Experimental and Descriptive Analyses of Mastery Criteria

Kristina K. Wong

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2021

Abstract

Experimental and Descriptive Analyses of Mastery Criterion

Kristina Wong

An acquisition criterion, more commonly known as "mastery criterion" is an instructor-established standard of performance that may signal the acquisition of a novel skill or the conclusion of a phase of intervention. When teaching new behaviors, researchers and practitioners in the field of Applied Behavior Analysis (ABA) apply some type of criterion for the learner to achieve. The purpose of the following studies was to evaluate the effects of acquisition criteria on skill acquisition in addition to other components of mastery such as response maintenance and stimulus generalization. In Experiment I (Wong et al., 2021), I conducted a systematic comparison of two applications of acquisition criteria. I selected four participants to teach 40 novel sight words using learn unit instruction. The participants were between the ages of 5 and 7 years old and diagnosed with developmental disabilities. I equated the target operants and quasi-randomly assigned 20 sight words in one acquisition criterion condition and 20 sight words in another acquisition criterion condition. In one condition, Set Analysis (SA), the acquisition criterion was applied to a set of four operants. The other condition, Operant Analysis (OA), applied acquisition criterion to individual operants. The level of accuracy and the replication of the accuracy remained the same across conditions, and more specifically, a 100% accuracy across one replication session was utilized under OA and SA. The results of skill acquisition showed that all four participants learned a greater number of sight words under the OA condition compared to the SA condition within the same time frame. Response maintenance results suggested that SA produced more durable responses for three out

of four participants. In Experiment II, I extended the findings of Experiment I by addressing some limitations and systematically replicating the procedures. I increased the number of replications of the acquisition criterion from 1 replication to two replications. I selected four new participants and taught them sight words under the OA application of acquisition criterion and the SA application of acquisition criterion. Similar to the findings of Experiment 1, the skill acquisition results showed all participants learned a greater number of operants under OA compared to SA. The response maintenance results showed that all four participants responded with 100% accuracy to a similar or higher percentage of operants under the OA condition compared to the SA condition, suggesting that the added replication to the acquisition criterion may have improved the durability of responses during four-week follow-up sessions. The findings of both Experiment I and Experiment II contributed to the small but growing body of literature demonstrating the parametric effects of acquisition criteria. However, small sample sizes in the existing acquisition criteria research limit the external validity of the findings. Thus, I conducted a descriptive analysis of every skill acquisition article published in 2017 to 2019 in three peer-reviewed behavioral journals, in order to address this limitation. I reported the general characteristics of over 200 articles targeting skill acquisition. Additionally, our analysis targeted the effects of acquisition-criterion levels and frequency of replications on response maintenance results and generalization results. Ultimately, the results provide evidence that acquisition criteria play an important role in the mastery of novel behaviors, which have practical implications for ABA clinicians and researchers.

Keywords: mastery, acquisition criteria, operant analysis, set analysis, response maintenance, stimulus generalization, frequency of replications, skill acquisition.

Table of Contents

List of Charts, Graphs, Illustrations

Acknowledgements

First and foremost, I am incredibly humbled by the grace of God and I thank Him for giving me this life that I lead. I spent so much of my life trying to take control and plan out my future, but I am thankful that God had bigger and better plans for me.

It is difficult to express how deeply grateful I am to all the people who have supported me throughout this experience. The following words will not do justice to the gratitude that I feel. I would like to thank the members of my dissertation committee who provided incredibly insightful and helpful comments regarding my research. Dr. Greer, you are truly an inspiration, and I am so blessed that I was accepted into the CABAS® program in 2016. The groundbreaking work that you've done for the program, for our kiddos, and for the field of ABA will forever inspire me to make the world a better place – one learn unit at a time. Dr. Dudek, thank you for always pushing me to understand the science of verbal behavior and the principles of behavior that set the foundation for everything we do. Your guidance in class and at work has helped me become more analytical and I've become a better student, teacher, and behavior scientist because of you. Dr. Richling, your passion and excitement for our line of research criteria gave me life! Thank you for shining a light on the importance of mastery criteria.

Dr. Fienup – It is safe to say that I would not have been able to come this far without your unwavering support and guidance… Or should I say, your support and guidance were *crucial* in helping me graduate ☺. You taught me everything I know about conducting quality research and writing publication-worthy papers. I am honored to have been an advisee of yours. There were many times throughout this program that I felt inadequate and incapable of succeeding but your encouragement kept me going and you've given me the tools to spread my wings and fly. I will never forget all your words of reassurance and optimism. You helped me

turn my mistakes into valuable learning lessons and made sure I celebrated even the tiniest of wins. You believed in me when I didn't believe in myself and for that I am forever grateful. Thank you for continually pushing me outside of my comfort zone and providing me with opportunities to pursue my passion in conducting research.

I am so thankful to my CABAS® mentors/ supervisors, Becca, Alie, Ellie, Elizabeth, Reggie, and Jessica. You all shaped me to become the best teacher I can be, not only with all those TPRAs but with your patience, compassion, and your knowledge. Thank you for being the best role models. I look up to you all every single day.

One of my favorite aspects of my CABAS® experience was working with my incredible mentees, Sarah, Tanya, Rachel, Laura, and Kyla. I was inspired by your love for our kiddos, your curiosity about the science of behavior, and your work ethics. It was so fun working alongside each and every one you. You ladies made me work my hardest to become a better mentor. Thank you for taking on any and every difficult task that was thrown your way and most importantly, thank you for helping me collect my data with fidelity!

Elizabeth and Tanya – I never liked the idea of mixing my personal life with my professional life but somehow, after several years, you have become two of the most special people in my life. I feel so lucky and undeserving of your friendship. Thank you for all the early morning workouts and the late-night venting sessions. I would not have made it to this point without you two.

To mom and dad – You two are the true heroes in my life and I am forever thankful for your unconditional and selfless love. Mom, since even before I was born, you put my needs above yours. You have never hesitated to drop anything for me, and you have always been there for me no matter what… that includes everything from driving 4 hours round-trip to pick up my

teddy bear that I left in a hotel, to spending several hours outside in 110 °F weather or below freezing weather watching me play golf. Furthermore, I would not have been able to survive the past several years without all the containers of your home-cooked meals that you delivered to me. I would not have been able to graduate without your "Magi Meals." Thank you for all the countless sacrifices you've made to help me pursue my dreams. Dad, you are the smartest, most brilliant doctor I know. Your charisma is felt by everyone when you walk into a room and I have always aspired to be just like you. I'm honored to have followed in your footsteps as a Columbia graduate. Although it wasn't medical school, it is pretty awesome to be Dr. Wong Jr ☺. Thank you for teaching me to never settle for mediocrity and for always pushing me to reach my fullest potential. Thank you for all the sacrifices you have made for me and for your unwavering confidence in my abilities to achieve anything that I want to do. I am the woman I am today because of you.

Andrew – you unknowingly set out on this wild journey with me after that one fateful night in East Village. Thank you for being my rock. God knew that I needed your steadfast love and your level-headedness in my life. Thank you for drying the many tears that were shed, the 2:00 AM pep-talks, the much-needed bear hugs to calm me down, and the practical advice in moments when I may have catastrophized things (just a little). Thank you for reading my papers, for your enthusiasm about my research even though you had no idea what I was rambling on about, and for your project management skills in order to help me meet the deadlines. If I didn't have that nifty project timeline you made me on our whiteboard, I would still probably be working on writing the introduction of this dissertation. More importantly, thank you for your excitement and joy in the little things. Your big smiles always make me stop and think about just how blessed I am to be loved by you. I thank God every night for bringing you to me.

Dedication

To all the kiddos that I have had the honor of working with these past several years and to all the kiddos that I will work with in the future. I promise to do everything I can to make the learning process as fun and as effective as possible!

Chapter 1: Scientific Foundations of Applied Behavior Analysis

Behavior analysis consists of three main branches, Behaviorism, Experimental Analysis of Behavior (EAB), and Applied Behavior Analysis (ABA). Behaviorism is the philosophy of the science of behavior, EAB is concerned with basic science, and ABA focuses on the development of technologies for improving socially significant behavior (Cooper et al. 2020). The field of ABA burgeoned in the late 1960s and early 1970s with the introduction of effective new tactics based on behavioral principles that improved a variety of behaviors ranging from gross motor skills (Johnson et al., 1966) to verbal behavior (Brigham & Sherman, 1968) to academic performance (Hall et al., 1968). Over the decades, researchers solidified ABA as a scientific field because great care was taken to identify functional relations between environmental events and behavior change. Researchers systematically applied interventions based on principles of behavior to drastically improve the quality of life for individuals. These scientific processes have taken place in large part due to the seven dimensions of ABA outlined by Baer and colleagues in 1968 (Baer et al. 1968), which provided a template for ABA interventions to follow.

ABA treatments are data-driven and built on the foundation of high-quality research. There is a plethora of compelling between-subject and within-subject research to support the effectiveness of interventions in ABA for improving behaviors of individuals diagnosed with intellectual disabilities and developmental disorders such as Autism Spectrum Disorder (ASD) (Landa, 2018; Lovaas 1987; National Autism Center, 2015; Rogers & Vismara, 2008). The education system in the 1960s was lacking scientific technology and started to improve with Skinner's application of a technology of teaching (Skinner, 1968). Two decades later, the

education system was still in a state of crisis despite the gradual increase in behavioral interventions and scientific teaching strategies (Greer, 1996). According to Greer (1991), the science of pedagogy needed to prioritize three major aspects in schooling, which included more opportunities for a learner to respond, increased learner responses, and increased data collection and systematic measurement of student performances. He stated, "The teacher who applies the existing science of teaching behaves more as an applied scientist than as a traditional teacher" (Greer, 1991, p. 29). Lovaas (1987) added major contributions to scientific teaching with his introduction of Discrete Trial Instruction (DTI), a technology of teaching that involves small three-term contingency units of instruction between an instructor and a learner (Smith, 2001). This was later expanded upon by Greer and McDonough (1999) with the learn unit. The learn unit is a fundamental measure of teaching, which involves interlocking operants (three-term contingencies) between an instructor and a learner (Greer & McDonough, 1991). There is a clear and defined antecedent presented by the instructor and the learner has the opportunity to respond within 5 s of the antecedent presentation. Immediately after the learner responds, the instructor delivers a differential consequence that is contingent on a correct or incorrect learner response. The consequence can be a form of reinforcement for correct responses or an error correction procedure for incorrect responses. The learn unit has been a strong predictor of effectual teaching. Since its introduction, student learning in a wide variety of domains such as academics, self-management, and verbal behavior have increased (Greer, 2002). Interventions such as the implementation of learn unit instruction are typically evaluated and replicated using single-subject experimental design, or within-subject designs, in order to analyze functional relationships between independent and dependent variables – how dependent variables change as a function of the systematic manipulation of independent variables. As a result of the field's

dedication to evidence-based science and careful investigations of socially significant behavior, the majority of ABA treatments fulfill the *analytic, behavioral,* and *effective* dimensions of ABA.

I emphasize the importance of the scientific properties that are the hallmark of ABA research because the field is clearly grounded in scientific methodology. However, there are components of the practice that are based on conventional wisdom and traditions, rather than scientific evidence. One such procedure is the application of mastery criteria in ABA teaching programs and interventions (see survey reported in Richling et al., 2019). Mastery criteria is widely used in ABA research and are considered a ubiquitous part of ABA programming (Fuller & Fienup, 2018; McDougale et al., 2020; Richling et al., 2019; Wong et al., 2021).

Since the beginnings of ABA research, researchers have referenced mastery criteria. In 1968, Hall and colleagues referenced a satisfactory rate of performance and when this rate was achieved, the instructor moved to a different phase of the study (Hall et al., 1968). Thirty years later, the report of steady-state mastery criteria in articles published in the *Journal of Applied Behavior Analysis* was about 20% (Sayrs & Ghezzi, 1997) and 46% in the *Journal of Experimental Analysis of Behavior* (Rehfeldt & Ghezzi, 1996). Fast forward another 20 years and the trend of reporting specific, percentage-based performance criteria, or mastery criteria, dramatically increased in ABA. Love et al. (2009) and Richling et al. (2019) conducted surveys of special education teachers and board-certified clinicians, respectively, which showed that almost all ABA practitioners adopt some form of mastery criteria in the administration of their treatments.

Conceptual Explanations of Mastery Criteria

The function of skill acquisition interventions in ABA is for learners to acquire new socially significant behaviors that the learners have in their repertoire long after an intervention is over. Instructors utilize a mastery criterion to signal the sufficient performance of learned behavior (i.e., behavior is mastered), as well as the conclusion of the teaching intervention. Mastery criteria can be defined as "the degree to which a response must be emitted accurately before it is considered acquired or mastered" (Greer & Ross, 2008, p. 296). Sufficient response strength and stimulus control allow newly acquired behavior to persist in the face of disrupters such as extinction periods and novel situations (Craig et al., 2014). This aligns with Behavioral Momentum Theory (BMT), which suggests greater behavioral mass will lead to behavior that is more resistant to change despite the influence of disrupting external factors (Nevin, 1992) Furthermore, the existing literature suggest that rich schedules of reinforcement build greater behavioral mass (Nevin et al., 1983) and when there is a comparison of two different schedules of reinforcement, the behavior that was reinforced in the thicker schedule of reinforcement was most resistant to change (Nevin & Wacker, 2013).

Potential disrupters of behavior are any events that alter a dimension of behavior that is being measured (Craig et al., 2014). In the case of mastery, disrupters may include response maintenance assessments (extinction experiences,) and stimulus generalization assessments (the introduction of new settings, stimuli, or people). Criteria used in skill acquisition interventions play a role in producing persistent behaviors because the acquisition criteria are proxies for response strength. When an instructor utilizes a high level of criteria, such as 100% accuracy, that means that 100% of the presented antecedents evoke the target behavior during training, thus showing a high response strength. In contrast, if the instructor utilizes a 50% accuracy criterion, only 50% of the antecedents evoke the target behavior and the learner contacts fewer instances of

reinforcement during training, thus creating a response strength that is much weaker than the previous example.

The Use of Mastery Criteria in ABA

Greer and Ross (2008) referenced a standard mastery criterion of 90% correct responding for two consecutive sessions or 100% correct responding for one session. Mastery criteria can also be defined as an instructor-determined discriminative stimulus or a specific requirement that signals an instructor to change teaching tactics (Fuller & Fienup, 2018; Richling et al., in press). More specifically, when teaching novel skills, mastery criteria reflect the guidelines for learner performance that is sufficient for the instructor to either stop teaching because the skill is deemed acquired or to implement a different level of prompting to facilitate skill acquisition.

On the surface, establishing a criterion for mastery may seem like a simple task. Conventional wisdom in education affirms that simply establishing 90%-100% correct responding for excellent performances, 80%-89% correct responding for good performances, and 70%-79% correct responding for satisfactory performance is adequate for measuring the mastery of target skills (NAEP, 2009; Schneider & Hutt, 2014). Within the field of special education ABA research, the most widely used dimension of mastery criterion also happens to be an aggregated level of accuracy for a set of teaching trials. Practitioners and researchers typically report the level of accuracy in the form of a percentage of correct responses across all teaching trials within an instructional session. Another dimension of mastery criteria that commonly coincides with the level of accuracy is the frequency of replications at which the performance reaches the predetermined level of accuracy (Fuller & Fienup, 2018). For example, an instructor may require a learner to respond with 80% accuracy for anywhere from one to three consecutive sessions (i.e., replications). According to the survey responses of approximately 200 Board

Certified Behavior Analysts (BCBAs), the most commonly reported mastery criterion was 80% accuracy across three consecutive sessions (Richling et al., 2019). Interestingly, as shown in Greer and Ross's (2008) utilization of mastery criteria, there is a difference between both of the criterion standards – one being 100% accuracy across one replication, and the other being 90% accuracy across two replications. The accuracy level and the frequency of replications differ, but both criteria standards are considered equivalent, indicating that the degree of impact between both dimensions of mastery criteria are equal.

The use of mastery criteria was prevalent throughout education, special education, sports, and organizational behavior management (OBM) (Richling et al., in press). The subsequent section reviews historical literature within the context of education where mastery criteria was applied and reported.

Review of Mastery Criteria in ABA Literature

General Education

In K-12 general education contexts, school-wide implementations of Positive Behavioral Interventions and Supports (PBIS) establish evidence-based and behavioral models to measure student performance in both academic and social domains. There are three tiers of intervention where students are placed. Among the core elements of each tier are data collection for decision making (such as identification of mastered skills, fading supports, or modifying tactics) and individualized instruction when data suggest the established criterion is or is not met. To determine the appropriate tiers for each student, curriculum-based measurement (CBM) or screening-based assessments are used, and the data are used to identify student deficits (Fuchs & Deno, 1991; Jimerson et al., 2016; Shinn, 1989). Data collected in the assessments are used to

compare the students' performance to state or national norms (Ardoin et al., 2005). The assessments focus on response accuracy as well as fluency.

Reading and math fluency criteria measure accurate responses within a given time frame. For students in grades 1-2, the recommended reading mastery rate is about 60 correct words read per minute with zero errors (Good, Simmons, & Kame'enui, 2001; Shapiro, 1996). Performance goals related to fluency are established by measuring "weekly growth" (Fuchs et al., 2003). Instructors set a goal of weekly improvement for students and the data collected allow instructors to make modifications in their teaching techniques if needed.

Higher Education

Research on mastery criteria among college-age students suggest that higher levels of mastery criteria lead to greater accuracy in academic performance in generalization and maintenance tests. One of the first studies on the effects of mastery criteria among college students was conducted by Johnson and O'Neill (1973). The college course was based on Keller's Personalized System of Instruction (PSI) and consisted of several small units that were self-paced (Keller, 1968). Students were required to achieve a predetermined mastery criterion on quizzes at the end of each small unit. The experimenters examined three different levels of mastery criteria applied to unit quizzes and the results showed that students performed much better when a higher mastery criterion was applied to the quizzes.

Similar results were demonstrated in subsequent mastery criteria studies among college courses that also used a PSI methodology (Carlson & Minke, 1975; Semb, 1974). Semb (1974) examined several variations of mastery criterion level and assignment length among college students taking an introductory child-development course. Students went through a series of experimental conditions that combined high or low mastery criterion with short or long

assignments. After the 6-week course, the results demonstrated that the high-criterion, short-assignment condition led to more accurate student performance in generalization and maintenance tests. These results supported the need for educators to establish a high mastery criterion level if they want student performance to have lasting accuracy.

Carlson and Minke (1975) expanded upon the mastery criteria research by evaluating fixed and ascending criterion levels. Based on previous research, there was already support for high levels of mastery criterion (Johnson & O'Neill, 1973; Semb, 1974). Carlson and Minke (1975) added an additional variable to the evaluation of mastery level instead of solely comparing static levels of mastery criterion (80% accuracy vs. 90% accuracy). There were three conditions in which the authors implemented successive approximations of mastery criterion levels on unit quizzes that progressively increased to the terminal mastery criterion level of 90% accuracy in one condition. The other two conditions were fixed mastery criterion levels of 80% accuracy or 90% accuracy. The results of the study showed that the 80% fixed mastery criterion level condition actually produced the best results in student performance. Reiser et al. (1986) found that while the most stringent mastery criterion level of 90% had some positive effects on quiz performance, there were no significant differences in the final examination results between the 70% criterion condition, the 80% criterion condition, or the 90% criterion condition. These results show some dispute to the findings of Johnson and O'Neill (1973) and Semb (1974) since the condition with the highest mastery criterion level did not produce the best results. Given this discrepancy, it is clear that more research is needed not only on different mastery criterion levels, but on different dimensions of mastery criterion that produces the best student performance on tests of generalization and the strongest stimulus control on tests of maintenance.

Different dimensions of mastery criteria and dependent variable effects were investigated by Fienup and Brodsky (2017). The authors compared a rolling versus a block mastery criterion as well as different stringency levels. The block mastery criterion condition was akin to how typical DTI is conducted, where the criterion is applied to a fixed number of trials and mastery can only be demonstrated after the end of each fixed trial session. In contrast, the rolling block mastery criterion condition could potentially allow students to achieve criterion quicker because students can demonstrate mastery at any trial. The mastery conditions were six consecutively correct responses in a row, 12 consecutively correct responses in a row and 100% accuracy in a 12-block trial session. These dimensions of mastery criterion were evaluated with college students who were learning neuroanatomy equivalence classes. The results of the study suggested that there was not much difference between rolling and blocked forms of mastery criterion. The results also showed that stringency of mastery criterion affected student performance in developing equivalence classes. The high stringency of mastery criterion condition produced significantly better performance compared to the low stringency condition and the rolling condition led to a greater percentage of students who passed all derived relation classes.

Special Education

ABA treatments for students with intellectual, learning, or developmental disabilities utilize mastery criteria to determine when an intervention can conclude. Overall, there are consistencies in mastery criterion practices that are widely used in ABA treatments among practitioners and researchers. However, some differences do exist. Richling et al. (2019) administered a survey to approximately 200 Board Certified Behavior Analysts to acquire valuable information on the mastery criterion they applied in their clinical practices. The survey

results indicated that the most widely used criterion to determine mastery was 80% accuracy across one or more sessions. Surprisingly, the results also suggest that justifications for the clinicians' selection of mastery criterion included employer policies and processes that were passed down from supervisors, not from scientific research. These results warrant further investigation because those who practice in field of ABA should implement procedures that are data-driven and based on science rather than tradition.

Following the results of the survey, Richling and colleagues (2019) compared the effects of an 80% accuracy criterion across three sessions, a 60% accuracy criterion across three sessions, and a 100% accuracy criterion on skill maintenance. The results of the study suggest that an 80% mastery criterion did not lead to response maintenance levels that were at or above 80% for any of the four participants in the study. In fact, two of the four participants responded at less than 60% accuracy during weekly follow-up probe sessions. The mastery criterion of 100% accuracy was the only level that showed response maintenance accuracy at or above 80% accuracy after weekly follow-up probe sessions.

McDougale et al. (2019) conducted a descriptive analysis to gather information on mastery criteria reported in three major behavioral journals between the years of 2015 and 2017. The authors then compared the mastery criterion practices with the responses in Richling et al. (2019). The majority of both clinicians and researchers used a level of accuracy within a session to report mastery criterion. However, the specific level of accuracy was disputed. While clinicians favored an 80% correct responding mastery criterion across three consecutive sessions, a 90% correct responding across two sessions was more commonly used by researchers. Another finding from the descriptive analysis was, less than half of the studies reported the delivery of maintenance probe sessions after the intervention was completed. The failure to report response

maintenance means that it is difficult to determine whether or not the mastery criterion used in ABA treatments affect accuracy of responses in the long-term. Thus, it is crucial for future research in mastery criteria to include assessments of response maintenance as a primary measure.

Fuller and Fienup (2018) included response maintenance as one of the primary dependent variables. The authors systematically evaluated three levels of mastery criterion (50% accuracy across one session, 80% accuracy across one session, and 90% accuracy across one session) and the effects on correct spelling responses during a skill acquisition phase and a maintenance phase. Only the 90% accuracy criterion reliably produced higher accuracy during maintenance probe sessions. The results suggest that higher levels of accuracy lead to more durable student performance, but much more research is needed to create a stronger evidence base.

Complexities of Mastery Criteria

A percentage of accuracy within a session across a particular number of replications appears to be a sufficient way of establishing mastery criteria. As identified in the previous section, many studies in a variety of different domains have reported a percentage of accuracy to identify mastery throughout the history of ABA. To this day, the majority of ABA practitioners and researchers utilize this application of mastery criteria to teach novel skills in their everyday practices. However, this cursory analysis of mastery criteria is problematic because it does not take into consideration all the nuances of skill acquisition and *mastery*. A mastery criterion of 90% may be adequate for someone who is learning how to spell but certainly not adequate for someone who is learning to stop at a crosswalk of a busy intersection. A 90% mastery criterion is not adequate for a baker following a task analysis to make a cake from scratch either. It may also be argued that depending on the level of verbal behavior of each individual learner, a session-

based criterion such as 80% or 90% accuracy across one session is not necessary. For an individual who demonstrates bi-directional naming (Greer & Ross, 2008; Greer & Longano, 2010; Miguel, 2016) and is learning letter sounds, a criterion of five correct responses in a row may be acceptable. For a typically developing adult who is learning how to schedule a video conferencing call on Zoom, the skill may be considered acquired if the individual independently schedules three consecutive meetings in a row.

The research on mastery criterion selection is scarce. To date, there are only a couple of researchers who have systematically evaluated different mastery criteria and the effects of skill acquisition and maintenance (Fuller & Fienup, 2018; Richling et al., 2019). The results of their studies suggest that only higher levels of accuracy (90% and 100%) are sufficient in predicting accurate responses during response maintenance sessions that were conducted 3- to 4-weeks following the termination of the teaching phases. The data provided evidence that the most commonly reported mastery criteria in the field of ABA were not effective in producing accurate retention. This violates one of the seven core dimensions of ABA, *effective*. Specifically, effective application of behavioral interventions should improve the target behavior to a practical degree (Cooper et al., 2020). As a scientific field, there needs to be a greater push for experimental evaluations of even our most commonly used procedures, such as establishing mastery criteria.

Furthermore, given the complex nature of mastery criteria identification, it is surprising that the majority of ABA practitioners and ABA researchers rely solely on reporting a set-based percentage of correct responses for mastery criteria. The survey responses of Richling et al. (2019) also indicated that this particular dimension of mastery criteria was chosen as a result of employer policies and directives that were passed down by supervisors. This breach of scientific

practice calls for two changes in the ABA field: a) greater emphasis on the systematic evaluation for identifying mastery criteria and b) more information on the use of mastery criteria by ABA practitioners and researchers in order to advance the experimental analyses of this topic.

A Need for the Scientific Evaluation of Mastery Criteria

The first step in scientifically evaluating and identifying mastery criteria begins with operationally defining what *mastery* means. In our everyday vernacular, mastery refers to an individual's "possession or display of great skill or technique" (Merriam-Webster, n.d.) However, as described earlier, mastery criterion in ABA is defined as a specific guideline to determine when a skill is sufficient for an instructor to either stop teaching or to implement changes in the teaching procedure. It seems questionable to identify a "mastery criterion" of anything less than 100% accuracy to describe a "mastered" skill. For example, an 80% mastery criterion for washing the dishes would certainly not identify the individual as a master dish washer. Furthermore, mastery criterion in ABA does not always identify an acquired skill, but rather acceptable performance to suggest the need to move to a less restrictive form of teaching (prompt-fading).

