three stimuli in a stimulus set on learning listener and speaker responses (Greer & Ross, 2008; Greer et al., 2020; Grow & LeBlanc, 2013; LaMarca & LaMarca, 2018; MacDonald & Langer, 2018). The involvement of multiple stimuli in a set may help promote the development of stimulus control by the essential characteristics of a stimulus. The results in this study extended past findings and identified that the set-size-3 was more efficient than the larger set size-6 and -12 on learning speaker responses. However, the maintenance outcomes question whether efficiency of initial acquisition should be the comparison researchers make (Wolery et al., 1991). A core goal of applied behavior analysis is to produce durable behavior change (Baer et al., 1968) and the data reported in this study demonstrated that efficacy and efficiency during acquisition phases did not necessarily correlate with the durability of responding (maintenance assessments). Similar outcomes have been demonstrated in studied evaluating acquisition performance criteria. For example, Fuller and Fienup (2018) taught children academic responses until performance was 50%, 80%, or 90% accurate or higher and found (1) children took longer to meet acquisition criterion at higher levels; and (2) only higher acquisition criteria were associated with high levels of behavior 3 to 4 weeks after teaching. Thus, we suggest researchers and clinicians focus on terminal goals (i.e., maintenance, generalization) for determining efficacy and only comparing efficiency for those procedures that produce intended outcomes.

Most researchers focusing on skill acquisition administered set-based acquisition criterion, such that the acquisition mastery was evaluated by the percentage of independent correct responses across all responding opportunities to the operants in a blocked set (Grow et al., 2011, 2014; Grow & Van der Hijde, 2017). Kodak et al. (2020) administered a consistent set-based acquisition criterion across different set sizes. Vladescu et al. (2021) further equated the number of independent correct responses required until mastery across different set sizes through the implementation of condition-specific set-based acquisition criteria. Nevertheless, the mastery for acquisition with set-based analysis relied on the acquisition rate of the slowest operant in a set (Wong et al., 2022). Controlling for the total stimuli to acquire, the implementation of set-based acquisition criterion might affect the instruction efficiency. We controlled the influences on the instruction efficiency from the operants in a set with the use of OA acquisition criterion (Wong et al., 2022; Wong & Fienup, 2022) that targeted the acquisition mastery of individual operant.

We also extended past research by equating the number of presentations per stimulus in a session across different conditions with OA acquisition criterion. Kodak et al.'s (2020) and Vladescu et al.'s (2021) studies involved different numbers of presentations per stimulus in an instructional session across different set sizes (e.g., four presentations per stimulus in the set-size-3 condition and one prestation per stimulus in the set-size-12 condition), in which the number of presentations per stimulus decreased as the stimulus set size increased. The higher efficiency of the bigger stimulus set sizes might be due to the more trials for the mastered words in the smaller set sizes. Such overestimation on the total number of trials required to master the smaller set sizes might be more evident in Kodak et al.'s study because of the implementation of the same set-based acquisition criterion across different set sizes (see explanation in the previous paragraph). The fastest acquisition rate of set-size-3 for all participants in current study confirmed the aforementioned concerns.

This experiment was limited in a few aspects. First, due to the COVID-19 pandemic, we only collected maintenance data for up to 10 days for Kevin. It would help verify the higher levels of correct responding during maintenance in the set-size-6 and -12 conditions than the set-size-3 condition with more data. Second, we did not replicate the procedures with novel sets of sight words for participants (intrasubject replication). Although all participants acquired the sight words in the set-size-3 condition with fewer number of trials and shorter duration than in the set-size-6 and -12 conditions (intersubject replication), it would help increase the reliability of the results with designs that allows a within stimulus replication (e.g., a concurrent multiple baseline design across stimulus sets). Because we used a static set of 12 words in each condition, the participants were exposed to all the 12 words throughout the instruction in the set-size-12 condition because there was no word to replace the mastered words. Thus, the higher maintenance level in the set-size-12 condition may be a result of the more overlearning trials (i.e., additional practice of a skill after reaching the acquisition mastery criterion; see Dougherty & Johnston, 1996) to the 12 words in the instruction. That is, during the set-size-3 and -6 conditions, once a word was mastered, it was removed from teaching until there was no new word to replace the mastered word; however, with the set-size-12 it continued to be part of instruction until all 12 words were mastered. Continued analysis with further control on the overlearning trials per stimulus across conditions (e.g., continue to replace the mastered words with new words in all conditions) would help verify the higher maintenance levels in the size of 12 condition.

The outcomes of current study suggest several areas for future research. Because previous and current studies focused on the effects of stimulus set sizes for speaker responses (e.g., tact), the generality of the results to multiple instruction programs requiring different responding topographies (e.g., listener responses) needed further verification. Future research may also replicate the procedure with more participants with different ages and language development levels. Considering the influence of the more overlearning trials per stimulus in the set-size-twelve condition on the maintenance levels, future researchers may involve dynamic sets of stimuli across different set sizes with OA acquisition criteria, such that the mastered stimulus will be replaced with a novel stimulus until the participant masters all the predetermined stimuli in a set. The results would help verify the higher maintenance levels in the set-size-12 condition. In addition, the results reported herein suggested measuring the "effectiveness" of different instruction procedures according to the corresponding effects on the maintenance levels rather than the acquisition rates as commonly used in the research on skill acquisition. Although all participants acquired the sight words fastest in the set-size-3 condition, the maintenance level was the lowest (0% accuracy for Indie and 17% accuracy for Nathan) in the set-size-3 condition at 30 days following the mastery. In contrast, although all participants required more trials and longer duration to master the targets in the set-size-`1 condition than the set-size-3, they demonstrated much higher levels of accuracy in the maintenance probes: 75% accuracy at 30 days following mastery for Indie and Nathan; 92% accuracy at 10 days for Kevin. Considering the opposite results between the skill acquisition and maintenance for small and large stimulus set sizes, we suggested future research to further discuss the definition of effectiveness and efficiency in learning novel skills (Wolery et al., 1991). Although further research is needed to clarify aforementioned concerns, the results reported herein clearly demonstrated the higher efficiency of set-size-3 than the set-size-6 and -12 in skill acquisition. These results provide valuable information for instructors regarding the arrangement of instruction antecedents and may in turn effectively facilitate students' learning in diverse educational settings.

**Author Note** This manuscript was submitted in partial fulfillment of a doctoral degree in Applied Behavior Analysis at Teachers College, Columbia University by the first author under the mentorship of the second author. We thank Dr. Jessica Dudek, Matthew Zajic, and Dr. Tom Cariveau for comments on an earlier version of this manuscript.

**Data Availability** Data can be obtained by contacting the corresponding author.

## Declarations

**Ethical Consent** A parent of each participant consented to the dissemination of research.

**Conflicts of Interest** The authors declare no conflict of interests.

## 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://www.jstor.org/stable/41823986

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

Cariveau, T., Helvey, C. I., Moseley, T. K., & Hester, J. (2022). Equating and assigning targets in the adapted allternating treatments design: Review of special education journals. *Remedial & Special Education, 43*(1), 58–71. https://doi.org/10.1177/0741932521996071

Cordeiro, M. C., Kodak, T., Reidy, K., Stopplewor th, A., Zelinski, K., & Jainga, A. (2022). Comparison of mastery criterion applied to individual targets and stimulus sets on acquisition of tacts, intraverbals, and listener responses. *Journal of Applied Behavior Analysis, 55*(4), 1258–1279. https://doi.org/10.1002/jaba.946

Dougherty, K. M., & Johnston, J. M. (1996). Overlearning, fluency, and automaticity. *The Behavior Analyst, 19*(2), 289–292. https://doi.org/10.1007/BF03393171

Fuller, J. L., & Fienup, D. M. (2018). A preliminary analysis of mastery criterion level: Effects on response maintenance. *Behavior Analysis in Practice, 11*, 1–8.

Ghaemmaghami, M., Hanley, G. P., & Jessel, J. (2021). Functional communication training: From efficacy to effectiveness. *Journal of Applied Behavior Analysis, 54*(1), 122–143. https://doi.org/10.1002/jaba.762

Greer, R. D., & Ross, D. E. (2008). *Verbal behavior analysis: Inducing and expanding new verbal capabilities in children with language delays*. Allyn & Bacon.

Greer, R. D., Pohl, P., Du, L., & Moschella, J. L. (2017). The separate development of children's listener and speaker behavior and the intercept as behavioral metamorphosis. *Journal of Behavioral & Brain Science, 7*(13), 674–704. https://doi.org/10.4236/jbbs.2017.713045

Greer, R. D.,Speckman, J., Singer-Dudek, J., Weber, J., Cahill, C, Du, L. & Longano, J. (2020). *Early Learner Curriculum and Achievement Record (ELCAR): A CABAS® developmental inventory*. CABAS® & the Fred S. Keller School.