The second step involves collecting more data and conducting more analyses of mastery criterion practices within our field. Broad surveys and descriptive analyses on mastery criterion (Love et al., 2009; McDougale et al., 2019; Richling et al., 2019) paved the way for more research to be conducted on how ABA practitioners and researchers differentiate mastery criteria based on learner population or skill. Building on our knowledge of mastery criterion practices allows us to carefully examine all dimensions of mastery criteria.

There are clear complexities surrounding the identification of mastery criteria and much more research needed in this area. Thus, it is crucial to advance the experimental research on this

topic. The purpose of my research is twofold. Experiments 1 and 2 systematically evaluate the unit of analysis within mastery criterion on the effects of skill acquisition and response maintenance. I compared two conditions, a set-based analysis of mastery criterion and an individual operant-based analysis of mastery criterion to investigate whether one condition produced quicker skill acquisition and more durable response maintenance over the other condition. In comparing these two conditions, I aimed to address questions regarding the efficacy of different units of mastery criteria. In Experiment 3, I reported a descriptive analysis of mastery criteria applied by ABA researchers on the effects of mastery criteria – also known as acquisition criteria on different components of skill mastery such as response maintenance and stimulus generalization. The objective is to gain more information on current criteria practices to set the foundation for more thorough evaluations of acquisition-criteria in our educational and clinical practices. The data suggest a need to continue analyzing information about the practices involving acquisition criteria in order to facilitate more systematic investigations in the future.

Chapter 2: The Application of Mastery Criterion to Individual Operants and the Effects on Acquisition and Maintenance of Responses

Wong, Bajwa, and Fienup (2021)

(reprinted in this dissertation document)

Journal of Behavioral Education
https://doi.org/10.1007/s10864-020-09420-3

ORIGINAL PAPER

The Application of Mastery Criterion to Individual Operants and the Effects on Acquisition and Maintenance of Responses

Kristina K. Wong[1] · Tanya Bajwa[1] · Daniel M. Fienup[1]

Accepted: 23 October 2020

Abstract

Instructors assess the acquisition of new skills by delivering blocks of trials containing multiple operants. Mastery is evaluated as percentage correct across all operants in the block. The purpose of the current study was to investigate this traditional mastery criterion arrangement compared to an analysis of mastery at the level of individual operants. In both conditions, mastery criterion was 100% accuracy in one session. In the Set Analysis (SA) condition, accuracy was evaluated as average correct responding across all 4 target operants, or sight words, in a set. In the Operant Analysis (OA) condition, we taught 4 sight words simultaneously, assessed accuracy per sight word, and substituted new sight words into the set each time a single sight word was mastered. Overall, all 4 participants learned textual responses to sight words quicker in the OA condition, the reliability of maintenance was similar across conditions for 2 of 4 participants, and 4 of 4 participants maintained a higher or same number of responses from the OA condition compared to the SA condition. Implications for skill acquisition are discussed.

Keywords Mastery criterion · Operant analysis · Response maintenance · Set analysis · Stimulus set size

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10864-020-09420-3) contains supplementary material, which is available to authorized users.

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

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

Published online: 02 January 2021

Introduction

Children with developmental disabilities rely heavily on direct and deliberate instruction to acquire novel functional skills. Applied Behavior Analysis (ABA) has emerged as an evidence-based treatment for teaching novel skills to children with developmental disabilities, such as Autism Spectrum Disorder (Lovaas 1987; National Autism Center 2009, 2015; Rosenwasser and Axelrod 2001). One technology of ABA instruction that is commonly used in educational settings is Discrete Trial Instruction (DTI). DTI follows a structured 3-term contingency framework[1] that includes instructor-delivered discriminative stimuli, prompts (as necessary), opportunities for the student to respond, and instructor-delivered consequences. This process is repeated until the student's performance meets or exceeds a pre-determined performance criterion, or mastery criterion. In the last several years, there has been a flurry of research examining how different components of DTI affect skill acquisition for children with disabilities (e.g., error correction procedures, Carroll et al. 2015; distribution of learning opportunities, Haq and Kodak 2015; effects of stimulus set size, Kodak et al. 2019; prompting procedures, Schnell et al. 2019).

An under-researched, but ubiquitous component of DTI is mastery criterion (Fuller and Fienup 2018; Richling et al. 2019). Mastery criteria are performance-based rules instructors use to determine when to terminate an instructional phase and engage in a new teaching phase. The new teaching phase could include teaching behavior with faded prompts, leaner schedules of reinforcement, or allocating time to teaching entirely new skills (Fuller and Fienup 2018). A recent descriptive analysis of mastery criterion identified discrepancies between how researchers and practitioners arrange components of mastery criteria (McDougale et al. 2019). The survey findings revealed that practitioners required observations of behavior over a longer period of time (80% accuracy *across three consecutive sessions*) and researchers required higher accuracy (*90% accuracy* across two consecutive sessions). Empirical research comparing different criteria has revealed that only stringent mastery criteria produce adequate response maintenance (90% accuracy levels, Fuller and Fienup 2018; 100% accuracy across three consecutive sessions, Richling et al. 2019). Indeed, there is room for continued investigation of how an instructor's use of mastery criterion affects the acquisition and maintenance of skills.

The unit of behavior at which mastery criterion is applied is another aspect of mastery criteria that warrants attention. In the ABA literature, instructors administer blocks of trials, composed of multiple operants and multiple opportunities for each response, and evaluate mastery as the percentage of correct responses across all operants and opportunities in the block. Instructors administer blocks of operants

[1] Albers and Greer (1991) expanded on the concept of 3-term contingency instruction—which focuses on the contingencies unique to the student—to include an instructor's interlocking 3-term contingencies. A learn unit constitutes the behavior of both the instructor and student, such that the child's readiness is the discriminative stimulus for the teacher to deliver instructions, which is the discriminative stimulus for the student to respond, the student's response is a discriminative stimulus for the instructor to provide differential consequences, and the student's ultimate independent response functions as a reinforcer for instructor responding.

for good reason, as there is a robust literature demonstrating that teaching students with disabilities conditional discriminations in randomized blocks of operants is both effective and efficient compared to other instructional arrangements (Grow et al. 2011,2014; Grow and Van der Hijde 2017). However, a side effect of teaching new operants in sets is the shift in focus of mastery on sets of operants—rather than individual operants—as evidenced by researchers reporting aggregated accuracy, or percentage correct across all operants taught in a session.

When an instructor applies mastery criterion at the level of the set ofoperants, two problematic outcomes may occur. First, in cases where the criterion level is below 100%,[2] errors concentrated with a particular operant may go unnoticed. For example, if an instructor teaches a set of four operants, each with five response opportunities per session (20 trials 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 in the set. Thus, the unit of analysis may affect the effectiveness of instruction.

A second potential problem with a set-based analysis is that mastery of the set of operants is inextricably linked to the operants that are acquired the slowest, thus affecting the efficiency of instruction. In our own experience, we have observed that during acquisition, some students rapidly acquire a few of the operants in a set and require additional sessions to master the remaining operants. Figure 1 displays two cases of analyzing behavior at the levels of the set (top panels) and the individual operants (bottom panels). On the left, the graphs display a case where the two units of analysis do not align. The student masters the set (criterion of 100% correct across 1 observation) in 9 sessions (top panel). When examining the individual operants in the bottom panel, individual operants meet that same criterion (100% correct) in 2, 3, 3, and 9 sessions, respectively. Thus, three of four operants continued in a skill acquisition teaching phase for 6 or 7 sessions after the participant demonstrated proficiency with those responses. On the right side of Fig. 1, the graphs display a case where the analysis of sets and individual operants align more closely. The participant masters the set in 4 sessions (top panel) and examining the data in the bottom panel reveals the participant masters individual operants in 2, 3, 4, and 4 sessions, respectively.

The purpose of this experiment was to directly compare the effects of the applying mastery criterion at the level of the set or individual operant and how this affects the acquisition and maintenance of operants. In the Set Analysis (SA) condition, we taught static sets of four sight words until the participant's performance met mastery criterion, which was evaluated as percentage of correct responses across all sight words in the set. Once a participant's responding met mastery criterion, we substituted a mastered set of words with a new static set of four sight words. This condition constituted a traditional application of DTI mastery criterion. In the Operant Analysis (OA) condition, we taught dynamic sets of four operants and analyzed performances at the level of the operant. As individual sight words met the

[2] Surveys by McDougale et al. (2019) and Richling et al. (2019) show that researchers and practitioners often use criterion with accuracy levels below 100%.

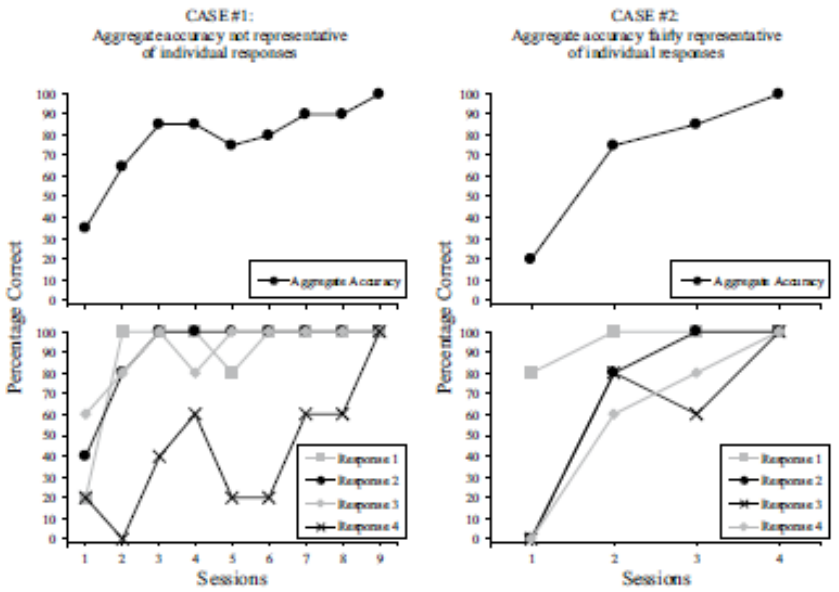


Fig. 1 The graphs display aggregated (tops panels) and disaggregated (bottom panels) data for two cases. In the first case, the participant acquires responses 1, 2, 4 within the first few sessions and struggles to acquire the 3rd response until session 9. In the second case, the participant rapidly acquires all responses within two to four sessions

mastery criterion, we substituted new sight words in the next session. This condition served as a more dynamic application of DTI mastery criterion. We chose sights words as target operants to teach because expanding students' vocabulary repertoire is educationally significant and the instruction did not deviate from the students' current reading curriculum. The objective of the current study was to examine how an instructor's application of mastery criterion affects the learning of his or her students.

Method

Participants

Four first-grade students participated in the study. All participants attended a behavior analytic special education school that implemented the Comprehensive Application of Behavior Analysis (CABAS®) model (Greer 2002). Arnie was 6 years old and he was educationally classified with a Speech and Language Impairment. Arnie's level of educational programming at the onset of the study included reading level B books from the Reading A-Z curriculum and textually responding to 25 printed common sight words. Arnie received behavioral analytic services for the past two years. Max was 7 years old and he was educationally classified with Other

Health Impairments. Max was also reading level B books from the Reading A-Z curriculum. At the onset of the study, he textually responded to 20 printed common sight words. Max had a shorter history of receiving behavioral analytic services. He began receiving services three months prior to the onset of the study. The third and fourth participants, Jason and Allison, were 6 and 7 years old, respectively, and they were both educationally classified with ASD. Jason's level of educational programming at the onset of the study included reading level H books from the Reading A-Z curriculum. He textually responded to over 800 printed sight words from Fry's Word List (Fry 2004). Jason received behavioral analytic services four months prior to the onset of the study. Allison read level L books from the Reading A-Z curriculum and she textually responded to over 800 printed sight words from the Fry's Word List as well. Allison received behavior analytic services for two years prior to the onset of the study.

Eligibility for participation was based on the following inclusion criteria: (a) attention to instructors and instructional tasks for at least 10 consecutive min at a time, (b) emission of three- to five- word mand and tact utterances, (c) emission of echoics for one- to four-syllable words, and (d) teacher praise functioned as a conditioned reinforcer. We developed the inclusion criteria to ensure participants could engage in the respective academic task. The experimenters evaluated the aforementioned inclusion criteria by conducting baseline observation sessions as a part of the Early Learner Curriculum and Achievement Record (ELCAR; Greer et al. 2019) with each participant prior to the start of the study. Additionally, each participant already received sight word instruction, so the intervention procedures did not deviate from the standard instruction they received on a daily basis, which used a traditional, SA application of mastery criterion.

Setting

The study was conducted in a center-based special education school in a suburb of a large metropolitan area. Each session of the study was conducted within a self-contained special education classroom in which the participants received daily academic instruction, or in a small conference room across the hallway from the classroom. The vast majority of the sessions were conducted within the classroom with a few rare exceptions when the classroom was too loud to deliver the sight word instruction. The experimenter's decision to conduct sessions in the conference room only to control for the level of distractibility during each session of the study. The experimenter conducted all sessions of the study at a table while sitting directly next to the participant. The table was pushed up against the corner of the classroom facing a blank wall to limit any distracting variables in the classroom.

Materials

The experimenters used a PowerPoint® slideshow presented on a 13.5" MacBook laptop to deliver sight word instruction. The sight words were presented in four font variations to promote generalization. We used the following fonts: Times

New Roman, Comic Sans MS, Century Gothic, and Calibri, all with the font size of 100 pt. Additional materials included a data sheet to record correct and incorrect responses, a pen, and preferred stimuli to deliver as consequences during instructional sessions.

Measurement

The dependent variable was the accuracy of textual responses to sight words. The criterion for mastery was 100% accuracy across one session. A correct response was defined as the participant's vocal production of a word with point-to-point correspondence to the sight word within 5 s of the presentation of the sight word. An incorrect response was defined as the participant's vocal production that did not correspond to the presented sight word or the absence of a response within 5 s of the presentation of the sight word. We assessed the accuracy of textual responses to sight words continually during skill acquisition and 3–4 weeks following mastery.

Procedure

Target Identification

Prior to sight word instruction, the experimenters identified 40 sight words that were not in each participant's repertoire. We assessed this by presenting the sight word and recording the participant's accuracy. The generic structure of the assessment consisted of the individual presentation of all 40 sight words on a PowerPoint® slideshow. Per opportunity, the experimenter presented the discriminative stimulus, which was the sight word on the computer screen. The experimenter allowed the participant 5 s to emit a response and provided no consequence for correct or incorrect responses. Once the participant responded to the discriminative stimulus or 5 s passed without a response, the experimenter presented the next sight word. This process continued until the experimenter presented all 40 sight words. We conducted three sessions of the assessment. The order in which the sight words were presented varied across sessions. We determined that an operant was not in the participants' repertoire if the participant emitted an incorrect response to the sight word presentation during all three sessions. A single correct response during any of the three sessions resulted in the experimenter discarding that sight word and choosing a new target operant to assess on three occasions. The experimenters ensured that the target sight words were not incorporated into the standard academic instruction outside of the study. Each participant received one preassessment session per day. The results of the preassessment (outlined in greater detail in the Results section) show that the participants all emitted zero correct responses to all 40 words during the three pre-assessment probe sessions under each condition. Thus, 20 words were assigned to OA and 20 words were assigned to SA.

Arnie and Max

We drew sight words for Arnie and Max from the Fry Sight Word List: Ninth and Tenth levels (Fry 2004). The Fry Sight Word List is a list of the 1,000 most common words in the environment and reading materials. The list is grouped into 10 groups of 100 words and are listed by the frequency in which the words occur. That is, the 100 words in the First 100 list are much more common than the 100 words in the Tenth 100 list. For Arnie and Max, their typical sight word instruction targeted words from the Fry Sight Word List: First and Second 100. For this study, the experimenters only considered words from the Ninth and Tenth levels to control for the participants coming into contact with more commonly found words outside of the study.

Twenty words were assigned to the SA condition and 20 words were assigned to the OA condition. The experimenters ensured that the target sight words were equated across conditions (OA and SA) through quasi-randomization. This process involved the experimenters numbering each sight word one through 40. The first 20 words were one-syllable words and the final 20 words were two-syllable words. Odd numbered sight words were assigned to one condition and even numbered sight words were assigned to the other condition. The experimenters also ensured that no two words contained the same initial letter within one instructional set. The inclusion criteria for target sight words included (a) one and two syllable words, (b) each one-syllable word contained four letters, (c) each two-syllable word contained six letters, (d) no two words that were phonetically or visually similar could be presented in the same instructional session, and (e) no two words with the same initial letter could be presented in the same instructional session. Furthermore, the experimenter ensured that there were an equal number of one-syllable and two-syllable words assigned to each condition.

Jason and Allison

The experimenters conducted similar target identification procedures for Jason and Allison. Because both participants had over 800 Fry Word List sight words in their repertoire, the experimenters chose more advanced three- and four- syllable sight words not contained within the Fry Word List. During target identification, the experimenters identified 40 sight words that were not in Jason or Allison's repertoire. As described above, the experimenters ensured that the target sight words were equated across conditions. The inclusion criteria for the target sight words were the same for Jason and Allison, with the change of including three- and four- syllable words. Each three-syllable word contained nine to 10 letters and each four-syllable word contained 12–13 letters.

General Procedure

We taught textual responses to sight words via learn unit instruction (Albers and Greer 1991) during the intervention phase. When the participant was attending to the experimenter, the experimenter delivered the antecedent, "Let's learn some new words!" and began the slideshow with the presentation of the first sight word. If the participant did not attend to the screen, the experimenter vocally prompted the participant by saying, "Look here" and pointed at the screen. Once the participant looked at the stimulus on the screen, the experimenter provided the participant 5 s to emit a textual response to the visual antecedent. Contingent on an independent correct response, the experimenter immediately praised the participant's response (e.g., "Awesome job, that is the word _____,") and provided a token or edible on a VR 3 schedule of reinforcement. Contingent on an incorrect response, the experimenter immediately began the correction procedure. The correction procedure began with the experimenter modeling the correct response. If the target word presented on the laptop was "hope," the experimenter pointed to the word and said, "hope." The participant emitted an echoic by saying, "hope" immediately after the experimenter delivered the model. Then, the experimenter provided the participant an independent opportunity to emit the correct response by representing the discriminative stimulus and providing him/her 5 s to emit a response. If the participant read "hope," the experimenter immediately presented the next sight word. If the participant still emitted an incorrect response, the correction procedure began again. The experimenter re-presented the correction procedure up to three times before moving on to the next operant. Correct responses during the correction procedure were not reinforced. It should be noted that all the participants emitted a correct independent response during the correction procedure before moving on to a new trial.

In both conditions, the experimenter presented 20 sight word learn units per session. The mastery criterion was 100% correct in one session, applied either at the level of the individual sight word or set of sight words. The accuracy level of 100% correct was consistent across both conditions to ensure that accuracy level was not the variable responsible for any changes in the rate of acquisition. We held the frequency component (across one observation) constant as well. Every session contained 4 different targets. When there were fewer than 4 operants left to master under the OA condition, we ensured that there were distractors present. The distractors were previously mastered words that were presented under the response maintenance condition.

Set Analysis

During this condition, the experimenter presented static sets of stimuli per session. We assessed mastery at the level of the set, which was 100% correct responding for all four sight words in one instructional session. When a participant's performance met the mastery criterion, the experimenter substituted a new static set of four sight words. The experimenter continued this process until the participant mastered 20 textual responses to sight words in this condition or the OA condition.

Operant Analysis

During this condition, the experimenter presented dynamic sets of sight words per session. We assessed mastery at the level of the individual sight word, or operant. The criterion was 100% correct responding for an individual sight word in one instructional session, across all 5 trial presentations. When a participant's performance met the mastery criterion for a particular sight word, that sight word was substituted for a new sight word in the next session. In this manner, the set of stimuli was dynamic and the specific four sight words being taught in any given session could vary. The experimenter continued this process until the participant mastered 20 sight words in this condition or the SA condition.

Response Maintenance

The experimenters measured the maintenance of responses during follow-up sessions 3 or 4 weeks after a specific sight word was mastered under both conditions. For sight words in the OA condition, we carefully documented the specific day a sight word was mastered and assessed maintenance 21 days from the mastery date for Jason and Allison and 28 days from mastery for Arnie and Max. Thus, across both conditions, we assessed sight words using the same time intervals.

The maintenance sessions were embedded in the skill acquisition sessions. Thus, there were other non-mastered words and/ or mastered words in each maintenance session. For the response maintenance sessions, the experimenters presented the mastered sight word a total of five times. The experimenter presented the sight words during the response maintenance sessions in the same manner as the intervention sessions with the exception of the consequence for incorrect responses. If the participant emitted an incorrect response, the experimenter continued to the next sight word presentation. The experimenter did not provide a correction procedure for incorrect responses during maintenance sessions to remove the opportunity for a participant to learn through the correction procedure during maintenance assessments. The criterion was 100% correct responding.

Procedural Modifications

After every session of the learn unit instruction (SA and OA conditions), the experimenter graphed the percentage of correct responses in order to analyze the data and implement any necessary instructional modifications if a participant's data did not demonstrate learning. The experimenters implemented a decision protocol (for a full description of the model, see Keohane and Greer 2005). Per the decision protocol, when there were two consecutive data points at 0% correct or six consecutive data points showing no trend or a descending trend, the experimenters implemented simultaneous prompting (Cengher et al. 2018) as a modified learning tactic in the following session. Simultaneous prompting involved the experimenter presenting the target sight word and vocally producing the sight word. For example, in the OA condition, Max emitted zero correct responses across two sessions for the word "safe."

Thus, the experimenter followed the decision protocol and implemented simultaneous prompting during the next instructional session for the word "safe". Under the OA condition during the initial comparison of OA and SA, the experimenters made one decision to implement a simultaneous prompting (5% of sight words) for Arnie, three decisions (15% of sight words) for Max, one decision (5% of sight words) for Jason, and no decisions (0% of sight words) were required for Allison. The experimenters did not implement any procedural modifications under the SA condition because the aggregate accuracy data paths never met a decision rule.

Experimental Design

The experimenters used an adapted alternating treatments design (Sindelar et al. 1985) with a best treatment option (Gorgan and Kodak 2019) to test the effects of SA and OA applications of mastery criterion. To begin the study, each participant received three pre-intervention probe sessions to ensure that he or she did not have any of the sight words in repertoire and to provide evidence that the participants did not acquire the sight words as a function of maturation or other external variables. Then, the intervention phase began. We counterbalanced the assignment of target sight words to conditions between Arnie and Max as well as Jason and Allison, given that each pair learned words at the same level. That is, the sight words Arnie learned in the OA condition were the words Max learned in the SA condition. Per participant, we counterbalanced the order of teaching each sight word from each condition such that if the OA condition came first on one day, the SA condition was run first the next day and the opposite order was conducted for the other participant in a pair. The experimenter presented one session of each condition per day. For the best treatment option aspect of the design, once a participant achieved mastery criterion for all 20 sight words in one condition (OA or SA), the experimenter taught the remaining to-be-mastered sight words in the condition that produced a faster rate of acquisition.

Interobserver Agreement and Treatment Fidelity

A trained independent observer collected interobserver agreement (IOA) data for 33.3% of the pre-intervention probe sessions for all four participants and 31.3%, 37.3%, 66.7%, and 55.6% of intervention and maintenance sessions for Max, Arnie, Jason, and Allison, respectively. The experimenters calculated IOA by dividing the number of agreements between the experimenter and the observer by the total number of trials (i.e., 20) and multiplied that number by 100. Agreement was 100% during all observations.

We collected treatment fidelity data on two aspects of the study: (a) implementation of the instruction, and (b) accuracy in making decisions whether sight words were mastered or not. To evaluate the fidelity of instruction, a trained independent observer completed a Teacher Performance of Rate and Accuracy (TPRA, Ingham and Greer 1992) form for 33.3% of the pre-intervention probe session for all participants, and 31.3%, 37.3%, 55.6%, and 66.7% of the intervention and maintenance

sessions for Max, Arnie, Allison, and Jason, respectively. On the TPRA form, an independent observer rated the accuracy of each antecedent and consequence delivered by the experimenter (per learning opportunity). Treatment fidelity was calculated by dividing the total number of correct experimenter responses by the total number of responses and multiplying that number by 100 to get a percentage of fidelity. Treatment fidelity was 100% for all participants across all phases of the study.

To measure the fidelity of the mastery criterion decisions made throughout the study, an independent observer analyzed 100% of the graphs and data sheets after the study was completed. The independent observer rated the accuracy of the experimenter's decisions that a sight word or set of sight words was mastered. Fidelity with mastery criterion decisions was 100%.

Results

Pre-Intervention Assessment

All four participants emitted correct responses for zero of the 40 sight words during the three pre-intervention assessments of textual responding to sight words. At this point, the experimenter assigned 20 sight words to each of the two conditions (SA and OA) in accordance with the rules stated above (see Target Identification section in Procedures).

Acquisition

Individual acquisition graphs can be found in Appendix A, B, C, and D, for Arnie, Max, Jason, and Allison, respectively. The graphs display data according to how the data were collected during the study. Graphs for the SA condition (left side of each panel) display aggregate accuracy across all sight words in a set. Closed circles represent regular instruction and open circles represent sessions with procedural modifications in place. Graphs for the OA condition (right side of each panel) display individual sight words as data paths. There are four graphs representing the OA condition. Each set of 4 sight words taught within the same instructional session are visually represented vertically across 4 panels. For example, Arnie was taught tone [top panel], else [second panel], snow [third panel], and fair [bottom panel] in session 1 of the OA condition. Note there was a delay in the implementation of a procedural modification, which occurred after 3 sessions of 0% correct responding rather than 2 sessions for Max (Appendix B) and Jason (Appendix C). These occurred in the OA condition. For Max, this occurred during the best treatment phase and not during the comparison of OA and SA.

Figure 2 displays the cumulative number of sight words mastered in both conditions for Arnie (top panel), Max (second panel), Jason (third panel), and Allison (bottom panel). Arnie achieved mastery criterion for all 20 sight words in the OA condition in 21 sessions. At that point, he had mastered 8 sight words in the SA

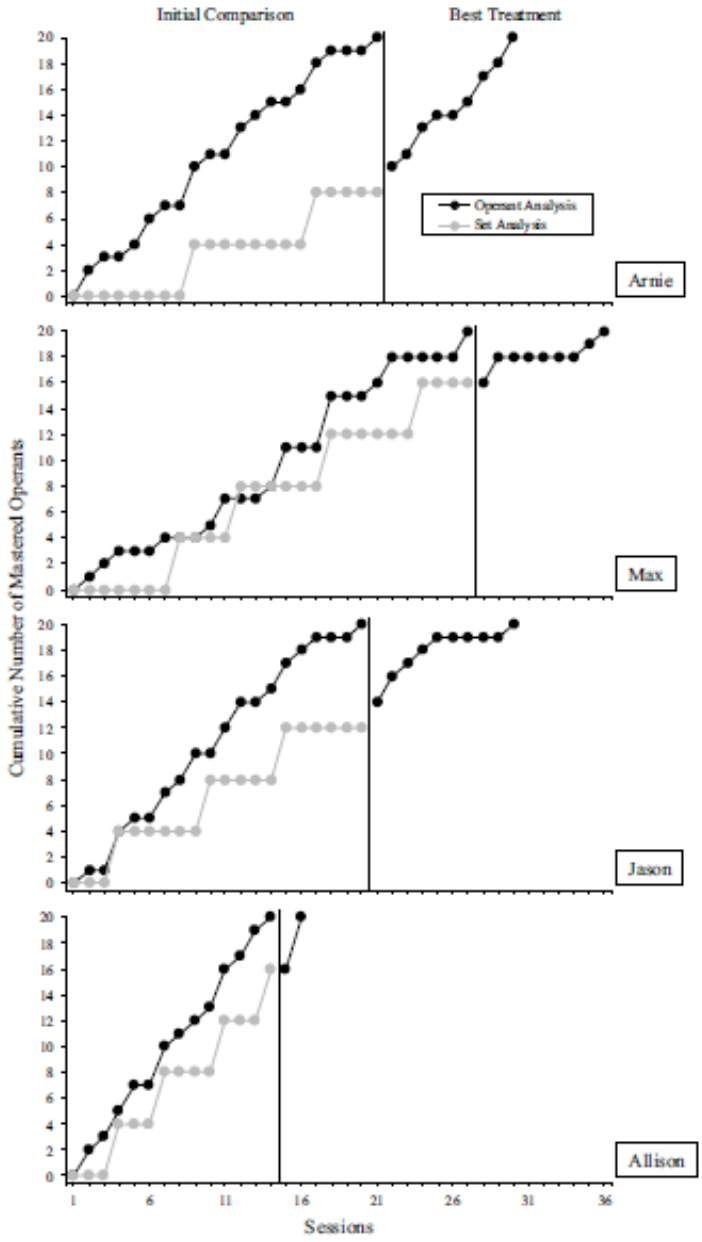


Fig. 2 The graphs display the cumulative number of operants (sight words) mastered under operant analysis (black circles) and set analysis (grey circles) mastery criterion conditions.

Table 1 Teaching trials required to master a target operants

	Arnie		Max		Jason		Allison	
	OA	SA	OA	SA	OA	SA	OA	SA
Number of Teaching Trials	140	340	380	480	195	300	215	280
Mean per Operant	17.50	42.50	23.75	30.00	16.25	25.00	17.92	23.33

condition. Arnie mastered 2.50 times as many sight words in the OA condition. Then, the remaining SA sight words were taught using the OA procedures and he mastered the remaining 12 sight words in 9 sessions. Max achieved mastery criterion for all 20 sight words in the OA condition in 27 sessions. At that point, he had mastered 16 sight words in the SA condition. Max mastered 1.25 times as many sight words in the OA condition. Then, the remaining SA sight words were taught using the OA procedures and he mastered the remaining 4 sight words in 9 sessions. Jason achieved mastery criterion for all 20 sight words in the OA condition in 20 sessions. At that point, he had mastered 12 sight words in the SA condition. Jason mastered 1.67 times as many sight words in the OA condition. Then, the remaining SA sight words were taught using the OA procedures and he mastered the remaining 8 sight words in 10 sessions. Allison achieved mastery criterion for all 20 sight words in the OA condition in 14 sessions. At that point, she had mastered 16 sight words in the SA condition. Allison mastered 1.25 times as many sight words in the OA condition. Then, the remaining SA sight words were taught using the OA procedures and she mastered the remaining 4 sight words in 2 sessions. Overall, all four participants acquired novel sight words faster in the OA condition. During the best-treatment option, Arnie, and Allison demonstrated similar rates of acquisition compared to the previous phase of OA instruction. Max and Jason demonstrated a slightly slower rate of acquisition compared to the previous phase of OA instruction.

Table 1 displays the total number of teaching trials in both OA and SA conditions as well.as the mean number of trials needed to master each target operant within both conditions. Each participant required many more trials to master the target operants in the SA condition. Arnie required 200 more teaching trials, Max required 100 more teaching trials, Jason required 105 more teaching trials, and Allison required 65 more teaching trials. On average, it took 25, 6.25, 8.75, and 5.41 more trials per operant to master each response in the SA condition. Overall, we found evidence of potentially unneeded instruction for all four participants during the SA condition.

Response Maintenance

We report maintenance data for sight words mastered *prior to the best treatment option phase of the experiment*. We analyzed data prior to the best treatment option phase in order to control for time. That is, per participant we examined the accuracy responding during follow-up assessments of sight words acquired during a fixed amount of time. Including data from the best treatment option phase would

have favored the OA condition given that each participant transitioned over to the OA condition and ultimately spent more sessions learning sight words under the OA condition.

We analyzed maintenance data in two manners. First, we examined the percentage of sight words that were accurately responded to with 5 out of 5 correct responses (100% accuracy) during the follow-up assessment. Behavior analysts are interested in procedures that reliably produce durable responding and this measure captures this; however, this analysis was limited in that the denominator (number of sight words initially acquired) differed between conditions. Figure 3 displays the percentage of sight words responded at 100% accuracy 4 weeks after the initial acquisition of the textual responses for Arnie and Max, and 3 weeks after the initial acquisition of the textual responses for Jason and Allison.[3] During the maintenance sessions, Arnie responded accurately to 75% (15 out of 20) of the sight words from the OA condition and 100% of the sight words from the SA condition (8 out of 8). Max responded accurately to 45% (9 out of 20) of the sight words from the OA condition and 56% (9 out of 16) of the sight words from the SA condition. Jason responded accurately to 100% (20 out of 20) of the sight words from the OA condition and 100% (12 out of 12) of the sight words from the SA condition. The final participant, Allison responded accurately to 95% (19 out of 20) of the sight words from the OA condition and 100% (16 out of 16) of the sight words from the SA condition. Overall, there was a clear advantage for accurate maintenance responding in the SA condition for one participant (Arnie). There was a moderate advantage for accurate maintenance responding in the SA condition for one participant (Max), and both OA and SA conditions were effective in producing high maintenance responding for two participants (Jason and Allison).

The second manner in which we analyzed follow-up responses was to count the total number of sight words the participants responded to at a predetermined accuracy level (100%). Since the two conditions produced a different number of mastered sight words (the denominator in first maintenance analysis, above), this measure characterizes the net effect, or total output, of using the respective procedures within a defined period of time—up to the best treatment phase. That is, this analysis characterizes the total number of sight words acquired during a set period of time—which may be of interest to clinicians trying to maximize learning. Figure 4 displays the number of sight words each participant textually responded to at 100% accuracy across one session during the 3–4 week follow-up assessment. Arnie responded accurately to 15 sight words from the OA condition and 8 sight words from the SA condition. Max responded accurately to 9 sight words from the OA condition and 9 sight words from the SA condition. Jason responded accurately to 20 sight words from the OA condition and 12 sight words from the SA condition. Allison responded accurately to 19 sight words from the OA condition and 16 sight words from the SA condition. Overall, Arnie, Jason, and Allison responded accurately to a higher number of sight words under the OA condition and Max had the

[3] We were unable to collect 4-week maintenance data for Jason and Allison due to the end of the school year.

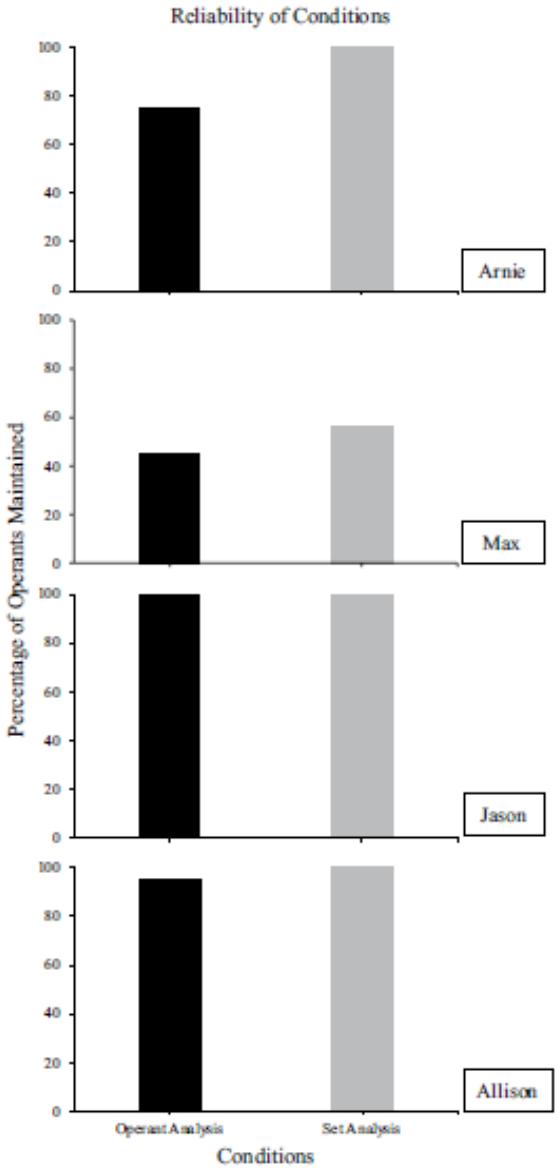


Fig. 3 The graphs depict maintenance responding 3 (Jason and Allison) to 4 (Arnie and Max) weeks following the mastery of sight word operants. Only data during the initial comparison are plotted. The graphs display the percentage of operants, or sight words, maintained at 100% accuracy.

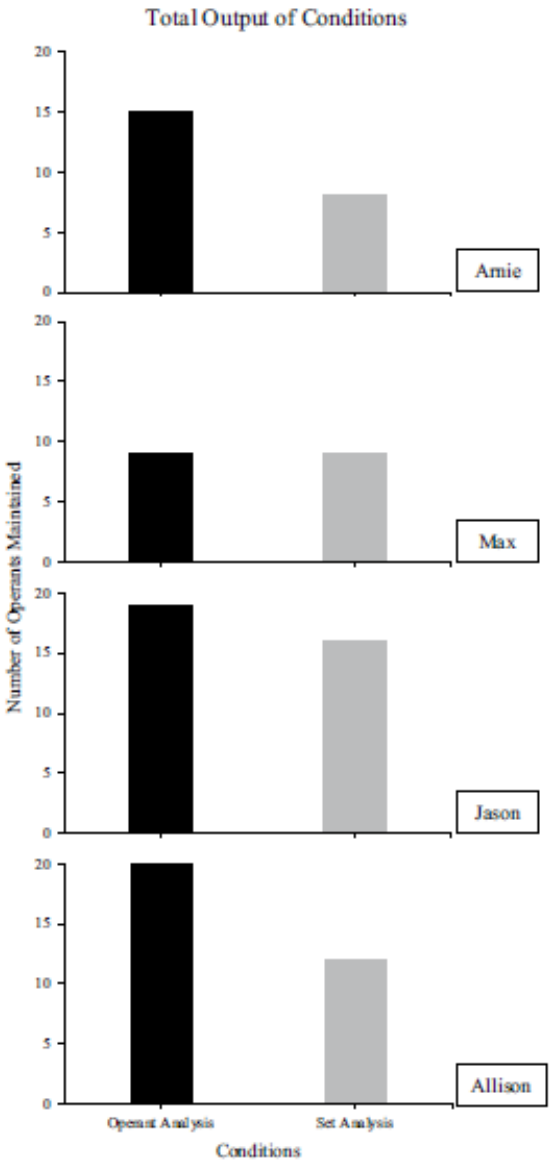


Fig. 4 The graphs depict maintenance responding 3 (Jason and Allison) to 4 (Arnie and Max) weeks following the mastery of sight word operants. Only data during the initial comparison are plotted. The graphs display the number of operants maintained at 100% accuracy. All data are from an equal time period of acquisition per participant prior to the crossover feature of the study.

same number of accurate sight words across conditions. This analysis was necessary because the students will have a stronger foundation in literacy when they acquire a greater number of sight words in his or her repertoire (Fry 1960).

Within-SA-Condition Analysis

Overall, participants acquired a higher number of sight words during the OA condition and responded accurately to a similar (Max) or higher number (Arnie, Jason, Allison) of sight words in the OA condition during follow-up sessions. We further analyzed the data in the SA condition to determine if there was evidence of potentially unnecessary training trials—called overtraining trials for the sake of this analysis. To do so, we disaggregated data from the SA condition and graphed the data per sight word akin to how we graphed OA condition data (see Fig. 1 for an example of disaggregating data). We applied the OA of mastery criterion to the SA condition data and then counted how many instructional sessions were conducted with an operant after the response met the OA condition criterion, if any. Then we multiplied the number of sessions by 5 trials because each session involved 5 trials of each sight word. This resulted in the number of overtraining trials for a single sight word. For example, the left panel of Fig. 1 displays the first teaching phase of Arnie's SA condition. The top panel illustrates the number of sessions Arnie needed to master the four sight words (fear, swim, cool, and tall). He responded with 35% accuracy in the 1st session and his accuracy gradually increased until he responded with 100% accuracy during the 9th session. The bottom panel illustrates the disaggregated data of the 4 sight words. Arnie responded with 100% accuracy for the operant "fear" (response 1) during the 2nd session and we continued teaching this word for 5 trials a session for an additional 7 sessions, or 35 overtraining trials. The words "swim" and "cool" (responses 2 and 3, respectively) met the OA criterion during the 3rd session and we continued teaching these sight words for an additional 6 sessions, or 30 overtraining trials for each sight word. There was no additional teaching for the word "tall" (response 4) since the mastery of this word brought the aggregate accuracy up to the OA criterion (100% correct). In total, at the level of individual sight words, there were 95 (35 + 30 + 30) overtraining trials during the mastery of this particular set of sight words. These trials could have been allocated to teaching new sight words. The 95 overtraining trials accounted for over 50% of total trials needed to Arnie to meet mastery criterion for this set of sight words. Had these sight words been taught with in the OA condition, Arnie would have been introduced to learn at least 3 new sight words within that same time period. With Allison's data from the third phase of the SA condition (see right side of Fig. 1), two sight words (responses 1 and 2) met the OA condition prior to the set and she acquired all the sight words relatively quickly, resulting in fewer overtraining trials (10 + 5 = 15, or 19% of the 80 total trials for this set were overtraining trials).

Table 2 displays the overtraining trials for each participant per SA phase, the percentage of overtraining trials, as well as the average number of overtraining trials. We calculated the average overtraining trials by dividing the total overtraining trials by the total number of sight words mastered in the SA condition. To understand the

Table 2 Sessions of potentially unnecessary overtraining trials for the SA condition

	Arnie	Max	Jason	Allison
Phase 1 Overtraining Trials (%)	95 (53%)	65 (41%)	20 (25%)	15 (19%)
Phase 2 Overtraining Trials (%)	85 (53%)	10 (13%)	35 (30%)	10 (17%)
Phase 3 Overtraining Trials (%)	n/a	30 (25%)	30 (30%)	15 (19%)
Phase 4 Overtraining Trials (%)	n/a	45 (38%)	n/a	10 (17%)
Phase 5 Overtraining Trials (%)	n/a	n/a	n/a	n/a
Total Overtraining Trials (%)	180 (53%)	150 (31%)	85 (28%)	50 (18%)
Number of Operants Mastered	8	16	12	16
Average Overtraining Trials Per Operant	22.5	9.4	7.1	3.1

Note "n/a" indicates that a participant did not master sight words in a particular phase because of the best treatment feature of this study. The percentages represent the percent of the total number of trials in each phase that were potentially uneccessary. To calculate Average Overtraining Trials Per Operant, we divided the Total Overtraining Trials by the Number of Operants Mastered. Each of the five phases represented the opportunity to master four sight words

scale of this dependent variable, an average of zero indicated no overtraining and any positive number meant some unnecessary teaching occurred during the SA condition. The largest percentage of overtraining trials occurred with Arnie. For each phase of instruction, an average of 53% of the total instructional trials were overtraining trials. With Max, Jason and Allison, an average of 31%, 28%, and 18% of the total instructional trials were potentially unnecessary and over-trained. Overall, we found evidence of potentially unneeded instruction for all four participants during the SA condition.

Discussion

Behavior analysis is the science of the behavior of individuals. As such, the application of behavioral principles is designed to improve an individual's specific operants or repertoires (Baer et al. 1968). While ABA interventions target specific operants for specific individuals, the manner in which researchers and clinicians apply mastery criterion is at an aggregated level, across multiple operants. Assessing mastery on an aggregate level may result in one of two problems. First, when criterion is set below 100% (e.g., 90% criterion level), one or more individual responses may contain errors, but the aggregate accuracy masks this. Second, aggregate accuracy across a set of stimuli may be an appropriate proxy for individual operants (e.g., see right panels in Fig. 1) or representative only of the slowest-to-acquire responses (e.g., see left panel of Fig. 1). When aggregate accuracy is not representative of individual performances, failure to master one operant means continuing to teach already proficient responses in lieu of teaching novel behavior. The current study was the first to directly test how an instructor's application of mastery criteria to sets and individual operants affected acquisition and maintenance of responses. The criterion of the SA condition was 100% accuracy across one 20 learn unit session targeting four operants. The criterion of the OA condition was 100% accuracy across

five learn units within one session per operant. The experimenters chose 100% criterion in both conditions of the study because higher mastery criterion levels produce higher levels of maintenance responding (Fuller and Fienup 2018). The experimenters chose 100% criterion across one session to signal termination of instruction because to date, there is no research on determining the most effective frequency of consecutive sessions at a criterion level. Thus, future research should evaluate criterion applied to a different number of observed sessions with both SA and OA.

Overall, the outcomes of the study suggest that applying mastery criterion at the level of the operant (the OA condition) produced quicker skill acquisition and similar or better follow-up performance, in regard to the total number of operants maintained. Participants learned between 1.25 and 2.5 times as many sight words in the OA condition, when mastery criterion was applied to individual sight words. Then, we analyzed responding three to four weeks after we terminated instruction and found that three participants responded accurately to a higher number of sight words (1.9, 1.7, and 1.2 times more words for Arnie, Jason, and Allison, respectively) and one participant responded accurately to a same number of sight words (Max). When examining the percentage of accurate textual responses to sight words, one participant responded accurately to a higher percentage of sight words in the SA condition, one participant responded accurately to a slightly higher percentage of sight words in the SA condition, and two participants responded accurately to a similar high percentage of sight words. The data on the total number of accurate textual responses to sight words along with the percentage of accurate sight words during the follow-up sessions were important because they showed a complete picture of the long-term effects of OA. Both analyses of maintenance responding were necessary due to the differences in the denominator of OA and SA. While the results for the percentage of accurate sight words suggested OA to be a similar or slightly less reliable method of mastery criterion application compared to SA, the total number of accurate sight words during follow-up sessions for 3 of 4 participants provided evidence for OA as a more effective method in teaching a greater number of sight words—an outcomes desired by ABA clinicians. As mentioned earlier, there is educational significance in the application of instructional methods that lead to the acquisition of a greater number of sight words (Fry 1960). In total, for three participants (Arnie, Jason, and Allison), the OA condition produced better educational outcomes and for one participant (Max), the traditional application of mastery criterion (SA condition) and the dynamic application of mastery criterion (OA condition) produced the same educational outcomes.

A nuanced picture of the effects of operant and set analysis emerges when comparing percentage and number of accurate sight words during follow-up sessions. The OA condition produced relatively rapid acquisition of textual responses and a high number of accurate sight words during follow-up sessions, but with some loss because not all sight words were maintained (see Arnie's outcomes). The SA condition produced relatively slower acquisition of sight words and a lower number of maintained sight words during follow-up sessions, but a higher percentage of accurate responses 3–4-weeks after instruction stopped for some participants. One reason why the SA condition produced more reliable follow-up data could be related to the total number of instructional trials and overtraining built into SA procedures

(see Table 1). Because SA is concerned with aggregated accuracy, a participant may be proficient with some operants over consecutive sessions until the set of operants meet criterion. Our analyses of both the total teaching trials (Table 1) and the potentially unnecessary training trials (Table 2) show that overtraining could be associated with improved retention given the higher number of total teaching trials needed to master the individual operants (see Dougherty and Johnston 1996). It should be noted, though, that the quickest-to-be-acquired operants undergo the most overtraining in a SA context and overtraining might be best applied to operants that are more difficult to acquire. This could be accomplished by extending the frequency component of mastery criterion (number of consecutive session). A follow-up study that applies the OA procedure and extends the frequency component to two consecutive sessions may provide the overtraining necessary for students to more reliably maintain difficult-to-acquire operants. The "within SA condition" analysis of the current study, nevertheless, showed that the SA condition produced many sessions of teaching operants that did not require additional teaching, for example, an average of 22.5 overtraining trials per operant for Arnie. Assuming that sight word instruction occurs once daily in a typical ABA setting, we may then deduce that OA allows a student with similar characteristics as Arnie to master a set of sight words in fewer days. For students with different characteristics, like Allison, the improvement may not be as dramatic.

Ultimately, the OA condition produced relatively rapid acquisition of novel operants for two reasons. First, because proficiency of an operant is more obvious when analyzing individual operants and OA allows one to terminate instruction quicker when an operant meets criterion. Second, the experimenter had more opportunities to make teaching decisions when learning was not occurring (see decision protocol in Procedure). The instructor analyzed data paths for each individual operant in the OA condition and the aggregated data path for the set of operants in the SA condition. Analyzing individual data paths allowed for quicker identification of problems to remediate, while in the SA condition, problems with acquisition were masked by the aggregate reporting of data. Once we identified the problems to remediate, we were able to implement evidence-based learning tactics (Cengher et al. 2018) that helped participants learn skills they were not able to learn initially.

McDougale et al. (2019) uncovered discrepancies in the application of mastery criterion with sets of operants in discrete trial instruction between ABA practitioners and researchers. This may be due to the sparse amount of research on mastery criterion. It is imperative that research in this area continues to shed light on components, parameters, and applications of mastery criterion that produce efficient skill acquisition and effective maintenance of responses. Recent studies on the accuracy level of mastery criterion in discrete trial instruction found that accuracy-levels lower than 90% did not produce response maintenance (Fuller and Fienup 2018; Richling et al. 2019). This provides the scientific evidence needed for ABA practitioners to implement higher accuracy levels during skill acquisition. There are other mastery criterion variables that warrant attention. The frequency component (criterion across a number of consecutive sessions) has not been studied. Another component is whether criterion-level performance is required across multiple instructors. This component addresses issues of stimulus generalization and, to the best of our

knowledge, there has been no research to examine mastery across different instructors or an evaluation of how mastery criteria affect generalization. Another area of research could target manipulations of the number of trials per session and the number of targets taught per sessions.

Future research should address the limitations of this analysis. In our experimental design, we implemented a best treatment option phase after the participant achieved criterion with 20 sight words in one condition. The experimenters implemented the best treatment option phase because of the ethical advantages of such a phase. However, this phase is also a limitation because the decision to end the initial comparison may have biased the efficacy results toward the OA condition. A more thorough analysis of the effects of skill acquisition between the two conditions without a best treatment option is needed in future investigations. Additionally, an experimental design that staggered and varied that number of pre-assessment sessions across both OA and SA conditions would have shown stronger experimental control.

We observed variable outcomes across the four participants in terms of acquisition, percentage of accurate sight words, and number of maintained responses. Clearly, there are between-subject variables that account for the between-subject variability. The participants in this study possessed specific ranges of repertoires that potentially explain between-subject variability and limit the generality of our findings to participants who are higher or lower functioning. Individuals whose baseline rate of acquisition is relatively fast may not experience increased efficiency with OA; the OA may be most appropriate for participants who have variable rates of learning and struggle to acquire some but not all responses (see right panel of Fig. 1 for an example of this response pattern). Future research should study different individual-participant variables, such as level of functioning and rate of acquiring responses. Students with less advanced skill repertoires may benefit the most from this procedure and students who learn quickly may not see as much of an effect on their rate of acquisition. Additionally, this study was conducted entirely within the context of teaching textual responses to sight words, which may limit the generality of outcomes to discrete academic responses or sight words, specifically. However, there are many different methods of instruction such as chaining and spiraling curricula.[4] Follow up studies should evaluate the effectiveness of OA and SA conditions on these different methods of arranging instruction.

The procedural modifications in the study included the implementation of simultaneous stimulus prompting contingent on persistent errors. However, the use of such decision protocols is not ubiquitous in the field of ABA and thus, serves as a limitation. In the absence of the decision protocol, we may have obtained different outcomes, despite the fact that the procedural modifications were applied to a small subset (0–15%) of sight words learned in the SA condition. It is possible that

[4] Spiral curricula build skills into larger repertoires across subsequent lessons. Many math curricula are spiral. For example, a child might learn to identify numerals and quantities, followed by counting, and finally addition. In this case, the curriculum spirals from initial discrimination skills to training discriminative functions (counting, adding) with increasing complexity.

the participants may have acquired all the sight words in the OA condition without procedural modifications given the result data shown in SA. Furthermore, the decision analysis was not implemented with individual targets within the SA condition because the aggregated data did not meet the decision rules, which led to a weaker isolation of the effects of OA and SA on mastery. The purpose of the study was to isolate and evaluate two different types of application of mastery criterion; thus no other variables should have been introduced. Future studies in this area of research should isolate the OA and SA analyses without the implementation of any other variables such as the decision protocol in order to gain stronger experimental control. Ultimately, more research is needed to address the nuances of this particular analysis, but the current trajectory of mastery criterion research is a promising start to reconcile the discrepancies between researchers and practitioners.

This study also highlights the importance of collecting maintenance data as a primary dependent variable during comparative analyses. While some studies collect maintenance data when comparing two instructional components (e.g., Kodak et al. 2019), many studies focus only on the initial acquisition of a skill and the speed of that acquisition (e.g., 54% of prompt comparison studies do not report maintenance data; Cengher et al. 2018). The current study presents data that show a clearer difference in the rate of skill acquisition between OA and SA; however, the maintenance results contain important nuances that must be considered by an instructor before adopting an OA mastery criterion. In the absence of maintenance data, we would not have a complete analysis of the application of mastery criterion. Indeed, assessing maintenance in any instructional component comparison study would lead to a better understanding of the durable effects of our interventions.