Grow, L., & LeBlanc, L. (2013). Teaching receptive language skills: Recommendations for instructors. *Behavior Analysis in Practice, 6*(1), 56–75. https://doi.org/10.1007/BF03391791

Grow, L. L., Carr, J. E., Kodak, T. M., Jostad, C. M., & Kisamore, A. N. (2011). A comparison of methods for teaching receptive labeling to children with autism spectrum disorders. *Journal of Applied Behavior Analysis, 44*(3), 475–498. https://doi.org/10.1901/jaba.2011.44-475

Grow, L. L., Kodak, T., & Carr, J. E. (2014). A comparison of methods for teaching receptive labeling to children with autism spectrum disorders: A systematic replication. *Journal of Applied Behavior Analysis, 47*(3), 600–605. https://doi.org/10.1002/jaba.141

Grow, L. L., & Van Der Hijde, R. (2017). A comparison of procedures for teaching receptive labeling of sight words to a child with autism spectrum disorder. *Behavior Analysis in Practice, 10*(1), 62–66. https://doi.org/10.1007/s40617-016-0133-0

Haq, S. S., Kodak, T., Kurtz-Nelson, E., Porritt, M., Rush, K., & Cariveau, T. (2015). Comparing the effects of massed and distributed practice on skill acquisition for children with autism. *Journal of Applied Behavior Analysis, 48*(2), 454–459. https://doi.org/10.1002/jaba.213

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

Kim, J. Y., Fienup, D. M., Draus, C. J., & Wong, K. K. (2023). Differential mastery criteria impact sight word acquisition and maintenance: Application to individual operants and teaching trial doses. *Journal of Applied Behavior Analysis, 56*, 388–399. https://doi.org/10.1002/jaba.970

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

LaMarca, V., & LaMarca, J. (2018). Designing receptive language programs: Pushing the boundaries of research and practice. *Behavior Analysis in Practice, 11*(4), 479–495. https://doi.org/10.1007/s40617-018-0208-1

Leaf, R., & McEachin, J. A. (1999). *A work in progress: Behavior management strategies and a curriculum for intensive behavioral treatment of autism*. DRL Books.

Lovaas, O. I. (2003). *Teaching individuals with developmental delays: Basic intervention techniques*. PRO-ED.

MacDonald, R., & Langer, S. (2018). *Teaching essential discrimination skills to children with autism: A practical guide for parents and educators*. Woodbine House.

Maurice, C., Green, G., & Foxx, R. M. (2001). *Making a difference: Behavioral intervention for autism*. New York: PRO-ED.

National Autism Center. (2015). *Findings and conclusions: National standards project, phase 2.*

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

Smith, T. (2001). Discrete trial training in the treatment of autism. *Focus on Autism and Other Developmental Disabilities, 16*(2), 86–92. https://doi.org/10.1177/108835760101600204

Vladescu, J. C., Gureghian, D., Goodwyn, L., & Campanaro, A. M. (2021). Comparing skill acquisition under different stimulus set sizes with adolescents with autism spectrum disorder: A replication. *Behavior Analysis in Practice, 14*(1), 193–197. https://doi.org/10.1007/s40617-020-00506-y

Wolery, M., Doyle, P. M., Ault, M. J., Gast, D. L., Meyer, S., & Stinson, D. (1991). Effects of presenting incidental information in consequent events on future learning. *Journal of Behavioral Education, 1*(1), 79–104. http://www.jstor.org/stable/41823972

Wong, K. K., & Fienup, D. M. (2022). Units of analysis in acquisition-performance criteria for "mastery": A systematic replication.

*Journal of Applied Behavior Analysis, 55*(3), 971–985. https://doi.org/10.1002/jaba.915

Wong, K. K., Bajwa, T., & Fienup, D. M. (2022). The application of mastery criterion to individual operants and the effects on acquisition and maintenance of responses. *Journal of Behavioral Education, 31*(3), 461–483. https://doi.org/10.1007/s10864-020-09420-3

Yaw, J., Skinner, C. H., Delisle, J., Skinner, A. L., Maurer, K., Cihak, D., Wilhoit, B., & Booher, J. (2014). Measurement scale influences in the evaluation of sightword reading interventions. *Journal of Applied Behavior Analysis, 47*(2), 360–379. https://doi.org/10.1002/jaba.126

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.