Acknowledgements We thank Dr. Mirela Cengher for helpful comments on an earlier version of this manuscript

Compliance with Ethical Standards

Conflict of Interest The authors that they have no conflict of interest

Ethical Approval The respective university's Institutional Review Board deemed this research as exempt educational research.

Informed Consent The research study was conducted retrospectively from data obtained for evaluating educational practices. The respective university's Institutional Review Board deemed this research as exempt educational research. Thus, requirements of informed consent were waived.

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.

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

Carroll, R. A., Joachim, B. T., Clarie, C., Peter, St., & Robinson, N. (2015). A comparison of error-correction procedures on skill acquisition during discrete-trial instruction. *Journal of Applied Behavior Analysis*. https://doi.org/10.1002/jaba.205

Cengher, M., Budd, A., Farnell, N., & Fienup, D. M. (2018). A review of prompt-fading procedures: Implications for effective and efficient skill acquisition. *Journal of Developmental and Physical Disabilities, 30,* 155–173. https://doi.org/10.1007/s10882-017-9575-8

Dougherty, K. M., & Johnston, J. M. (1996). Overlearning, fluency, and automaticity. *The Behavior Analyst, 19,* 289. 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. https://doi.org/10.1007/s40617-017-0201-0

Fry, E. (1960). Teaching a basic reading vocabulary. *Elementary English, 37,* 38–42.

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

Greer, R. D. (2002). Designing teaching strategies: An applied behavior analysis systems approach. *Academic Press.* https://doi.org/10.1002/bin.160

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.

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,* 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,* 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,* 62–66. https://doi.org/10.1007/s40617-016-0133-0

Haq, S. S., & Kodak, T. (2015). Evaluating the effects of massed and distributed practice on acquisition and maintenance of tacts and textual behavior with typically developing children. *Journal of Applied Behavior Analysis, 48,* 85–95. https://doi.org/10.1002/jaba.178

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,* 153–164. https://doi.org/10.1901/jaba.1992.25-153

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,* 252–271. https://doi.org/10.1037/h0100749

Kodak, T., Halbur, M., Bergmann, S., Costello, D. R., Benitez, B., Olsen, M., & Cliett, T. (2019). A comparison of stimulus set size on tact training for children with autism spectrum disorder. *Journal of Applied Behavior Analysis, 53,* 265–283. https://doi.org/10.1002/jaba.553

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

McDougale, C. B., Richling, S. M., Longino, E. B., & O'Rourke, S. A. (2019). Mastery criteria and maintenance: A descriptive analysis of applied research procedures. *Behavior Analysis in Practice.* https://doi.org/10.1007/s40617-019-00365-2

National Autism Center (2009). *Findings and conclusions.* National standards project. Randolph, MA: Author.

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

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,* 701–717. https://doi.org/10.1002/jaba.580

Rosenwasser, B., & Axelrod, S. (2001). The contributions of applied behavior analysis to the education of people with autism. *Behavior Modification, 25,* 671–677. https://doi.org/10.1177/0145445501255001

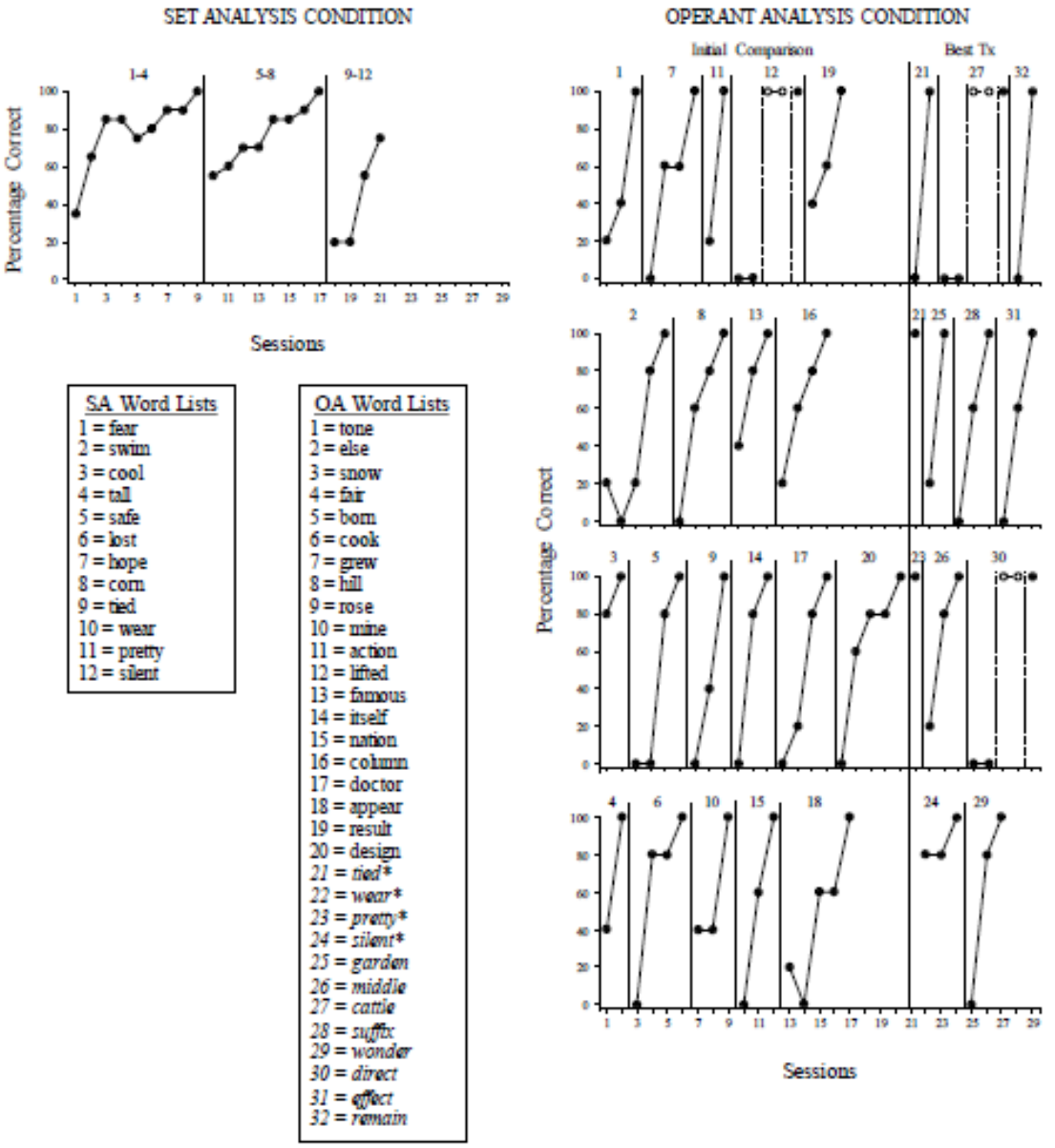
Schnell, L. K., Vladescu, J. C., Kisamore, A. N., DeBar, R. M., Kahng, S., & Marano, K. (2019). Assessment to identify learner-specific prompt and prompt-fading procedures for children with autism spectrum disorder. *Journal of Applied Behavior Analysis, 53,* 1111–1129. https://doi.org/10.1002/jaba.623

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

Online supplemental material below can be found at: https://static-content.springer.com/esm/art%3A10.1007%2Fs10864-020-09420-3/MediaObjects/10864_2020_9420_MOESM1_ESM.pdf

Appendix A

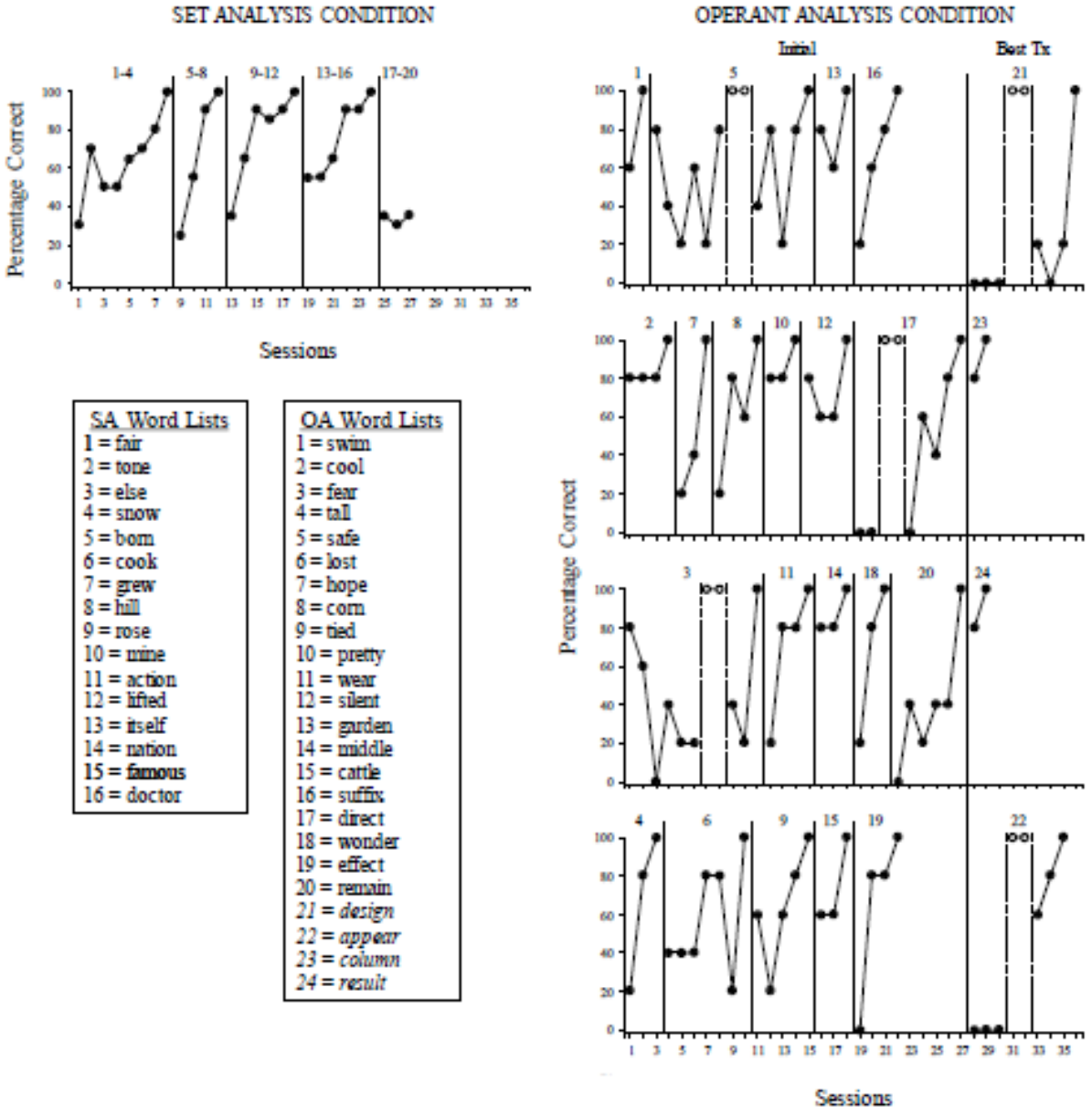
Arnie's Acquisition Data



Note. Data from the SA condition are displayed on the left and OA data are displayed on the right. Broken condition lines signify a procedural modification according to the decision protocol and open circles represent data while the modification was in place. In the OA Word List, italicized words are those from after the crossover and a * indicates a word that had begun intervention during the SA condition prior to the crossover.

Appendix B

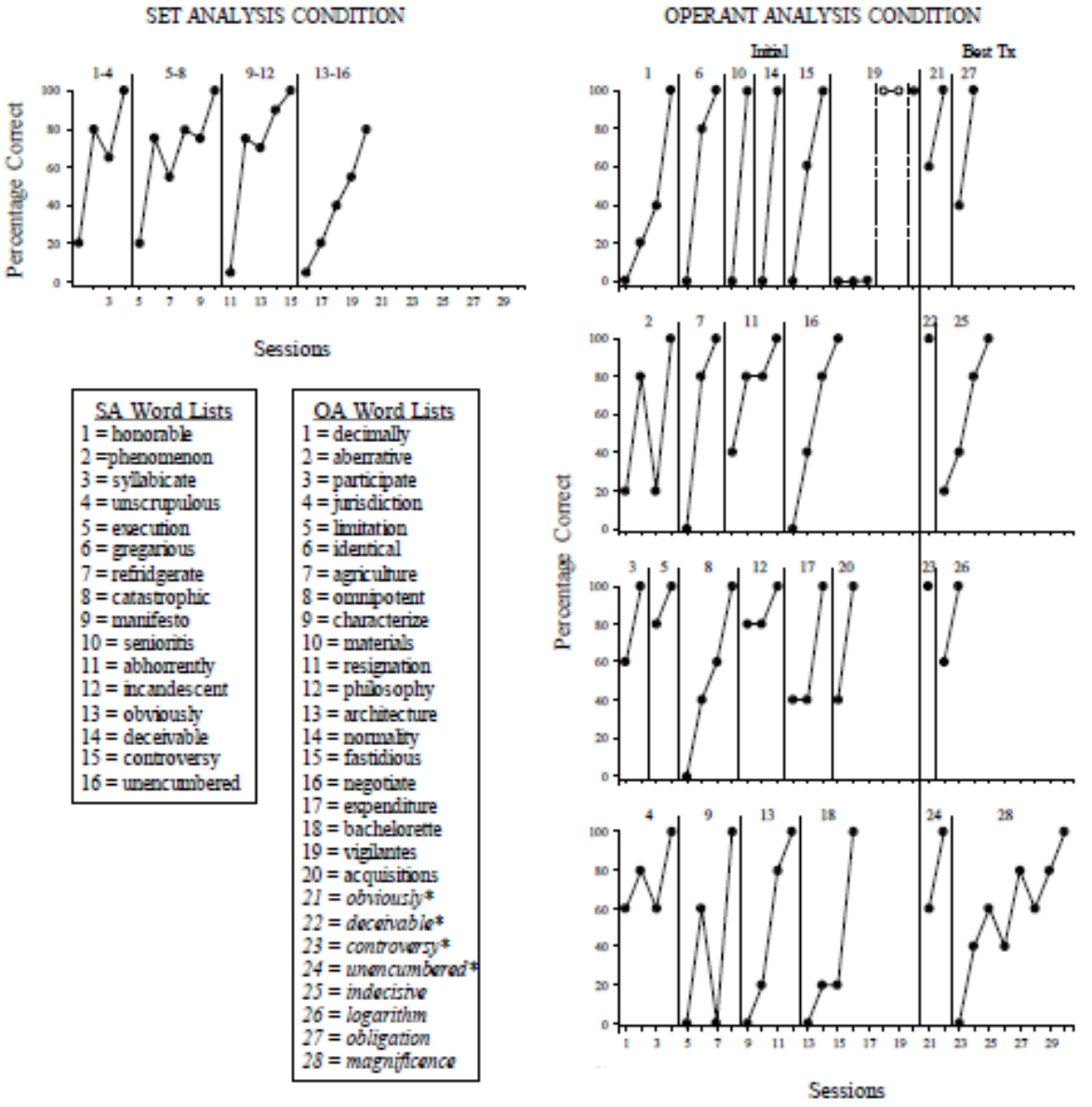
Max's Acquisition Data



Note. Data from the SA condition are displayed on the left and OA data are displayed on the right. Broken condition lines signify a procedural modification according to the decision protocol and open circles represent data while the modification was in place. In the OA Word List, italicized words are those from after the crossover.

Appendix C

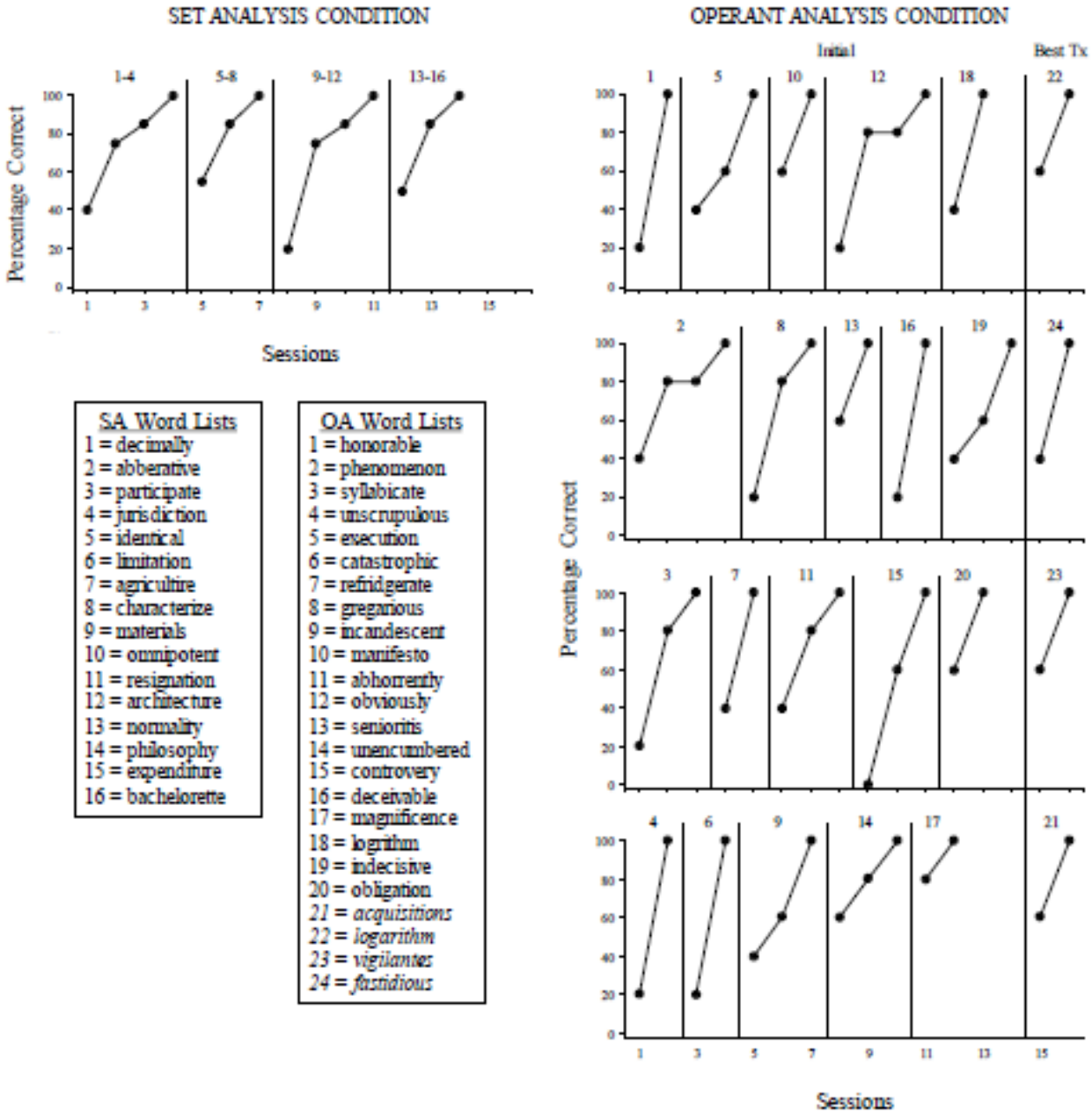
Jason's Acquisition Data



Note. Data from the SA condition are displayed on the left and OA data are displayed on the right. Broken condition lines signify a procedural modification according to the decision protocol and open circles represent data while the modification was in place. In the OA Word List, italicized words are those from after the crossover and a * indicates a word that had begun intervention during the SS condition prior to the crossover.

Appendix D

Allison's Acquisition Data



Note. Data from the SA condition are displayed on the left and OA data are displayed on the right. Broken condition lines signify a procedural modification according to the decision protocol and open circles represent data while the modification was in place. In the OA Word List, italicized words are those from after the crossover.

Chapter 3: Mastery Criterion Units of Analysis: A Replication and Extension

Abstract

This study systematically replicated and extended the findings of Wong et al. (2021) by comparing two units of analysis for assessing mastery during sight word instruction for four 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, I taught sets of four sight words simultaneously and assigned a 100% mastery criterion across two consecutive sessions for each set of words. In the Operant Analysis (OA) condition, I also taught sets of four sight words simultaneously but assigned a 100% mastery criterion across two consecutive sessions for each individual word and replaced individual words as performance met criterion. The results of this study replicated those of Wong et al. (2021) and suggest the rate of acquiring novel textual responses to sight words was faster under the OA condition for all participants. Additionally, this study extended previous research by showing enhanced response maintenance as a function of increasing criterion from one replication of 100% accuracy to two. Future directions and important educational implications are discussed.

Keywords: mastery criterion, operant analysis, response maintenance, set analysis, stimulus set size, frequency of replications

Mastery Criterion Units of Analysis: A Replication and Extension

Learn Unit instruction is an effective instructional methodology to teach individuals who do not learn incidentally in their naturalistic environment. For these individuals, learn unit instruction offers many direct learning opportunities and feedback in a controlled environment. Individuals with developmental disabilities such as Autism Spectrum Disorder (ASD) benefit from receiving learn units to acquire a variety of novel skills in academic, self-management, and communication domains (Greer et al., 1999). Learn unit instruction incorporates interlocking operants between an instructor and a learner (Albers & Greer, 1991). The core components of a learn unit include an instructor-presented discriminative stimulus contingent on a learner observing response, an opportunity for the learner to respond, and an instructor-delivered consequence that is contingent on a correct or incorrect learner response.

An important preparation process occurs prior to the delivery of individual learning opportunities. This involves the instructor outlining a set of target novel skills that are socially or educationally significant to the learner. Once those educational responses are identified, instructors must establish a criterion for mastery, which signals the termination of teaching for the identified set of novel skills. Mastery criterion is applied to nearly all ABA programming (Richling et al. 2019) and yet there is very little research on why particular mastery criteria are chosen. In a survey of 200 Board Certified Behavior Analysts (BCBAs), the primary information source contributing to the practitioners' use of specific mastery criteria were "previous supervised experience" and "employer policies" (Richling et al. 2019). Based on these answers, mastery criteria practices in the field of ABA seem to be based on tradition and customs that have been passed down over time.

The research on mastery criteria in ABA is sparce. Studies conducted in the 1970s with college students suggested high levels of performance on quizzes were required for students to succeed in cumulative exams (Carlson & Minke, 1975; Johnson & O'Neill, 1973; Semb, 1974). Recently, three studies have examined mastery criterion effects with children with disabilities. In a preliminary analysis conducted by Fuller and Fienup in 2018, a 50%, an 80%, and a 90% accuracy level (each across one 20 trial session) was applied to spelling programs for three students with ASD. The authors evaluated each criterion level on maintenance responding 3-4 weeks after the mastery criterion was achieved. The results suggest that higher mastery criterion levels predict greater accuracy in maintenance responding. The 90% accuracy criterion predicted the highest level of responding during 4-week maintenance probe sessions. Richling et al. (2019) also compared different criterion levels (60%, 80%, and 100%, each across three consecutive sessions) (Experiment 2) and 80%, 90%, and 100% criterion levels across three consecutive sessions (Experiment 4). The results showed only the 100% accuracy criterion led to accurate responses above 80% during response maintenance probe sessions (Experiment 2) and provided evidence against the efficacy of criterion levels less than 100% that can produce high response maintenance results. Recently, Pitts and Hoerger (2021) replicated Richling et al.'s 80%, 90%, and 100% across three consecutive session performance criteria and found that both the 90% and 100% criterion levels produced maintenance at or above 90% accuracy. These three studies, while not producing identical outcomes, converge with research conducted with college students demonstrating the necessity of high levels of performance during skill acquisition (Carlson & Minke, 1975; Fienup & Brodsky, 2017; Fuller & Fienup, 2018; Johnson & O'Neill, 1973; Richling et al., 2019; Semb, 1974).

Fuller and Fienup (2018), Richling et al. (2019), and Pitts and Horger (2021) focused specifically on the level of accurate responding in determining mastery criterion. Another aspect of mastery criteria is the unit of analysis to which mastery criteria are applied. Learn Unit instruction is typically administered in blocks of trials with multiple operants in each block and there are multiple opportunities for the learner to respond to each operant. Teaching individuals with developmental disabilities such as ASD using blocks of trials is advantageous and commonly used when delivering learn unit instruction. In a "tacting zoo animals" program for example, an instructor may select four target operants to teach in a 20-trial session. The four target operants may be zebra, monkey, tiger, and lion. In the 20-trial session, there are typically five exemplars of each target stimulus and the instructor may assign a mastery criterion of 80% accuracy during one 20-trial session. This can be referred to as a set-based analysis of mastery criterion because the criterion is applied to an entire set of operants – or accuracy aggregated across all operants and opportunities to respond. Once the performance criterion is achieved, the instructor will select a set of four new zoo animals to teach and this process is repeated until the terminal goal of a particular number of zoo animals is acquired (or long-term learning objective).

Wong et al. (2021) found evidence to show the inefficiency of a set-based analysis of mastery criteria. The authors introduced a novel analysis of mastery criterion, called Operant Analysis (OA) and systematically compared it with Set Analysis (SA), the set-based analysis of mastery criterion. The OA condition required the mastery criterion to be applied at the individual operant level. It involved a dynamic process of replacing mastered operants with novel operants when each individual operant achieved the established mastery criterion. In this analysis of mastery criterion, the acquisition of individual operants was not affected by the acquisition of other operants. The results suggested that all participants in the study learned textual responses to

novel sight words at a much quicker rate with OA compared to SA. In regard to response maintenance, the results suggested similar or more operants maintained (responded at 100% accuracy) under the OA condition 3-4 weeks after the initial mastery of the operants. The skill maintenance results also showed that the percentage of operants maintained after 3-4 weeks was higher under the SA condition for two participants. The differences between the raw number of operants maintained and the percentage of operants maintained were due to the difference in the total number of operants acquired in each condition. Ultimately, the results demonstrated that a greater number of sight words acquired and maintained in the long-term (3-4 weeks), which is a desired outcome for ABA practitioners. However, this nuanced picture of the effects of both OA and SA conditions on response maintenance call for more experimental manipulations to evaluate whether OA can be superior in skill maintenance in both the number of operants maintained and the percentage of operants maintained.

The results of Wong et al. (2021) highlight an important issue with skill acquisition in general. ABA practitioners and researchers alike aim to teach individuals in an effective and efficient manner. Instructional methods that allow individuals to achieve goals faster should be applied. However, the rate of acquisition is not the only priority in teaching. A fast rate of acquisition is of little value when the learner fails to emit accurate responses to the same or similar discriminative stimuli in the long-term after instruction has concluded. Thus, the investigation of beneficial instructional methods requires the consideration of multiple components, including response maintenance, which is a crucial dependent variable to consider.

The purpose of this study is to systematically replicate the procedures of Wong et al. (2021) and extend the previous findings by addressing a couple of key limitations. I taught sight words to four participants under the OA condition and the SA condition. The conditions were

arranged in the same manner as the original study. The primary difference in the procedure is that the mastery criterion for both conditions was 100% accuracy across two consecutive sessions rather than 100% accuracy across one session. This modification of the conditions was done to evaluate whether or not the participants would respond accurately to a similar or higher percentage of operants during maintenance probe sessions compared to SA. Wong et al. (2021) also implemented the decision protocol (Keohane & Greer, 2005) during the skill acquisition phase, which led to questions about whether outcomes were due to OA, the decision protocol, or some combination of these variables. In this study, I omitted procedural modifications to allow for a more direct manipulation of mastery criterion alone.

Method

Participants

Four elementary students participated in the study. Three students attended a public elementary school and placed into a self-contained special education classroom and one student attended a center-based self-contained classroom. All classrooms implemented the Comprehensive Application of Behavior Analysis (CABAS®) model (Greer, 2002). Eligibility inclusion criteria included the following: (a) attention to instructors and instructional tasks for 10 consecutive min at a time 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 consisted of at least tangibles, edibles, and social praise. The experimenters included these criteria to ensure participants could engage in the respective academic task. I 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 for sight words. They also received 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 and met the criterion for the bi-directional naming (BiN) cusp, allowing him to learn language incidentally. His educational level at the onset of the study included reading Level D stories proficiently from the Reading A-Z curriculum. Patrick could accurately identify 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 met the criterion for unidirectional naming (UniN), which means she acquired listener responses incidentally. Her current education level at the onset of the study included reading Level E stories from the Reading A-Z curriculum. She could accurately identify 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 did not demonstrate incidental language learning and possessed no incidental language learning (NiN). At the onset of the study, William was working on reading Level AA story books from the Reading A-Z curriculum, and he had less than 10 words in repertoire from the Fry Sight Word list. Zara, the fourth participant of the study was an 8-year-old female student in third-

grade, and she received behavior analytic services in a CABAS® classroom for two years. Zara was educational classified with ASD. Like William, Zara possessed no incidental language learning in her repertoire (NiN). At the onset of the study, Zara was reading Level A story books from the Reading A-Z curriculum and had approximately 20 words in repertoire from the Fry Sight Word list. All participants had a history of instruction that closely mimicked the SA mastery criterion condition.

Setting

The setting of the study took place in different locations depending on the participant. For Patrick, Katie, and William, I conducted every in-person session of the study within the participants' self-contained kindergarten through second grade classroom of a public elementary school. 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 in front of the participants. 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.

The study took place in two settings for William due to hybrid in-person/remote learning models during his school year. All pre-intervention probe sessions were conducted in the classroom in the same fashion mentioned in the previous paragraph. About 30% of the intervention sessions and post-intervention sessions took place over Zoom® video calls. William sat at his kitchen table next to his mom, with a laptop on the table. About 70% of the intervention sessions and post-intervention sessions took place in the classroom setting as described in the above.

Zara was the only participant who attended a center-based special education classroom. All phases of the study were conducted in a conference room with one table and six adult-sized

chairs surrounding the oval-shaped table. The conference room was approximately 12 m from her classroom. Zara sat on an adult-sized chair adjacent from an instructor at the table. All sessions of the study took place in-person. Although there were multiple periods of prolonged school closings, Zara did not participate in any virtual sessions.

Materials

The experimenters 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 and data sheets and treatment fidelity data sheets.

Measurement

The dependent variable was the participants' 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 mastered and response maintenance four weeks following the end of instruction. An accurate textual response was defined 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 experimenters calculated the

percentage of accurate responses after each instructional session and four weeks following the initial acquisition of the sight word. I also reported the cumulative number of mastered operants in both conditions in the study.

Procedure

I identified up to 40 novel sight words to teach and after equating the target operants, I assigned an equal number of words into each experimental condition in a quasi-randomized manner (Wong et al., 2021). I delivered three baseline assessments of the sight words to the participants over the course of three days. The intervention phase included the delivery of the OA and SA conditions in an alternating and counterbalanced fashion. After a student's performance met the established mastery criterion for sight words, I conducted weekly response maintenance probes for up four weeks (7 days, 14 days, 21 days, and 28 days) after the initial acquisition session to assess accuracy of textual responses to the acquired sight words.

Target Identification

Prior to the onset of the study, I selected 40 novel sight words to teach the Patrick and Katie, and 24 sight words to teach William and Zara. The assignment of each word was done in a quasi-randomized 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 in 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 and Zara), b) each four-syllable word contained 12-13 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.

The pre-intervention assessment procedure involved the experimenter presenting all sight words individually on the PowerPoint® slideshow and collecting data on correct and incorrect responses. During the assessments, the experimenter sat next to the participant at the desk and opened up the slideshow. The experimenter 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 experimenter 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 pre-intervention 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 experimenters substituted the known word for another target sight word that met the inclusion criteria. Furthermore, the experimenters ensured that the sight words taught in the study were not incorporated into the daily academic programming the participants received. After the three baseline sessions, the experimenters 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. The experimenters assigned 12 words to the OA condition and 12 words to the SA condition for William and Zara.

General Teaching Procedure

I used Learn Unit instruction (Albers & Greer, 1991) throughout the teaching phase and replicated the general instructional procedure used to teach sight words by Wong et al. (2021).

In both conditions, I taught four target sight words in each 20 learn unit session. The mastery criterion was 100% accuracy across two consecutive sessions. 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. In the OA condition, when there were less than four target operants left to master, I presented distractors or previously mastered words that were being assessed under the response maintenance condition during their appropriate 1-week, 2-week, 3-week, or 4-week post-mastery assessment day. Thus, there were always at least 15-20 learn units in each session.

Set Analysis. I taught static sets of four target stimuli until the participant responded with 100% accuracy across two consecutive sessions. When the participant met the established criterion, I introduced four novel stimuli to teach in the set. The teaching process continued until all words in the condition were acquired.

Operant Analysis. I taught dynamic sets of four target stimuli per session. These sets were dynamic because the mastery criterion of 100% accuracy across two consecutive sessions was applied at the level of the individual operant. Once an operant achieved the established mastery criterion, I introduced a novel operant in the next session. There was a constant cycle of new operants being taught to the participant until all words in the condition were acquired.

Response Maintenance

I measured the accuracy of textual responses to the acquired sight words during maintenance probe sessions 1 to 4 weeks after each specific sight word was acquired under both OA and SA conditions. I conducted the response maintenance sessions in the exact manner as Wong et al. (2021). The criterion for maintained operants was 100% correct responding during the session for each word. That is, response maintenance was calculated on a per-operant basis. Response maintenance data were calculated for each individual operant under both conditions of

the study. If the participant textually responded to every presentation of the word correctly, the operant was considered maintained. All response maintenance sessions were unconsequated.

Interobserver Agreement and Treatment Fidelity

A trained independent observer collected trial-by-trial interobserver agreement (IOA) data. The experimenters calculated trial-by-trial IOA 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 experimenters collected IOA for 33.3% of the pre-intervention probe sessions for the four participants, 47% of the intervention and maintenance sessions for Patrick, 37% of the intervention and maintenance sessions for Katie, 62.5% of the intervention and maintenance sessions for William, and 71% of intervention and maintenance sessions for Zara. The interobserver agreement was 100% during all observations for Patrick, Katie, and William. The interobserver agreement was 99.7% (range of 95%-100%) for Zara across all phases of the study.

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. I collected treatment fidelity data for 33.3% of the preintervention assessment sessions and 47%, 37%, 62.5%, and 71% of intervention and maintenance sessions for Patrick, Katie, William, and Zara respectively. I calculated treatment fidelity by diving 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

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, before the school closed, I taught Patrick and Katie both 12 sight words that had 4-week maintenance data. Additional sight words had been mastered in the OA condition for each participant; however, to directly compare acquisition and maintenance data of equal sizes, I report all acquisition data gathered prior to the school closure and focus maintenance assessments only on the first 12 sight words mastered. This modification affected William's and Zara's analysis in that I changed the goal from teaching 20 sight words per condition to 12 sight words per condition.

Zara's response maintenance data were affected by prolonged school closings that took place a few times during the intervention. Due to quarantine periods and extended holiday breaks, there were large gaps in time between acquisition sessions at three points in time (between the 6th and 7th session, between the 16th and 17th session, and between the 20th-21st session of intervention). Furthermore, only 6 and 8 words were assessed under 4-week follow-up sessions due to the closings. For Zara's data, I focused on those operants mastered prior to the school closing, but report all data nonetheless.

Pre-Intervention Assessment

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

Acquisition

Figure 1 displays the cumulative number of sight words that Patrick, Katie, and William mastered. Black circles represent words mastered under the OA condition and the gray circles represent words mastered under the SA condition. Open circles represent sight words that were not assessed during maintenance probe sessions for Patrick and Katie. Patrick (top panel) mastered 20 sight words in the OA condition after 27 sessions. During the same time frame, Patrick mastered 12 words in the SA condition. When comparing only 12 mastered operants for both conditions, Patrick needed 10 additional sessions to master 12 operants under the SA condition, or 40% fewer sessions to master 12 operants in the OA condition. Katie (middle panel) mastered 20 sight words in the OA condition after 22 sessions. During the same time frame, Katie mastered 12 words in the SA condition and eventually mastered 16 words under the SA condition after 29 sessions. When comparing only 12 mastered operants for both conditions, Katie needed 6 additional sessions to master 12 operants under the SA condition, or 32% fewer sessions to master 12 operants in the OA condition. William mastered all 12 sight words after 21 sessions in the OA condition and 29 sessions in the SA condition, or 28% fewer sessions to master 12 operants in the OA condition.

Zara's acquisition data are reported in Figure 2. Her data are reported separately due to the caveats mentioned above regarding school closings. Additionally, Zara was taught 11 words under the OA condition due to an experimenter error. Zara mastered 11 sight words after 20 sessions in the OA condition. Zara mastered 12 sight words after 33 sessions in the SA condition. It took 1.8 sessions to master one sight word in the OA condition and 2.75 sessions to master one sight word in the SA condition.

Table 1 provides data comparing the acquisition of 12 words in both SA and OA for Patrick, Katie, and William. Zara's data are also reported by adjusting the OA learn units to reflect 12 mastered operants. The adjustment is explained in further detail below.

I chose to report the total number of learn units to acquire 12 sight words in order to conduct an equal comparison between both conditions where there were also available 4-week maintenance data. In order words, I calculated the mean learn units to criterion for each operant. The table shows that all participants needed many more learn units to acquire the sight words under the SA condition. There was an 82% increase in the number of learn units needed to acquire all operants under SA compared to OA for Patrick. There was a 49% increase in the number of learn units needed to acquire all operants under SA compared to OA for Katie. There was a 76% increase in the number of learn units needed to acquire all operants under SA compared to OA for William. Because Zara acquired 11 operants in the OA condition, I calculated the number of learn units needed to acquire one operant and multiplied that total by 12 to get an estimated report of the total number of learn units needed to acquire 12 operants. Based on this estimate, there would be a 63% increase in the number of learn units needed to acquire all operants under SA compared to OA for Zara.

Response Maintenance

I examined the percentage of operants maintained four weeks following the acquisition of 12 sight words per condition for each participant. An operant was considered maintained when the participant emitted correct textual responses for all presentations of the sight word across two consecutive sessions. Figure 3 displays the percentage of operants maintained with 100% accuracy at 4-weeks. Patrick (top panel) maintained all 12 (100%) of the operants acquired under both OA and SA conditions, making both conditions effective in predicting durable maintenance

responses. 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. These data suggest both conditions are effective in predicting durable maintenance responses. William's overall response maintenance data were lower than Patrick and Katie, but he maintained a higher percentage of operants learned under the OA condition compared to the SA condition. He maintained 8 (67%) of the operants mastered under the OA condition and 7 (58%) of the operants mastered under the SA condition. William's data suggest that the operants mastered under the OA condition were comparable to the operants mastered under the SA condition.

Zara's response maintenance results are reported separately. Figure 4 displays the percentage of operants maintained with 100% accuracy at the 4-week follow-up session (left panel) as well as the total number of operants maintained with 100% accuracy at the 4-week follow-up session (right panel). Zara's data are reported with two graphs because experimenters conducted 4-week follow-up sessions for a different number of operants in both conditions due to prolonged school shut-downs and holiday breaks during the maintenance assessment period. The experimenters assessed the response maintenance of 9 words acquired under OA and 8 words acquired under SA. The left panel of Figure 3 shows that Zara maintained 67% of the operants under OA (black bar) and 62.5% of the operants under SA (gray bar). The right panel of Figure 3 shows that Zara maintained the same number of operants under both conditions. She maintained 6 out of 9 operants under OA and 5 out of 8 operants under SA.

Within-SA Condition Analysis

It is clear that all participants acquired operants quicker in the OA condition (see Figure 1) and they required fewer teaching trials to acquire them (see Table 1). The response maintenance data also showed that OA was effective in producing accurate responses four weeks

after the termination of instruction. I determined that further analysis of the data in the SA condition was meaningful in order to find evidence for potentially unnecessary learn units (overtraining trials). Following the conclusion of the study, I disaggregated data from the SA condition by applying the OA mastery criterion to the first 12 sight words mastered in the SA condition and adding all the learn units it took to achieve the OA mastery criterion for each participant. I analyzed the SA words in this manner to investigate how many trials it would have taken the participants to master all 12 words, had the unit of mastery criterion analysis been applied at the individual operant level. It allowed us to calculate how many potentially unnecessary learn units 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, William, and Zara in their respective SA conditions for the first 12 words that were mastered. I also calculated the average number of overtraining trials per operant by dividing the total number of learn units by the total number of sight words acquired in the SA condition. To clarify the meaning of each number, if the average was 0, that means there was no overtraining. Any number higher than 0 indicates the presence of unnecessary teaching. On average, 28% of the learn units delivered to Patrick were potentially unnecessary, 30% of the learn units delivered to Katie were potentially unnecessary, 35% of the learn units delivered to William were potentially unnecessary, and 36% of the learn units delivered to Zara were potentially unnecessary. I delivered an average of 13, 10, 18, and 28 extra learn units per operant to Patrick, Katie, William, and Zara 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 mastery criteria during learn unit instruction. All participants acquired sight words faster when I applied the OA mastery condition. More importantly, when the 100% accuracy was applied across two consecutive sessions (as opposed to one session in Wong et al., 2021), accuracy during maintenance probe sessions was high under the OA condition suggesting that OA is an optimal application of mastery criteria compared to SA. This study isolated the effects of mastery criterion unit of analysis by eliminating the decision protocol from previous research and found robust effects of applying mastery criteria to individual operants. Even though I did not implement the decision protocol, it is important to note that it is an evidence-based algorithm that is effective in solving student learning problems within acquisition programs (Keohane & Greer, 2005). As such, it may have been in the students' best interest to utilize the decision protocol.

Unlike in Wong et al. (2021), response maintenance outcomes were undifferentiated in this study when the mastery criterion was raised from 100% in one session to 100% across two sessions. Undifferentiated response maintenance suggests that both procedures for mastering operants are effective and give rise to examining whether there are efficiency differences between the two techniques. Indeed, participants required many fewer sessions and learn units to master operants in the OA condition. Wong et al. (2021) did not produce durable response maintenance following a mastery criterion of 100% across one session for individual operants. The simple decision to increase the frequency of sessions to two led to lasting results under the OA condition. ABA practitioners, instructors, and researchers should consider the effects of different frequency of replications at a particular accuracy level on long-lasting behavior change.

In the Behavioral Momentum Theory literature, behaviors that are more resistant to change are reinforced with a rich schedule of reinforcement compared to a thinner schedule of

reinforcement (Nevin, 1992). This concept is relevant to the topic of mastery criteria because mastery criteria determine the proportion of antecedents that evoke behavior and are subsequently reinforced. When mastery criterion levels are higher, then a higher proportion of antecedents that evoke behavior exist, which lead to greater response strength. Increasing the level of accuracy during training leads to more persistent behavior in the face of disruption, such as extinction experiences (i.e., response maintenance assessments). Similarly, increasing the frequency of replications that the criterion level is performed at, increases the overall stringency of the criterion. The results of this study suggest that utilizing a high criterion level (100% accuracy) in addition to an increase in the frequency of sessions from one to two led to more durable response maintenance compared to a high criterion level across one replication. It is interesting to note that the change in the frequency of sessions did not necessarily improve the response maintenance results, perhaps due to the overtraining that already exists under the SA condition.

During the systematic comparison of OA and SA, I did not implement the decision protocol in either condition. While the decision protocol is an effective verbally governed algorithm that allows instructors to solve learning problems (Keohane & Greer, 2005), when it was applied in Wong et al. (2021), there were unintended differences in the number of decisions made in each condition. More decisions were made in the OA condition and thus, it was more challenging to suggest that the unit of mastery criterion analysis alone was the reason for the differences in the cumulative number of operants mastered. Foregoing the decision protocol in this study helped to isolate the effects of OA and SA. 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 implement a teaching tactic if the student is not learning.

Previous literature support differential rates of learning contingent on a learner's repertoire of verbal behavior cusps (Greer & Ross, 2008; Hotchkiss & Fienup, 2020; Longano & Greer, 2010). Children acquire various cusps as they progress through verbal behavior milestones, which allow them to contact new environmental contingencies and learn in new ways, often at an accelerated pace (Greer & Ross, 2008). One important cusp is BiN (Greer & Ross, 2008; Greer & Speckman, 2009; Miguel, 2016). When a learner possesses BiN in their repertoire, they acquire novel object-word relations without direct teaching. A subtype of BiN is UniN. When untaught listener behavior emerges, a learner possesses UniN in their repertoire. Learners who do not learn any language incidentally possess no incidental naming (NiN). Learners who possess BiN in their repertoire acquire novel skills at faster rate compared to learnes with UniN or NiN (Greer & Logano, 2010). There were different in the participants' learning slopes (rise/run). Patrick and Katie possessed higher levels of Naming (BiN and UniN) compared to William and Zara (NiN) and both Patrick and Katie acquired sight words under OA in fewer sessions than William and Zara. Between Patrick and Katie, however, Katie actually required fewer sessions in both conditions to learn 12 sight words compared to Patrick even though she did not have BiN in her repertoire. Nevertheless, conclusions cannot be drawn from such a small sample size. Further evaluations with more participants are necessary to consider how verbal developmental repertoires affect skill acquisition.

There are some limitations that are worthy of discussion. In this study, I compared very stringent mastery criteria: 100% mastery criterion levels across two sessions (at the individual operant level and the set level). The mastery criterion of 100% accuracy across two sessions is

not indicative of the most widely used mastery criterion across ABA practitioners or ABA researchers; however, at least one empirical study found that very stringent criteria (e.g., 100% across 3 replications) were required to produce durable responding (Richling et al., 2019). For the participants in this study, the mastery criterion for programs utilizing learn unit 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 mastery criteria (e.g., 90% across two sessions). Nevertheless, I 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. However, the evaluation was abruptly stopped due to the world-wide 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. There were also large time gaps, sometimes up to three weeks between Zara's acquisition sessions due to prolonged school closures. Thus, her acquisition data may have been affected. Furthermore, due to these prolonged closures, I was only able to assess nine operants under the OA condition and eight operants during the SA condition during 4-week response maintenance sessions.

The results of this study demonstrate the need to continually examine well-established acquisition criterion procedures in our field and to question the effectiveness of our teaching practices (Richling et al., 2019). The rules instructors set during instruction have great effects on performance and should not be implemented without thorough examinations of their efficacy.

Moreover, the rules instructors set should be based on scientific evidence rather than traditions passed down from prior practices. In the realm of mastery criteria, there is a need for systematic and experimental manipulations to identify how different criteria affect response maintenance and other components of mastery. This replication study prioritizes the skill acquisition and response maintenance components of overall mastery. However, future analyses should consider the effects of mastery criterion on response generalization and stimulus generalization – both important goals of instruction.

Table 1

The total number of learn units required

	Patrick		Katie		William		Zara	
	OA	SA	OA	SA	OA	SA	OA	SA
Number of Learn Units	275	500	255	380	330	580	408	660
Mean Learn Units to Criterion per Operant	23	42	21	32	28	48	34	55

Note. The total number of learn units reported were to acquire 12 target operants (Patrick), 12 target operants (Katie), and 12 target operants (William) for each experimental condition. OA represents Operant Condition and SA represents Set Analysis. The mean numbers of learn units to criterion per operant are also displayed. Zara's OA learn units reflected the projected total number to acquire 12 operants based on 370 learn units to master 11 operants. Zara's SA learn units reflected the total number of learn units acquire 12 operants.

Table 2

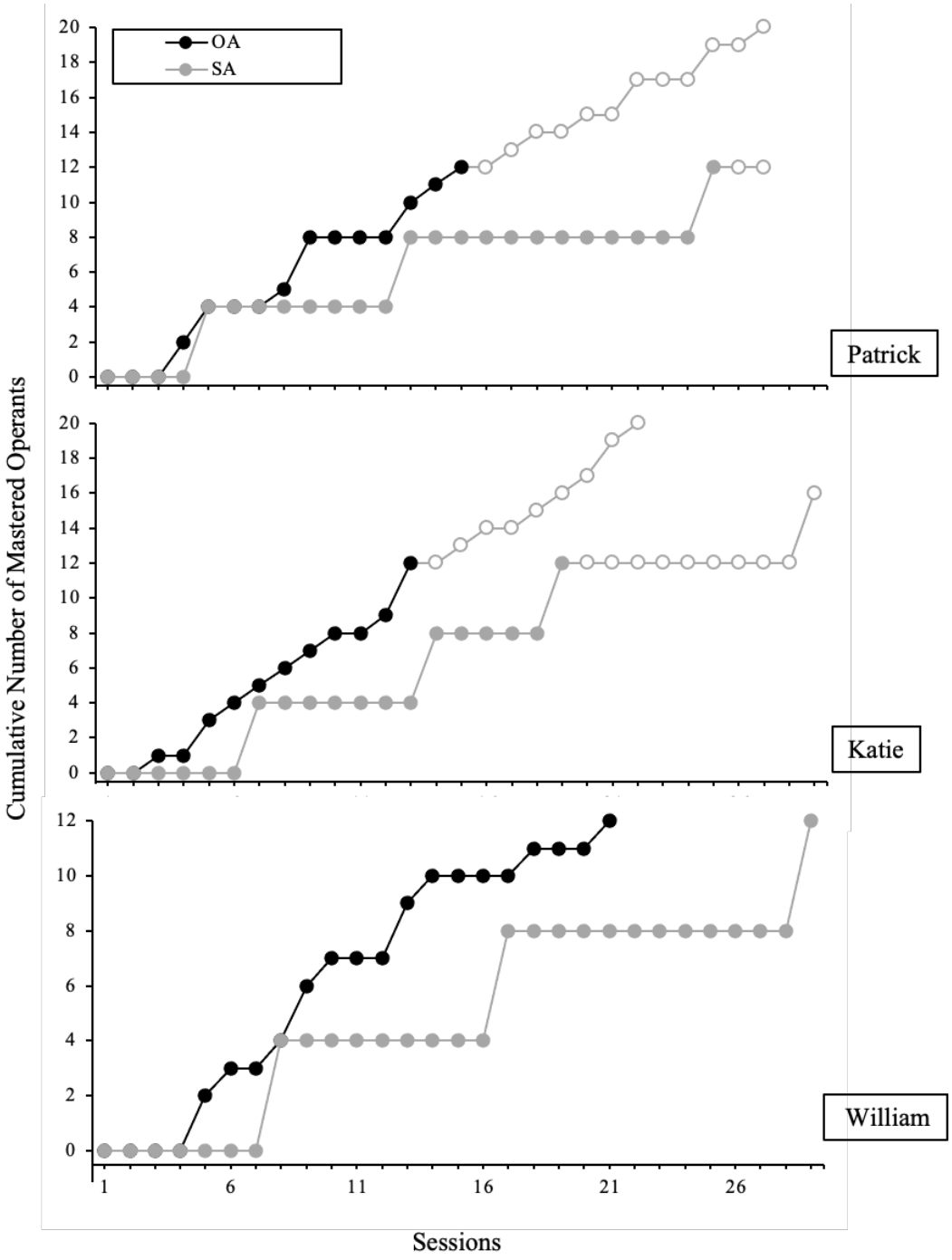
Sessions of potentially unnecessary overtraining trials for the SA condition

	Patrick	Katie	William	Zara
Phase 1 Overtraining Trials (%)	10 (10.0%)	45 (32.1%)	45 (28.1%)	235 (69.1%)
Phase 2 Overtraining Trials (%)	50 (31.3%)	45 (32.1%)	50 (27.8%)	90 (37.5%)
Phase 3 Overtraining Trials (%)	100 (41.7%)	25 (25.0%)	115 (47.9%)	5 (0.1%)
Total Overtraining Trials (%)	160 (27.7%)	115 (29.7%)	210 (34.6%)	330 (35.6%)
Number of Operants Mastered	12	12	12	12
Average Overtraining Trials Per Operant	13	10	18	28

Note. The percentages represent the percent of the total number of learn units in each phase that were potentially unnecessary. To calculate Average Overtraining Trials Per Operant, I divided the Total Overtraining Trials by the Number of Operants Mastered. Each of the three phases represented the opportunity to master four sight words.

Figure 1

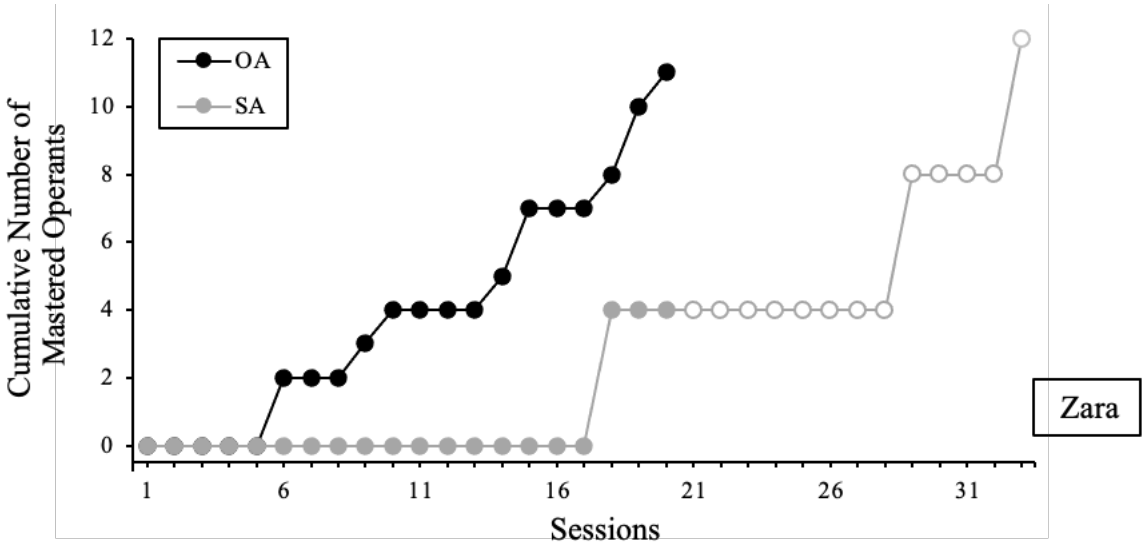
The Cumulative Number of Operants Acquired Under Each Condition



Note. The graphs display the cumulative number of operants (sight words) acquired under operant analysis (black circles) and set analysis (gray circles) mastery conditions. The open circles represent the operants that were not assessed during four-week maintenance sessions.

Figure 2

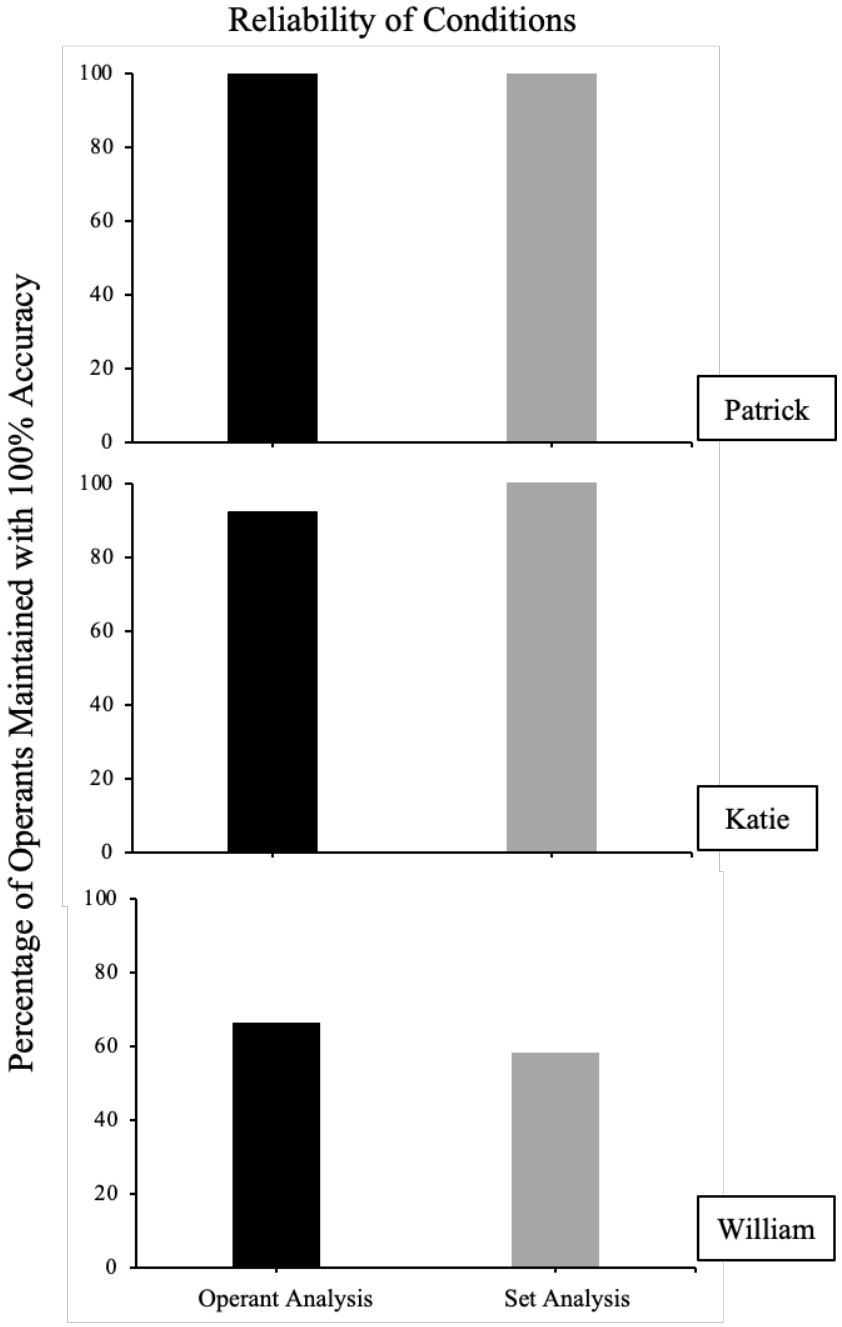
The Cumulative Number of Operants Acquired Under Each Condition for Zara



Note. The graphs display the cumulative number of operants (sight words) acquired under operant analysis (black circles) and set analysis (gray circles) mastery conditions. The open circles represent the operants that were taught after a 17-day gap between the 20th session and the 21st session. There were two additional gaps between instructional sessions, however, the gap affected both Operant Analysis and Set Analysis conditions equally. These two additional gaps are not represented in the graph.

Figure 3

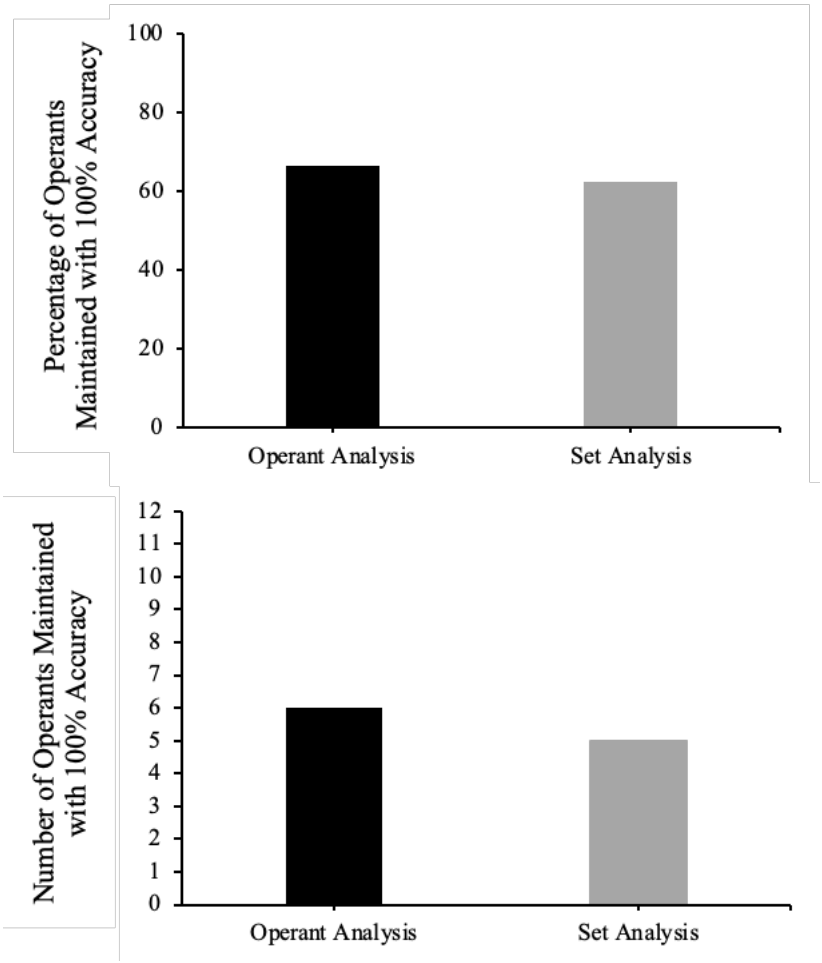
The Percentage of Operants Maintained Under Each Condition for Patrick, Katie, and William



Note. The graphs display response maintenance results four weeks following the acquisition of sight word operants for each participant. The percentage of operants, or sight words, maintained at 100% accuracy are shown.

Figure 4

Response Maintenance Performance Under Each Condition for Zara



Note. The graphs display response maintenance results four weeks following the acquisition of sight word operants for Zara. The percentage of operants, or sight words, maintained at 100% accuracy are shown on the left panel. The total number of operants maintained at 100% accuracy are shown on the right panel.

Chapter 4: A Descriptive Analysis of Mastery Criterion Effects on Response Maintenance and Stimulus Generalization

Abstract

As researchers begin empirically analyzing the effects of skill acquisition performance criteria, or mastery criteria, on response maintenance, it is clear that there are parametric effects of the performance criteria. That literature is limited by small sample sizes combined with large-scale implications. To address this limitation, the current analysis expanded upon those findings by analyzing acquisition criterion practices of Applied Behavior Analysis researchers reported in three peer-reviewed behavioral journals within the past three years. General characteristics of acquisition criterion practices were described. In addition, the analysis targeted the effects of acquisition-criterion levels and frequency of replications on response maintenance results and generalization results. Across different populations, interventions, and teaching tactics, the results highlight that higher acquisition criteria produced higher response maintenance results and higher generalization results. The results also indicated nuanced effects and occasions when acquisition criterion level alone is not sufficient in producing high accuracy results during tests of response maintenance. Future directions and important implications were discussed.

Keywords: mastery, acquisition criteria, mastery criteria, response maintenance, generalization, skill acquisition

A Descriptive Analysis of Mastery Criteria in Applied Behavior Analysis Research

Many educators and researchers in the field of Applied Behavior Analysis (ABA) implement procedures that focus on the acquisition of responses not already in one's repertoire. Skill acquisition programs utilize performance criteria – or mastery criteria – to signal an individual's acquisition of a novel skill and/or when teaching can conclude. The practice of assigning mastery criteria to skill acquisition programs is nearly universal in ABA research (McDougale et al., 2019; Rehfeldt & Ghezzi, 1996). In fact, 100% of Board Certified Behavior Analysts (BCBAs) who participated in a nation-wide survey of ABA practices reported that they use some form of mastery criteria in their programing (Richling et al., 2019). For a practice to be so widely adopted among ABA researchers, educators, and clinicians, there is a disconcerting lack of empirical evidence for the efficacy of specific components of mastery criteria and their effects on skill acquisition and other aspects of responding such as stimulus generalization and response maintenance. To date, there are only five publications that evaluate mastery criteria as an independent variable in skill acquisition (Fuller & Fienup, 2018; Longino et al., 2021; Pitts & Hoeger, 2021; Richling et al., 2019; Wong et al., 2021). With such insufficient research on the practice of mastery criterion, an introduction to the concept of mastery is necessary.

What is mastery? What does it mean to master a skill?

Merriam-Webster defines mastery as an individual who possesses or displays great skill, knowledge, or technique (Merriam-Webster, n.d.). Mastery is synonymous with terms such as expertise, proficiency, achievement, and success and can be defined as possessing or displaying great skill, knowledge, or technique (Merriam-Webster, n.d.). When identifying individuals who possess mastery in various domains, there is a universal acceptance of one defining characteristic – masters of certain skills are objectively the best among their counterparts. In the sports domain,

Tiger Woods and Jack Nicklaus are masters of golf, Bruce Lee is a master of martial arts, and in the biomedical field, Sir John B. Gurden is a master of nuclear transplantation and cloning. In virtually every domain, there are individuals who excel and achieve monumental success, and these individuals are regarded as masters in their respective domains.

Thus, mastery occurs when an individual performs at or beyond a standard benchmark of achievement. In a vast range of skill domains, there are standard benchmarks for performance that indicate a progression of competency. For example, as a taekwondo jeja (student) demonstrates proficiency in specific techniques, they progress through a belt ranking system that denotes a rank of expertise in the sport. In the realm of technology corporations, companies such as Oracle outline categories and subcategories of skill levels (trainee, novice, proficient, expert) to classify employees. The office of human resources at the National Institute of Health also has a proficiency scale to measure skill competency among their employees. These scales and benchmarks function as a guide for the expectations of the top performers in each skill level. They also demonstrate how skill levels belong on a continuum.

"Mastery" in the context of ABA does not always coincide with the conventional definition of mastery, which refers to superior expertise. Within the ABA context, there are four pillars of mastery (Richling et al., in press). Simply stating that a skill is mastered once an individual emits accurate responses during instruction does not cover the range of expected educational outcomes that should be entailed in such a definitive term as "mastery." Acquiring the skill to some pre-determined level of proficiency is only the first pillar of mastery. Once the skill is acquired, the rate at which responses are emitted is another important pillar. Correct responses should be emitted at an adequate rate. The third pillar of mastery is response maintenance, which means the target responses are emitted long after the instruction has

concluded and placed under extinction conditions. The final pillar of mastery is stimulus generalization, which means that the target responses are emitted in completely novel and changing situations.

Theories of Skill Acquisition and Competency

Mastery and expertise development have been explained through different viewpoints and learning theories (Adams, 2011; Dreyfus, 1980; Ericsson, 2009; Hoffman, 1998). Much of the work on expertise development and skill acquisition has been centered around a complex process of learning whereby a learner advances through several phases of growth. Bransford and Schwartz (2009) assert that the process of expertise development is perennial, adaptive, and inherently social.

In the 1970s, Noel Burch developed the "Stages of Competence Model." Within this model of learning, individuals fall into four distinct and sequential stages including: unconsciously unskilled, consciously unskilled, consciously skilled, and unconsciously skilled (Adams, 2011). At the unconsciously unskilled stage, the learner does not know how to perform the skill and does not recognize their own deficiency. The next stage, consciously unskilled, the learner begins to recognize problems and issues in their current level of performance. Then, at the consciously skilled stage, the learner is capable of emitting target behaviors but is not fluent. Much effort is put into a correct response. In the final stage, the unconsciously skilled learner performs the skill effortlessly and naturally. The learner emits correct responses without much effort and emits correct responses while multitasking at the same time.

The Dreyfus Model of adult skill acquisition is another model based on stages that focuses on the learner building on previous experiences to achieve expertise. This learning model has been applied in sectors such as sports, public health, military, computer programming, and

medical practices (Honken, 2013). There are five stages in the Dreyfus Model: novice, advanced beginner, competent, proficient, and expert. In the novice stage, the learner is new to the skill and must learn simple rules for determining actions in a context-free environment. Once the learner acquires the simple rules of the skill, they progress to the advanced beginner stage by which they learn to apply the rules in new contexts and environments. In the third stage, the competent stage, the learner encounters additional situations where the acquired rules may or may not apply. The learner must decide when and where to apply the rules. In the proficient stage, the learner is more confident in different situations and has a repertoire of discriminating among a variety of situations. To reach this stage, the learner must have many experiences and practice. At the final stage, the expert learner performs quickly and accurately. They behave according to natural instinct or intuition.

In an effort to operationally define expertise, Robert Hoffman (1998) developed several developmental milestones of learning. The characteristics of this model are similar to the Dreyfus Model and each stage describes different levels of skill knowledge. In Hoffman's model, the learner begins at the naivete stage with zero knowledge of the skill domain. Eventually, the learner gains more knowledge about the domain and progresses through the, initiate, apprentice, journeyman stages until they reach the expert and master stages. A distinction between the expert and master stages is the master is qualified to teach learners at lower levels.

How does a learner become a master of a skill? Swedish psychologist Anders Ericsson studied experts in a variety of domains and found that the most accomplished musicians accumulated over 10,000 hours practicing their skill before the age of 20. Their 10,000 hours of practice was about 8,000 hours more than amateur musicians of the same age (Kramp &

Ericsson, 2006). This led to a widely popularized 10,000 Hour Rule promoted by Malcolm Gladwell who stated that "10,000 hours is the magic number to greatness" (Gladwell, 2008). This rule unfortunately glosses over the important implications of Ericsson's findings. Ericsson provided a theoretical framework for expert performance, which focused on the importance of deliberate practice, constructive feedback, motivational factors, and social influence (Ericsson, 2006; Wang & Zorek, 2016). Deliberate practice involves the learner working on skills that are outside of their repertoire and are attainable in a short period of time. It is also essential for learners to seek training from teachers and coaches – learners who are farther along on the expertise continuum. Teachers provide constructive feedback that aids in the growth of their learners. Effective teachers also foster the growth of independent learners. When an independent learner becomes a self-teacher, they can monitor their own progress and develop their own plans for success. This method of becoming an expert or master performer has been applied in all aspects of life. Hoffman emphasized that experts are not born as experts; they are made.

Mastery and Skill Acquisition in Education

The established theories of skill acquisition and expertise development are adopted within the education system as well. Within the conventional academic model, students are grouped together by age (not necessarily skill level). A student follows a lesson plan delivered at one general pace, they complete homework, and they take periodic exams that assess knowledge of the concepts. The grades that the student receives throughout the school year dictate their level of knowledge of all the academic subjects they were taught. The traditional grading system in the United States takes numerical and letter forms. Generally, grades from 0%-59% correspond with an F letter grade, 60%-69% correspond with a D, 70%-79% correspond with a C, 80%-89% correspond with a B, and 90%-100% correspond with an A. While this traditional educational

model and grading system have been in place since the early 1940s (Schneider & Hutt, 2014), there can be a problematic issue as they relate to mastery. When a student receives a grade on a test, any percentage point below 100 indicates a gap in the student's knowledge. For example, a student who receives a 90% on a test is missing 10% of the tested content and a student who receives a 75% on a test is missing 25% of the tested content. These two students continue through the curriculum without reconciling the gaps in their knowledge and the gaps can grow wider and wider as the student graduates from each grade.

The field of ABA approaches education and skill acquisition from the position of operant conditioning and the strength of operant behavior. In contrast to the general education model, ABA practitioners do not group individuals together based on arbitrary classifications such as age. Every strategic behavioral intervention is focused on improving socially significant behaviors of singular individuals. Instead of teaching programs at a general pace for a group of learners to follow regardless of whether or not the concepts are acquired, ABA practitioners implement skill acquisition programs until an individual achieves a predetermined criterion for performance, also known as a mastery criterion. Mastery criterion is a foundational part of ABA practice and evidence of mastery criteria can be found in some of the earliest applied research studies in the 1960s (Hall et al., 1968, Johnson & O'Neill, 1973). In a recent survey of 200 board certified ABA practitioners, 100% of respondents use some form of mastery criteria in their practices. Between 2015-2017, 76% of ABA research studies on skill acquisition explicitly reported a form of mastery criteria to indicate when teaching interventions could conclude as a result of the participants' "mastered" performance. Unfortunately, the ubiquity of these practices also highlighted the fragmented nature of mastery criteria application. Given the lack of research behind the most effective form of mastery criteria, it would seem that both ABA researchers and

practitioners alike establish criteria that are not based empirical research and wistfully believe the criteria will produce behavioral momentum in the future – a practice that is not scientific or based on the principles of behavior.

A conceptualization of mastery within operant behavior from a basic behavioral viewpoint may facilitate ABA researchers and practitioners to understand the underlying principles behind mastery criteria, which may lead to more evidence-based practices regarding mastery criteria.

Mastery and Behavioral Momentum Theory

Mastery requires sufficient response strength and stimulus control to be durable across time. During the early stages of behavioral psychology, Thorndike (1913) described response strength as the connection between a stimulus and the occurrence of a response in the presence of that stimulus. In 1938, Skinner stated that an operant's strength is directly proportional to the frequency of emission. Since then, response strength has been predominantly measured by the rate of a response. That is, the number of times a response is emitted within a unit of time corresponds with the strength of the response (e.g., faster rate = stronger response). Nevertheless, there are other dimensions of behavior that provide evidence of response strength (Simon et al., 2020). In the context of skill acquisition, the percentage of correct trials during instruction may also reflect response strength. That is, a learner who emits responses with 100% accuracy demonstrates a strong response strength because the antecedent stimuli evoke accurate responses 100% of the time. Furthermore, when a response contacts reinforcement in the presence of an environmental stimulus, the response is undeniably associated with that stimulus. Thus, the analysis of stimulus control alongside the analysis of response strength is necessary. Behavioral Momentum Theory (BMT) suggests that strength of stimulus control and strength of response is

associated with the rate of reinforcement (Nevin et al., 1983). Rich schedules of reinforcement positively correlate with greater persistence in behavior (Craig et al., 2014; Nevin, 1992). BMT equates the dynamics of operant behavior to Newton's second law of motion, which states that an object's velocity and acceleration is directly related to an object's mass and the force applied to the object. The more mass a stimulus or object has, the more force needed to accelerate an object. In operant behavior, response strength represents the mass of a behavior and it increases with high rates of reinforcement. In the BMT literature, higher behavioral mass leads to behavior that is more resistant to change despite the influence of disrupting external factors such as the termination of teaching or the cessation of reinforcement. In the context of skill acquisition, programming for high behavioral mass (building response strength) is the ultimate goal as this predicts persistence in the face of disruption (e.g., time since teaching, extinction conditions).

The process of acquiring novel responses does not occur in a vacuum. There are temporal features that are interconnected in all behavioral contingencies. Thus, when teaching novel skills, one must consider all the effects of the arrangement of instruction and the dimensions of the current teaching contingency on future behavior. One crucial dimension of the teaching context is the criterion for mastery that instructors establish. Based on BMT, mastery criteria have two functions: proxies for response strength and predictors of persistent behaviors in the face of disrupters.

Mastery criteria are proxies for response strength because if a teacher sets the criterion at 50% accuracy versus 100% accuracy, behavior in the latter condition is likely to have more strength as 100% of antecedents correctly evoke behavior that is reinforced. Thus, behavior with 100% accuracy is more reliably evoked by antecedents and contacts more instances of reinforcement than behavior with a 50% accuracy criterion. As a result, mastery criteria function

to predict the behavior's momentum during periods of disruption such as maintenance and generalization tests. It is interesting to note that results of response maintenance tests have shown that behaviors acquired to a high criterion level persist to some degree despite BMT literature providing evidence of greater probability of extinction for behaviors that have been reinforced on a dense, continuous reinforcement schedule compared to behaviors that have been reinforced on a variable or partial schedule of reinforcement (Lerman & Iwata, 1996; Uhl & Young, 1967).

Different areas of literature in ABA such as Precision Teaching (PT), Comprehensive Application of Behavior Analysis in Schooling (CABAS®), and Personalized System of Instruction (PSI) utilize mastery criteria for skill acquisition. The following sections will describe mastery criteria in applied settings within the framework of behavioral momentum.

Precision Teaching

In the PT literature, the ultimate form of mastery is evident when an individual performs a behavior accurately and with speed (Lindsley, 1971). Fluency, which refers to the number of correct responses emitted per minute within an assessment period (Lindsley, 1991) is the true marker of mastery because it ensures that the individual will emit correct responses in spite of distractions in ever-changing environments (Binder, 1998). In terms of stimulus control, fluent behavior occurs within a short latency after the discriminative stimuli are presented. The likelihood that fluent behavior will change in the face of a disrupter is unlikely because operant behavior that is emitted at a frequent rate and contacts frequent reinforcement is likely to persist over time (Nevin, 1996). Thus, PT literature suggests that fluent behavior produces a number of beneficial outcomes, such as retention, stability, and endurance (Kelly & Holloway, 2015). These characteristics make fluent behavior similar to the *unconsciously skilled* and the *expert* levels of skill acquisition theories proposed by Dreyfus (2004). PT practitioners assert that fluent

performances are automatic to the individual and are performed without effort or hesitation. Thus, the curriculum within the PT method establishes a set time-based mastery criterion that will lead to stronger stimulus control for learners in the long-term. For example, a mastery criterion for addition and subtraction math facts may be 60-70 correct responses per minute. Another aspect of PT includes a large variety of exemplars for learners to use to ensure an adequate number of learning opportunities available for learners to achieve fluency. An issue with behavioral fluency, however, is the lack of research supporting the effectiveness of fluency training in persisting behavior in the face of disrupters such as tests of generalization because of the restricted stimulus control developed through training (Doughty et al., 2004; Meindl et al., 2013).

CABAS®

The primary technology of instruction within the CABAS® model of education is the learn unit. The learn unit is the smallest, most fundamental measure of teaching (Greer & McDonough, 1999). It measures both teach and student behaviors because it consists of at least two 3-term contingencies between the teacher and the student. Learn units are delivered during all instructional programs until the target skill is under the natural contingencies for the student (Greer et al., 1999).

To produce mastery and fluency of skills, there is an emphasis on establishing a high-level performance – or high acquisition criterion level – for each instructional program. Within CABAS®, the standard mastery criteria for the majority of academic, communication, and self-management domains are 90% accuracy across two sessions or 100% accuracy across one session (Greer & Ross, 2008). For tests of verbal behavior cusps and capabilities, the mastery

criterion is 80% accuracy across one probe session. The high level of accuracy is considered to relate to the strength of stimulus control, which is also supported by BMT (Nevin, 1992).

Personalized Systems of Instruction

Historically, PSI has been implemented primarily with university students. This particular model of instruction is based on student mastery of a series of quizzes within a unit of content. According to Keller (1968), a student was required to emit correct responses with 100% accuracy on each individual quiz prior to moving on to the next quiz in the unit. Then, a student was required to respond with 100% accuracy on the terminal unit quiz before moving on to the next unit. Students progressed at their own pace and exited their courses with complete mastery of the contents that were taught. Mastery criteria functionally controlled the academic performance of college students (Johnson & O'Neill, 1973; Semb, 1974) and higher mastery criteria within PSI produced optimal results with adult undergraduate students. This finding also extended to skills beyond the academic realm, including health-related fitness (Hannon et al., 2008; Pritchard et al., 2012) and sports (Cregger & Metzler, 1992). Furthermore, PSI research provided evidence for the function of mastery criteria predicting future performance during test disruption (i.e., response maintenance probe sessions and generalization probe sessions). Zencius et al. (1990) found that students who achieved a 100% accuracy criterion performed at the same high levels during four- and 10-week follow-up sessions and tests of generalization across novel setting.

Dimensions of Mastery Criterion in Applied Settings

There are many models of skill acquisition in ABA, particularly in educational settings, and the application of mastery criteria plays a large role in each existing model. A major similarity among mastery criteria practices across the educational models of instruction, is the

requirement of high accuracy levels across a particular number of replications (Fuller & Fienup, 2018; Richling et al., 2019). There are two main dimensions of mastery criteria at play. The accuracy level represents the proportion of antecedents that evoke behavior. More specifically, an 80% accuracy criterion means 80% of the antecedents presented evoked the target responses. The second dimension is the frequency of replications at which the target responses are emitted, which represents the duration that responses are required to be emitted at a particular strength level. McDougale et al (2019) analyzed over 150 research articles in major ABA journals between 2015 and 2017 and found that the most commonly applied mastery criteria used in ABA research were session-based. More specifically, the majority of ABA research studies (54%) utilized a particular percentage of accurate trials within a session to determine mastery. Less than 1% of the research articles utilized a certain rate of response per unit of time to determine mastery. Of the articles that reported a session-based mastery criterion, the majority used a percentage that was between 90%-99% accuracy. Additionally, the majority of ABA researchers require two replications at a specific level of accurate performance to determine mastery. These results contrasted with the results of a survey delivered to 200 Board Certified Behavior Analysts (BCBAs) on their mastery criteria practices (Richling et al., 2018). According to the survey results, the majority of BCBAs reported that they applied an accuracy level of 80% for their mastery criteria. Furthermore, the BCBAs required the accuracy to be observed across three sessions to determine mastery. These variations of mastery criteria practices between ABA researchers and ABA practitioners call for more experimental evaluations of mastery criteria. Furthermore, it is necessary to conduct more research on the effects of mastery criteria, not just on skill acquisition but also with response maintenance and response generalization, which encompass true mastery.

Much of the existing parametric empirical research that analyzed the application of mastery criteria on student learning, response maintenance, and generalization exist with undergraduate students in the college setting. When mastery criteria were applied at low, medium, and high rates of accuracy, student performance changed as a function of the different rates of criteria (Johnson & O'Neill, 1973). More specifically, students performed better on tests under the high mastery criteria condition compared to the low mastery criteria condition. Similarly, Semb (1974) evaluated the effects of mastery criteria levels and length of assignments on student performance. The author found that a 100% accuracy criterion in conjunction with short assignments was most effective in predicting the best student outcomes. Recently, this area of research extended into a novel type of student performance: derived relations. Not surprisingly, researchers found that students performed better when higher, more stringent mastery criteria were implemented (Brodsky & Fienup, 2018). Ultimately, the existing research provides evidence that students perform better and retain more skills when they are taught under conditions of high accuracy criteria.

Recently, researchers have examined the effects of acquisition criterion levels with students with developmental disabilities. Fuller and Fienup (2018) conducted important and necessary comparisons between three levels of skill acquisition mastery criteria. The authors taught a set of spelling responses and manipulated the targeted response strength by applying a 50%, 80%, or a 90% level of accuracy in each condition. In other words, the three levels of accuracy correlate with the proportion of antecedents that evoke behavior. When the authors implemented extinction procedures four weeks after the initial acquisition of the spelling responses, the results demonstrated a clear difference in response maintenance accuracy across the three conditions. All participant behaviors persisted with greater strength for words acquired

under the 90% mastery condition (when 90% of the antecedents evoked the behavior). Behavior levels dropped during the extinction procedures under the 50% and 80% mastery conditions. The findings of Fuller and Fienup suggest that when the frequency of replications are held constant, a higher level of mastery criterion is necessary in order to produce durable responses in the future.

Similarly, Richling et al. (2019, Experiment 2) compared 60%, 80%, and 100% acquisition-criterion levels during skill acquisition training. The authors also manipulated the duration element to the proportion of antecedents evoking behavior by requiring the accuracy to occur three consecutive sessions for teaching auditory-visual conditional discrimination tasks. The results of this evaluation support the efficacy of a higher level of criteria for skill acquisition interventions. Richling et al. (2019, Experiment 2) found that only the 100% x 3 condition produced response maintenance at over 80% accuracy. The 60% and 80% conditions produced lower accuracy. Richling et al. (2019, Experiment 4) compared an 80%. 90%, and 100% criterion level. In this study, the results showed that even the 90% criterion level was not sufficient in producing adequate response maintenance results, which differed from Fuller and Fienup's (2018) finding that 90% accuracy was adequate in producing durable responses in the future. However, recently Pitts and Hoerger (2021) replicated Richling et al.'s study and found that both 90% x 3 and 100% x 3 criterion conditions produced maintenance at 90% accuracy or higher and only the 80% x 3 condition produce inadequate maintenance. The results of Longino et al. (2021) also found that the 90% accuracy across 3 session acquisition criterion and the 100% accuracy across 3 session acquisition criterion led to the highest response maintenance one month later.

Some of the differences in results may be due to procedural differences of the intervention sessions but nevertheless, more studies are necessary in order to gain a better

understanding of the effects of acquisition-criteria on skill acquisition as a whole. The handful of studies described above (Fuller & Fienup, 2018; Longino et al., 2021; McDougale et al., 2019; Pitts & Hoerger, 2021; Richling et al, 2019), took important first steps in examining acquisition-criteria (mastery criteria) as an independent variable. However, a limitation of these studies includes the small sample sizes, which may bring into question the external validity of the findings. Moreover, there are still many more aspects of acquisition-criteria and skill acquisition to investigate. To start, there needs to be more information regarding the effects of acquisition criteria on other aspects of skill acquisition in addition to response maintenance. Thus far, there has been little to no research on the effects of acquisition criteria on the generalization of acquired skills in novel settings, across novel stimuli, or novel individuals. Acquisition criteria may also produce differential effects depending on the number of opportunities to respond within a session. Another important area of examination is the type of skill targeted in the teaching interventions. Perhaps, a higher level of acquisition-criteria is necessary for different types of skills (e.g., skills involving the safety of an individual). Acquisition criteria may also vary depending on the learner's level of verbal behavior.

It is clear that many questions regarding acquisition criteria remain. Thus, the purpose of this study is to systematically analyze skill acquisition articles published in *Journal of Applied Behavior Analysis* (JABA), *Behavioral Interventions* (BIN), *and Behavior Analysts in Practice* (BAP) between the years 2017-2019. I chose to include the years 2017 through 2019 to extend the findings of McDougale et al. (2019), which analyzed articles from 2015-2017. Data collected on specific components of each published article included the type of mastery criteria implemented (if reported), the structure of the intervention sessions, maintenance results (if reported), generalization results (if reported), the target skill taught, and developmental

information of the participants (age and diagnoses if applicable). Using those data, descriptive analyses were conducted to answer the following research questions:

1. What are the general characteristics of skill acquisition articles published in JABA, BIN, and BAP between 2017-2019 in terms of acquisition criteria, skill categories, and maintenance/ generalization results?
2. Do acquisition criteria have an effect on response maintenance?
    a. Is there an association between the specific acquisition-criterion level and the percentage of accurate responses during maintenance assessments?
    b. Is there an association between the specific acquisition-criterion level across a specific number of replications and the percentage of accurate responses during maintenance assessments?
3. Do acquisition criteria have any effects on stimulus and response generalization?
    a. Is there an association between the specific acquisition-criterion level and the percentage of accurate responses during generalization assessments?
    b. Is there an association between the specific acquisition-criterion level across a specific number of replications and the percentage of accurate responses during maintenance assessments?

Method

Inclusion Criteria

I chose to include articles from three behavior analytic, peer-reviewed journals which were likely to include behavioral acquisition procedures derived from behavior analytic procedures. The journals included were the, *Journal of Applied Behavior Analysis* (JABA), *Behavioral Interventions* (BIN), and *Behavior Analysis in Practice* (BAP). Every article

published within these three journals between the years 2017 to 2019 was considered for this descriptive analysis. Each article was then reviewed and coded for several variables. First, articles were reviewed to determine if the focus was primarily on skill acquisition interventions. Only articles fitting this criterion were included in this descriptive analysis. I defined skill acquisition interventions as any procedure targeting the acquisition of one or more novel skills or any procedure specifically aimed at increasing or improving the accuracy of one or more behaviors for human participants. Studies that specified the purpose of decreasing disruptive behavior, problem behavior, self-injurious behaviors, or stereotypic behaviors were not included in this descriptive analysis. Systematic reviews, meta-analyses, discussion articles, and technical articles were also not included.

Article Selection

For the years 2017 and 2019, I manually searched the articles in each journal using the Columbia University Libraries search engine, which permitted access to every full-length article in each journal. Figure 1 summarizes the process of article selection, including the number of articles identified in each step. For each issue of the three journals, I downloaded all articles and subsequently read each article's title and corresponding abstract to determine if the intervention targeted skill acquisition. I targeted key phrases such as, "skill acquisition," "improving," "increasing," "teaching/ taught," etc. If the description in the abstract was not clear, I read the dependent variable section for further clarification in order to determine inclusion for the descriptive analysis. There were 234 research articles published in JABA between the years 2017-2019, and this initial inclusion review process yielded 82 skill acquisition articles, which were then included in the descriptive analysis. Of the 134 research articles published in BIN between the years 2017-2019, this process yielded 45 skill acquisition articles. Of the 219

research articles published in BAP between the years 2017-2019, this process yielded 85 skill acquisition articles for inclusion in the analysis. Following this inclusion review stage, I further eliminated research articles that did not include an explicitly stated acquisition criterion, also known as mastery criterion, for the acquisition intervention. I defined an explicitly stated acquisition criterion as any statement that identified the parameters for stopping the final teaching phase and beginning the post-training phase of the study. This process led to the exclusion of 424 research articles resulting in a total of 163 skill acquisition articles in the current descriptive analysis. This included 73, 36, and 54 articles from *JABA, BIN,* and *BAP*, respectively. Once all skill acquisition articles were identified, I collected data on eight variables.

Data Extraction

Target Skill

First, I identified the target skill (dependent variable). Next, I identified the specific acquisition criterion used in the teaching intervention (percentage of correct responses, number of correct consecutive trials, duration of target behavior, etc.). I collected data per case rather than per article. This was done to account for instances of different mastery criteria assigned to participants, for instances of different mastery criteria assigned for multiple target skills, and to account for the differential maintenance and generalization results for each participant. For example, there were four participants in a study conducted by Gallant et al. (2017). The authors reported an acquisition criterion of 100% accuracy in one session for three participants and a 100% accuracy across two sessions criterion for one participant. This article also targeted one skill in the maintenance assessment and one skill in the generalization assessment. Thus, I analyzed four cases of data for this article in the maintenance analysis and four cases of data in the generalization analysis with three cases falling under one specific mastery criterion and one

in another. In another example, Pachis and Zonneveld (2019) studied three participants. In the tests of maintenance, two target skills were targeted. As a result, I analyzed six cases of data for this article in our maintenance analysis (one case per participant, per each of the two target skills assessed).

Participants

I collected data on the age of each participant and categorized them into the following categories: (a) 12 years or younger, (b) adolescent (13-17 years old), (c) college students (explicitly stated in participant section as undergraduate students), (d) adult (18-64 years old), (e) older adult (65 years or older). I also collected data on the diagnoses of the participants (if applicable). Additional demographic data were not obtained for the current analysis.

Skill Categories

I categorized each skill acquisition into one of the following skill categories: a) academics, b) academic related engagement behavior, c) functional life skills, d) safety skills, e) social communication, f) sports/ physical activity, g) treatment implementation, h) elementary verbal operants. Academic skills were defined as skills directly related to school subjects or specific content areas such as reading, writing, math, science, history, etc. Academic related engagement behaviors were defined as skills related to increasing attention and on-task behaviors. Functional life skills were defined as skills related to daily living and improving quality of life. Safety skills were defined as skills related to the safety and well-being of the participants. Social communication skills were defined as skills aimed to improve social interactions with adults and peers, including increasing eye-contact and increasing the number of initiations to converse. Skills that were categorized into social communication also included complex mands (e.g., manding for information), and complex tacts (e.g., tacting what peers are

sensing). Sports/ physical activities were defined as skills related to sports and athletics. Treatment implementations were defined as skills related to delivering an intervention or tact to another group of individuals. Elementary verbal operants included basic tacts, mands, intraverbals, and echoics.

Acquisition Criteria

There were 136 articles that reported using a specific percentage of accuracy as a criterion for skill acquisition or "mastery." For each of these articles, I extracted the specific criterion percentage and categorized it into one of the following ranges: a) under 80%, b) 80%-89%, c) 90%-99%, d) 100%. If any article utilized multiple acquisition-criterion levels, I only categorized one criterion level. I chose to categorize the lowest level criterion level for this analysis because it was the most conservative approach. In addition, for this specific analysis only, I did not include Fuller and Fienup (2018) and Richling et al. (2019) into the count because these studies systematically evaluated three different acquisition-criterion levels as their independent variable within their studies.

Identification of Maintenance Assessments

For this variable, I reported the time frame in which the assessments were conducted following the participants achieving the mastery criteria for the intervention. If one maintenance assessment was conducted three weeks following the conclusion of the intervention, I reported the time frame as 1 three-week follow-up. I coded, "none reported" if authors did not explicitly state they conducted a test(s) of maintenance**.**

Maintenance Results

For articles that explicitly reported assessments for maintenance, I coded the specific results for each participant and for each skill specified by the authors. When there were instances

of multiple maintenance session data, I calculated the mean score for each target skill of each participant.

The mastery criteria were categorized by level of accuracy (under 80%, 80%-89%, 90%-99%) and response maintenance results were categorized by level of accuracy (under 80%, 80% and above, 90% and above). Mastery criteria were also categorized by the frequency of replications required (i.e., one session, two sessions, or two or more sessions). After all the target skills were identified, they were grouped into the eight skill categories: a) Academics, b) Engagement (academic related behaviors), c) Functional life skills, d) Safety Skills, e) Social communication, f) Sports/ Physical activities, g) Treatment implementation, h) Elementary Verbal Operants.

*Identification of **Generalization** Assessments*

For this variable, I reported the type of generalization assessments that were conducted following the participants achieving the mastery criteria for the intervention (i.e., generalization across settings, across instructors, across behaviors, across peers, and across stimuli). I coded "none reported" if authors did not explicitly state they conducted generalization assessments.

***Generalization** Results*

For articles that explicitly reported assessments for generalization I coded the specific results for each participant and for each generalized skill specified by the authors. When there were instances of multiple generalization session data, I calculated the mean score of each target skill of each participant.

The mastery criteria were categorized by level of accuracy (under 80%, 80%-89%, 90%-99%) and generalization results were categorized by level of accuracy (under 80%, 80% and above, 90% and above). Mastery criteria were also categorized by the frequency of replications

required (i.e., one session, two sessions, or two or more sessions). After all the target skills were identified, they were grouped into the eight skill categories: a) Academics, b) Engagement (academic related behaviors), c) Functional life skills, d) Safety Skills, e) Social communication, f) Sports/ Physical activities, g) Treatment implementation, h) Elementary Verbal Operants.

Interobserver Agreement

To obtain interobserver agreement (IOA), a second, trained, and independent observer extracted data on all variables for all the skill acquisition articles acquired. After data were extracted and coded from both the independent observer and I. Then I highlighted any and all discrepancies among the variables for each case. At this stage, IOA was calculated by dividing the number of agreements by total cases multiplying by 100. There was a total of 823 cases of data during the data extraction process. Initial IOA was 90.77% (747 cases of agreement and 76 cases of disagreement).

To ensure accurate data analysis, a second stage of IOA corrected all disagreements. A separate meeting took place between the primary author and another independent observer to discuss the discrepancies. The independent observer then read through the articles with disagreements and extracted the data for the relevant variables. The extracted data from this independent observer were used to settle the disagreements between the first 2 data collectors. This additional step led to 100% agreement on all variables for all articles.

Results

General Characteristics

Across three behavior analysis journals between the years of 2017-2019, there were 587 research articles with 212 (36%) articles targeting skill acquisition specifically. Of the 212 skill acquisition articles published, 163 (77%) articles explicitly reported an acquisition criterion.

Across these articles, researchers reported forms of acquisition criteria that included a percentage of correct responses (83.4%), a consecutive number of correct trials (4.9%), duration (0.6%), responding within an established threshold (1.8%), a particular trend of responding (i.e., stable or increasing) (3.1%), or a correct response on the first opportunity (6.1%). Table 1 reports the percentage of articles targeting each skill across the 163 skill acquisition articles. A high percentage of articles targeted academics (19.8%), followed by treatment implementation (18.4%) and social communication skills (17.9%).

The majority of the skill acquisition articles that reported an acquisition criterion utilized a percentage-based criterion (136 out of 163 articles, or 83.4% of articles). The acquisition criterion percentages ranged from 60% accuracy to 100% accuracy. There was a relatively even proportion of articles utilizing acquisition-criterion levels between the 80%-89%, 90%-99%, and 100% accuracy (see Figure 2). Approximately 30% of articles reported specific acquisition-criterion levels in each of the three ranges (80%-89%, 90%-99%, and 100%) and 2% of the articles reported a criterion under 80% accuracy. The remainder of the analyses focused on the number of cases that reported percentage correct acquisition criteria.

Acquisition Criteria and Response Maintenance

Eight-four of the 163 skill acquisition articles (51.5%) that reported an acquisition criterion assessed for and reported response maintenance results. In these articles,there were a total of 535 cases reported. For my analyses, the *maintenance outcomes* were categorized into the following accuracy groups: (a) under 80%, (b) 80% or above, (c) 90% or above. I counted and summed all the *cases* within each acquisition-criterion category that produced the maintenance results identified above.

Effect of Acquisition-Criterion Level

In Figure 3, I reported the percentage of cases that produced response maintenance outcomes under 80%, between 80%-89%, and 90%-100% (y-axis) as a function of the acquisition-criterion level when the number of replications was held at one (x-axis). That is, when the frequency of replications is held at one, the data represent the influence of criterion-level alone. The x-axis reports the four acquisition-criterion levels. The white bars (top panel) show that 100% of the cases with an acquisition-criterion level under 80% produced response maintenance results of *under 80%.* The percentage of cases that produced under 80% response maintenance (y-axis) decreased as the acquisition-criterion level increased (x-axis). The gray bars (middle panel) demonstrate the opposite effect when examining the percentage of cases producing *80% or above response maintenance* with the percentage of cases increasing as the acquisition-criterion levels increased. For cases that utilized a mastery criterion between 90%-99%, over 90% of the cases produced response maintenance results at or above 80%. The black bars (bottom panel) also show a positive relationship between the levels of acquisition criterion and the percentage of cases that produce *90%-100% response maintenance*. Overall, the black bars, show a fewer percentage of cases in the 80%-89%, 90%-99%, and the 100% acquisition-criterion levels that produced response maintenance results of 90% or above compared to the gray bars. There were no cases with an acquisition criterion level under 80% that produced response maintenance responses at or above 80% accuracy.

A chi-square test of independence was performed to examine the relation between acquisition-criterion levels and response maintenance outcomes. The relation between these variables was significant, X^2 (3, N = 451) = 56.48, p = 0.00. Table 2 reports the observed counts and the expected counts of the total cases of response maintenance at or above 80% (termed "high" maintenance for this analysis) and the observed counts and the expected counts of the

total cases of response maintenance below 80% accuracy during maintenance tests (or, "low" maintenance). The expected counts represent the frequency of cases that are expected, on average, if acquisition-criterion level and response maintenance outcome were independent variables. The adjusted residuals greater than the absolute value of 1.96 indicate significant differences between observed and expected values. For the cases that utilized an acquisition-criterion level of less than 80% accuracy, there were many more cases that led to low maintenance outcomes compared to the expected count. There were also fewer cases that led to high maintenance outcomes compared to the expected count. Similarly, for the cases that utilized an acquisition-criterion level of 80-89% accuracy, there were more cases that led to low maintenance outcomes compared to the expected count and fewer cases that led to high maintenance outcomes compared to the expected count. The opposite association is demonstrated for cases that utilized an acquisition-criterion level of 90-99% and 100%. There were fewer cases that led to low maintenance outcomes compared to the expected count and more cases that led to high maintenance outcomes compared to the expected count.

These outcomes suggest parametric effects of acquisition criterion-level on response maintenance in accordance with empirical studies (Fuller & Fienup, 2018; Pitts & Hoerger; 2021; Richling et al., 2019). As criterion-level increases, there is a systematic increase in adequate levels of response maintenance (80% or higher) and systematic decrease in inadequate levels of response maintenance (below 80%). However, when moving from 80% or higher to 90% or higher threshold for response maintenance, there is a decrease in the effects of criterion-levels. This decrease suggests that criterion-level is necessary, but not sufficient for producing very high levels of response maintenance. Comparing Figure 3's middle and bottom panels examines how criterion levels produce different outcomes depending on how one defines

"acceptable" response maintenance outcomes. Across both panels, there is a positive relation between criterion level and the percentage of cases producing 80% or higher maintenance (middle panel) or 90% or higher maintenance (bottom panel). However, data in the bottom panel are generally at a lower level. Additionally, there was no statistical difference between the 90%-99% acquisition-criterion levels for producing high maintenance results.

Effects of Acquisition-Criterion Level and Frequency of Replications

Figure 4 displays the percentage of cases that produced response maintenance outcomes under 80%, 80% or higher and 90%-100% (y-axis) as a function of the acquisition-criterion level when the number of replications was held at one and at two or more (x-axis). I analyzed the same acquisition-criterion levels described above and compared the results of the levels reported across one replication and across two or more replications. The outcomes in Figure 4 largely reflect those displayed in Figure 3. To examine the effects of frequency (a criterion level observed in one session v. replicated in two or more), I compared the 1-frequency and 2+ frequency for any given criterion level. Some notable differences are observed. When examining cases with inadequate maintenance (below 80% accuracy, top panel), adding multiple replications to the criterion level of <80% produced fewer inadequate maintenance outcomes; however, with other criterion levels outcomes were roughly the same. The middle panel displays the percentage of cases with maintenance at or above 80%. A similar effect is seen with the <80% criterion level – a 20 percentage point increase in adequate maintenance is observed when multiple replications of the criterion level are required. The bottom panel displays a more stringent threshold for adequate maintenance (90% or above) and demonstrated that multiple replications for 80%-89% and 90%-99% produces more cases obtaining adequate maintenance, but this did not extend to the most stringent criterion level of 100%.

Acquisition Criteria and Stimulus Generalization

Seventy-six of all the 163 skill acquisition articles (46%) that reported an acquisition-criterion assessed for and reported generalization results. In these articles, there were a total of 388 cases reported. For my analyses, the generalization results were categorized into the following accuracy groups: (a) under 80%, (b) 80% or above (c) 90% or above. I counted and summed all the cases within each acquisition-criterion category that produced the generalization results identified above.

Effect of Acquisition-Criterion Level

In Figure 5, I reported the percentage of cases that produced generalization outcomes under 80%, at or above 80%, and 90%-100% (y-axis) as a function of acquisition-criterion level when the number of replications was held at one (x-axis). When the frequency of replications is held at one, the data represent the influence of criterion-level alone. The x-axis reports the four acquisition-criterion levels (under 80%, 80%-89%, 90%-99%, and 100%). The white bars (top panel) show that there were zero cases with an acquisition criterion level under 80%. For cases with an acquisition-criterion level between 80%-89% accuracy, 31% produced response generalization results of *under 80%.* The percentage of cases that produced under 80% response generalization (y-axis) decreased as the acquisition-criterion level increased (x-axis). The percentage of cases that produced under 80% response generalization (y-axis) decreased as the acquisition-criterion level increased (x-axis). The gray bars (middle panel) demonstrate the opposite effect when examining the percentage of cases producing 80% or above generalization with the percentage of cases increasing as the acquisition-criterion levels increased. For cases that utilized a mastery criterion between 90%-99%, over 80% of the cases produced generalization results at or above 80%. The black bars (bottom panel) also show a positive relationship between the levels of acquisition-criterion and the percentage of cases that produce

90%-100% generalization results. Overall, the black bars show a fewer percentage of cases among the 80%-89%, 90-%-99%, and the 100% acquisition-criterion levels that produced generalization results of 90% or above compared to the gray bars. There were no cases with an acquisition-criterion level under 80% that produced generalization responses at or above 80% accuracy.

A chi-square test of independence was performed to examine the relation between acquisition-criterion levels and generalization outcomes. The relation between these variables was significant, X^2 (3, N = 306) = 36.05, p = 0.00. Table 3 reports the observed counts and the expected counts of the total cases of generalization at or above 80% (termed "high" degrees of generalization for this analysis) and the observed counts and the expected counts of the total cases of generalization below 80% accuracy during generalization tests (or, "low" degrees of generalization). The adjusted residuals greater than the absolute value of 1.96 indicate significant differences between observed and expected values. For the cases that utilized an acquisition-criterion level of less than 80% accuracy, there were many more cases that led to low degrees of generalization outcomes compared to the expected count. There were also fewer cases that led to high degrees of generalization outcomes compared to the expected count. Similarly, for the cases that utilized an acquisition-criterion level of 80-89% accuracy, there were more cases that led to low degrees of generalization outcomes compared to the expected count and fewer cases that led to high degrees of generalization outcomes compared to the expected count. The opposite association is demonstrated for cases that utilized an acquisition-criterion level of 90-99% however, the difference between the observed counts and the expected counts is less apparent and ultimately, not significant. For cases that utilized an acquisition-criterion level of 100%, there were fewer cases that led to low degrees of generalization outcomes compared to the

expected count and more cases that led to high degrees of generalization outcomes compared to the expected count.

Similar to the parametric effects of acquisition-criterion level demonstrated with response maintenance responding, as criterion level increases, there is a systematic increase in the adequate degrees of generalization responding (80% or higher) and a systematic decrease in inadequate degrees of generalization (below 80%). There was also a decrease in the effects of criterion-levels when moving from 80% or higher results to 90% or higher results of generalization. Less than 72% of the 100% acquisition criterion cases produced generalization results at or above 90%. The generally lower level of cases shown in the bottom panel suggests that criterion-level alone may not be sufficient for producing high degrees of generalization accuracy. Furthermore, there were no statistical differences between 90%-99% and 100% criterion levels or <80% and 80%-89% criterion levels.

Effects of Acquisition-Criterion Level and Frequency of Replications

Figure 6 displays the percentage of cases that produced generalization results under 80%, 80% or above, and 90%-100% (y-axis) as a function of the acquisition-criterion level when the number of replications was held at one and at two or more (x-axis). I analyzed the same acquisition-criterion levels and compared the results of the levels reported across one replication and across two or more replications. Overall, the outcomes shown in all three panels are similar to the corresponding outcomes in Figure 5. When comparing the 1-frequency and the 2+ frequency for any given criterion level, the outcomes are roughly the same. In the middle panel, the criterion levels of 90%-99% and 100% produced very similar outcomes as well.

Discussion

The analyses of the current study report important information regarding skill-acquisition criteria in the field of ABA research. Surveys have found that acquisition criteria are widely used by both ABA practitioners and ABA researchers and there are variations of acquisition criteria that are utilized (Love et al., 2009; McDougale et al., 2019; Richling et al., 2019). While empirical studies of acquisition criterion effects are emerging, they are limited in terms of sample sizes (e.g., Fuller & Fienup, 2018; Pitts & Hoerger, 2021; Richling et al., 2019; Wong et al., 2021). This study addressed the limitations of the empirical studies by examining acquisition-performance-criterion effects across a diverse set of skill acquisition articles published in three prominent ABA journals in an attempt to examine the external validity of the empirical research to date. Largely, the outcomes of this study corroborate the empirical findings across a large set of articles and specific cases for maintenance outcomes and extend the literature by finding comparable criterion effects on stimulus generalization.

General Characteristics

Overall, the results of the general acquisition-criteria characteristics are in line with McDougale's descriptive analysis of articles published between 2015 and 2017 and the research practices within this topic have not changed drastically. The percentage of published skill acquisition articles increased slightly from 33% to 36% and the percentage of skill acquisition articles that reported an acquisition criterion increased slightly from 74% to 77%. One noticeable difference is evident in the percentage of articles reporting a percentage-based acquisition criterion for the teaching interventions. A percentage-based acquisition criterion was the most commonly utilized criterion form among researchers with 54% of articles reporting a percentage-based criterion in 2015-2017. That number increased to 83%, which means more researchers are

relying on a level of accuracy to determine skill acquisition as opposed to a certain number of correct trials in a row, a rate of response per unit of time, or another form of acquisition criterion. The percentage of articles that utilize an 80%-89%, a 90-99%, or a 100% acquisition criterion is relatively similar. These data dispute McDougale's (2019) evidence that researchers tend to use stringent acquisition-criterion levels. In fact, the current study's results provide evidence that ABA researchers and ABA practitioners utilize similar levels of acquisition criteria (80%-89%). Given the results of four recent experimental comparisons of acquisition-criterion levels (Fuller & Fienup, 2018; Pitts & Hoerger, 2021; Richling et al., 2019), it is disappointing that there are not more articles utilizing stringent mastery criterion levels and ABA researchers should reevaluate the level of acquisition-criterion they utilize. However, those studies are still relatively new and perhaps more researchers in the future will consider utilizing stringent acquisition criteria.

Acquisition Criteria and Response Maintenance

About half of the research articles targeting skill acquisition did not assess participants for response maintenance, which is problematic because there is no measure for how persistent the acquired skill is after the conclusion of the teaching intervention. If the goal of ABA practices is to target socially and educationally significant behaviors of individuals (Baer et al., 1968), then long-term assessments of the changed behavior are critical.

According to BMT, a disrupter is any environmental event that has the potential to change some dimension of the behavior being measured (Nevin, 2011). Response maintenance assessments are a potential disruptor event where a researcher can measure degrees of persistence in the face of that disruptor. A primary difference between teaching conditions and response maintenance assessments is the presence of reinforcement and time since teaching

ceased. Overall, the associations between acquisition-criterion levels and response maintenance results show that higher acquisition-criterion levels will reliably produce higher accuracy during response maintenance assessments, which supports the experimental findings of Fuller and Fienup (2018), Richling et al. (2019), Pitts and Hoerger (2021), and Semb (1974). There are no statistical differences between the effectiveness of an acquisition criterion of 90%-99% with an acquisition criterion of 100% when identifying high maintenance results. The results also provide evidence that the proportion of antecedents that evoke behavior plays a role in persistent behavior in the face of extinction experiences. Furthermore, the results support existing BMT literature that have found a positive correlation between rich rates of reinforcement and the degree of persistent behavior (Nevin, 1992). While the results illustrate an overall positive correlation between the percentage of cases with a higher acquisition criterion and high response maintenance accuracy, there is a drop in the percentage of cases that produce high response maintenance threshold (80% or greater and 90% or greater) when acquisition-criterion levels increase from the 90%-99% category to the 100% category. When the response maintenance threshold is at 90% or greater, there is a decrease in the effectiveness of criterion levels alone. These data suggest there is another variable that comes into play when training for persistent behaviors. This calls for future research on the interaction between criterion levels and specific populations, specific levels of verbal behavior of the participants, specific types of behavior, and specific interventions. The results have important implications for practitioners and what they decide are adequate levels of mastery during long-term assessments. If 80% response maintenance accuracy is adequate for the individual, then an acquisition-criterion level between 90%-100% will likely produce successful outcomes. The data from this study demonstrate this effect regardless of intervention. However, if a response maintenance accuracy of 90% or above

is adequate, then acquisition criterion alone is not sufficient to produce those outcomes and the practitioners must evaluate other important variables that come into play.

A limitation of the analyses of response maintenance is the lack of differentiated evaluations based on the different time frames response maintenance were conducted. As a result, the maintenance data included assessments conducted one day after the conclusion of the intervention or 3 months after the conclusion of the intervention. Future analyses could evaluate differential response maintenance outcomes conducted at different points in time. Additionally, for cases that included multiple follow-up assessments, the response maintenance results were averaged across all sessions. Future analyses could look into only reporting the data point of the final maintenance session.

Acquisition Criteria and Generalization

A stimulus generalization assessment is another potential disruptor event where a researcher can measure degrees of persistence in the face of that disruptor. A primary difference between teaching conditions and stimulus generalization maintenance assessments is the topography of antecedent and discriminative stimuli (e.g., tacting different looking stimuli, a different individual delivering discriminative stimuli, or contacting discriminative stimuli in a different context). This is valuable to conduct because it measures the likelihood that the individual can perform newly acquired skills in environments that are not identical to the controlled teaching environment. It allows instructors to determine whether teaching procedures may need to be modified and training sessions may need to become more "loose" (Stokes & Baer, 1977). However, the importance of generalization seems to be overlooked as more than half of the published skill acquisition articles do not include an assessment of stimulus generalization to novel settings, novel stimuli, or novel individuals.

The overall results indicate that similar to response maintenance, the percentage of cases with higher generalization results are positively correlated with the use of higher acquisition criterion levels. However, about 30% of the cases with the highest level of criterion (100%) still did not produce generalization results at or above 90%. Over 40% of the cases with an acquisition criterion of 90%-99% did not produce generalization results at or above 90%. These data suggest that acquisition criterion alone is not sufficient in producing reliably high generalization results and there are aspects of the teaching procedures that may also be involved with how well an individual generalizes a novel skill they learned. To date, Semb (1974) is the only study that evaluates the impact of acquisition criteria on generalization. Thus, the findings of this descriptive analysis provide a novel contribution to the existing literature.

High acquisition-criterion levels that are achieved across multiple sessions suggest more restrictive stimulus control, yet the percentage of cases that produced higher generalization accuracy increased when the frequency of replications with high accuracy levels increased from one to two. A limitation of this analysis is that data were not collected on whether stimulus or response generalization procedures were programmed within the teaching interventions (procedures that trained individuals across novel individuals, situations, or stimuli). Thus, the analysis did not reflect the differences in intervention procedures.

Future Directions

The current analysis only reviewed three empirical ABA journals from only three years, which limited the generality of the findings. Additionally, the three years overlapped with one year of McDougale et al. (2019), which limited the comparisons between the two studies. Future analyses should collect data from a more diverse selection of journals and include additional years.

Ultimately, the associations among the response maintenance outcomes and generalization outcomes of these analyses were observed despite differences in participant age, interventions, and the types of behaviors that were targeted. The findings of the current analysis demonstrate that acquisition criteria play an important role in the mastery of novel behaviors across populations, interventions, and target behavior types. These criteria are a fundamental intervention component and moderators of intervention effectiveness. And future statistical analyses could be conducted to determine strength of the relations between acquisition-criteria and components of mastery.

Acquisition criteria determine the response strength of a behavior and impact the momentum of the behavior in the face of disruptions that are bound to occur in an individual's environment. An analysis that should be conducted in the future includes an evaluation of whether acquisition-criteria differ contingent on the type of novel skill being taught. The findings may lead to even more questions and may prompt further investigations into the process of establishing the most effective acquisition criterion for different types of skills. Furthermore, the results of this study are correlational and there is a need for more systematic empirical work to be conducted. Some systemic evaluations may include holding the frequency component of acquisition criteria constant and examining the effects of criterion levels on response maintenance, stimulus generalization, response generalization, and other dependent variables. Alternatively, researchers may hold the criterion level constant while examining the effects of the other relevant variables. Finally, it would be worthy to conduct systematic evaluations across participants of different age ranges/developmental levels and different types of behaviors (e.g., discrete responses, dynamic responding, scripted curricula, shaping, etc). These future studies

will provide a solid foundation for an effective technology of skill acquisition criteria and ultimately lead to more productive behavioral interventions that produce lasting change.

Table 1

Types of Skills Taught in Skill Acquisition Articles

Skill Category	Percentage of Articles
Academics	19.8
Engagement (Academic Related Behaviors)	2.9
Functional Life Skills	15.5
Safety Skills	7.2
Social Communication	17.9
Sports/Physical Activity	5.8
Treatment Implementation	18.4
Elementary Verbal Operants	12.6

Note. All the skill-acquisition articles analyzed were published in the *Journal of Applied Behavior Analysis*, *Behavioral Interventions*, and *Behavior Analysis in Practice* between 2017 and 2019.

Table 2

Results of a Chi-Square Test for Association Between Acquisition-Criterion Levels and Response Maintenance Outcomes

			Response Maintenance Outcomes		Total
			80% or less	80% or above	
Criterion Levels	80% or less	Count	14	3	17
		Expected	3.6	13.4	17.0
		Adjusted Residual	6.3*	-6.3*	
	80%-89%	Count	42	104	146
		Expected	30.8	115.2	146.0
		Adjusted Residual	2.8*	-2.8*	
	90%-99%	Count	3	55	58
		Expected	12.2	45.8	58.0
		Adjusted Residual	-3.2*	3.2*	
	100%	Count	36	194	230
		Expected	48.4	181.6	230.0
		Adjusted Residual	-2.9*	2.9*	
Total		Count	95	356	451
		Expected	95.0	356.0	451.0

Note. The asterisks denote adjusted residual values that are statistically significant.

Table 3

Results of a Chi-Square Test for Association Between Acquisition-Criterion Levels and Generalization Outcomes

			Generalization Outcomes		Total
			80% or less	80% or above	
Criterion Levels	80% or less	Count	7	3	10
		Expected	2.9	7.1	10.0
			2.9*	-2.9*	
	80%-89%	Count	57	76	133
		Expected	38.7	94.3	133.0
			4.7*	-4.7*	
	90%-99%	Count	9	35	44
		Expected	12.8	31.2	44.0
			-1.4	1.4	
	100%	Count	16	103	119
		Expected	34.6	84.4	119.0
			-4.8*	4.8*	
Total		Count	89	217	306
		Expected	89.0	217.0	306.0

Note. The asterisks denote adjusted residual values that are statistically significant.

Figure 1

Schematic Diagram of the Article Selection Process and the Number of Articles and Cases Analyzed

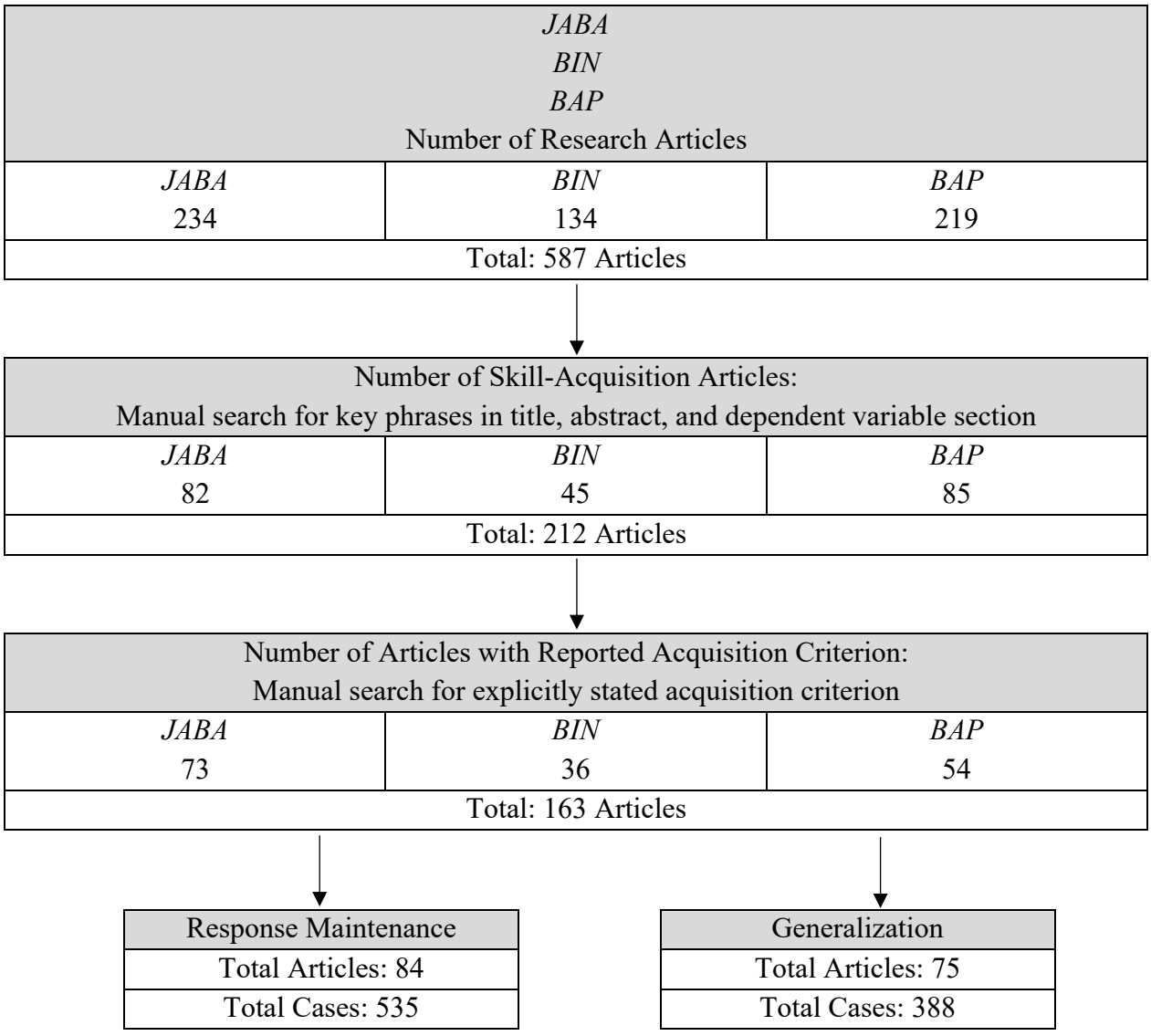
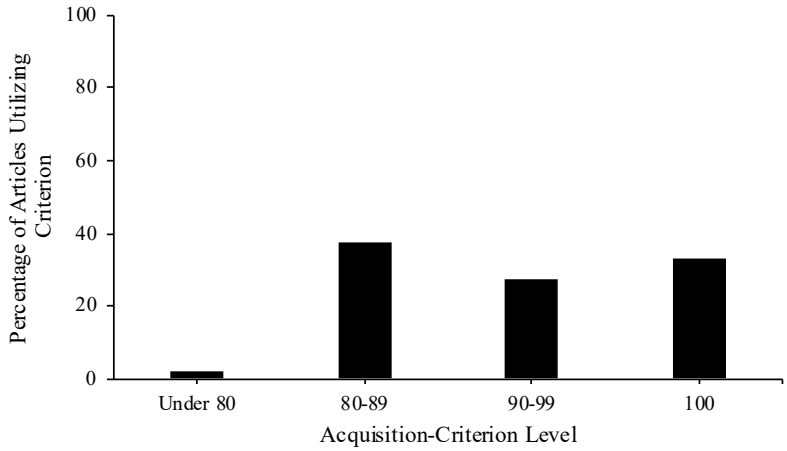


Figure 2

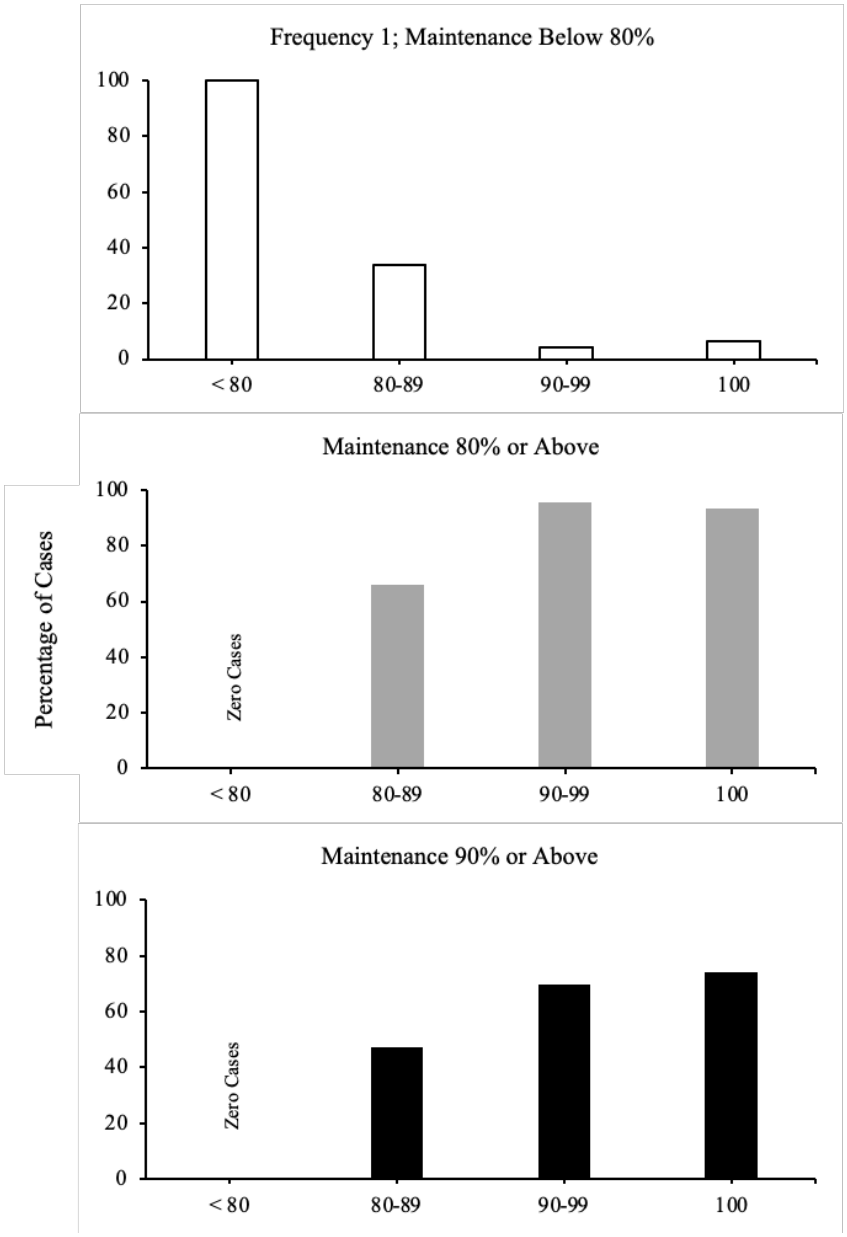
The Percentage of Articles Utilizing Specific Acquisition-Criterion Levels



Note. The acquisition-criterion levels are based on percentages. The percentage of skill acquisition articles utilizing criterion is shown on the y-axis, and the categories of acquistion-criterion levels are shown on the x-axis.

Figure 3

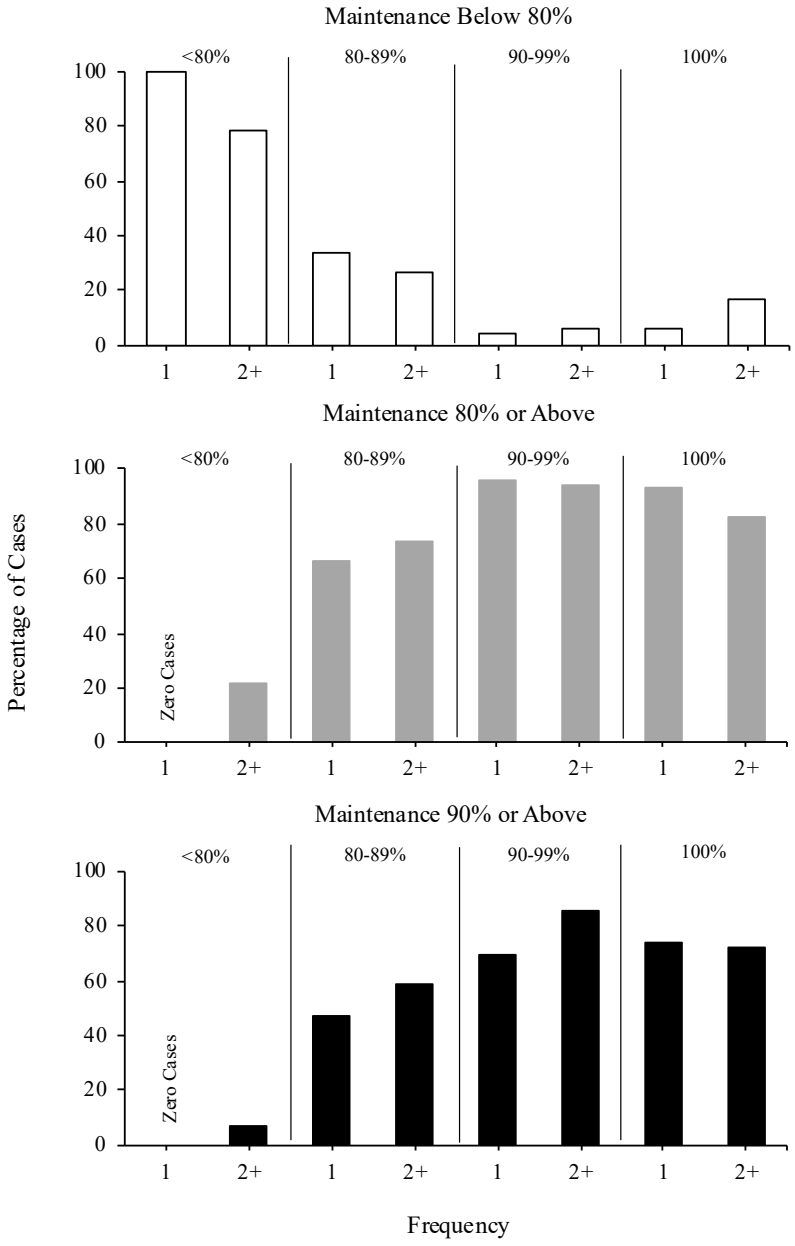
The Effects of Acquisition-Criterion Level on the Percentage of Cases Producing Various Response Maintenance Results



Note. The acquisition-criterion levels are based on percentages. There were zero cases with an under 80% acquisition-criterion level that produced response maintenance results at or above 80%

Figure 4

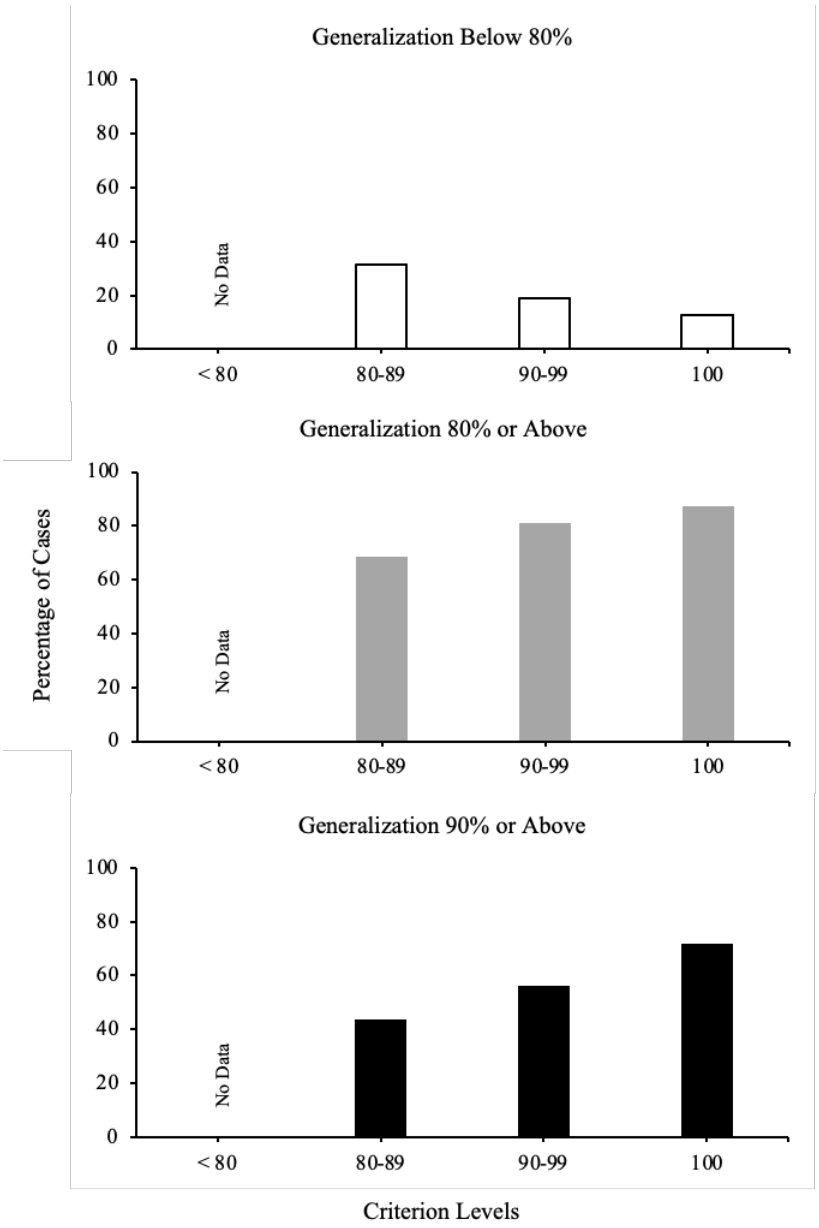
The Effects of Acquisition-Criterion Level and Frequency of Replications on the Percentage of Cases Producing Various Response Maintenance Results



Note. The bars on the left of each frame represent frequency of one replication and the bars on the right of each frame represent frequency of two or more replications.

Figure 5

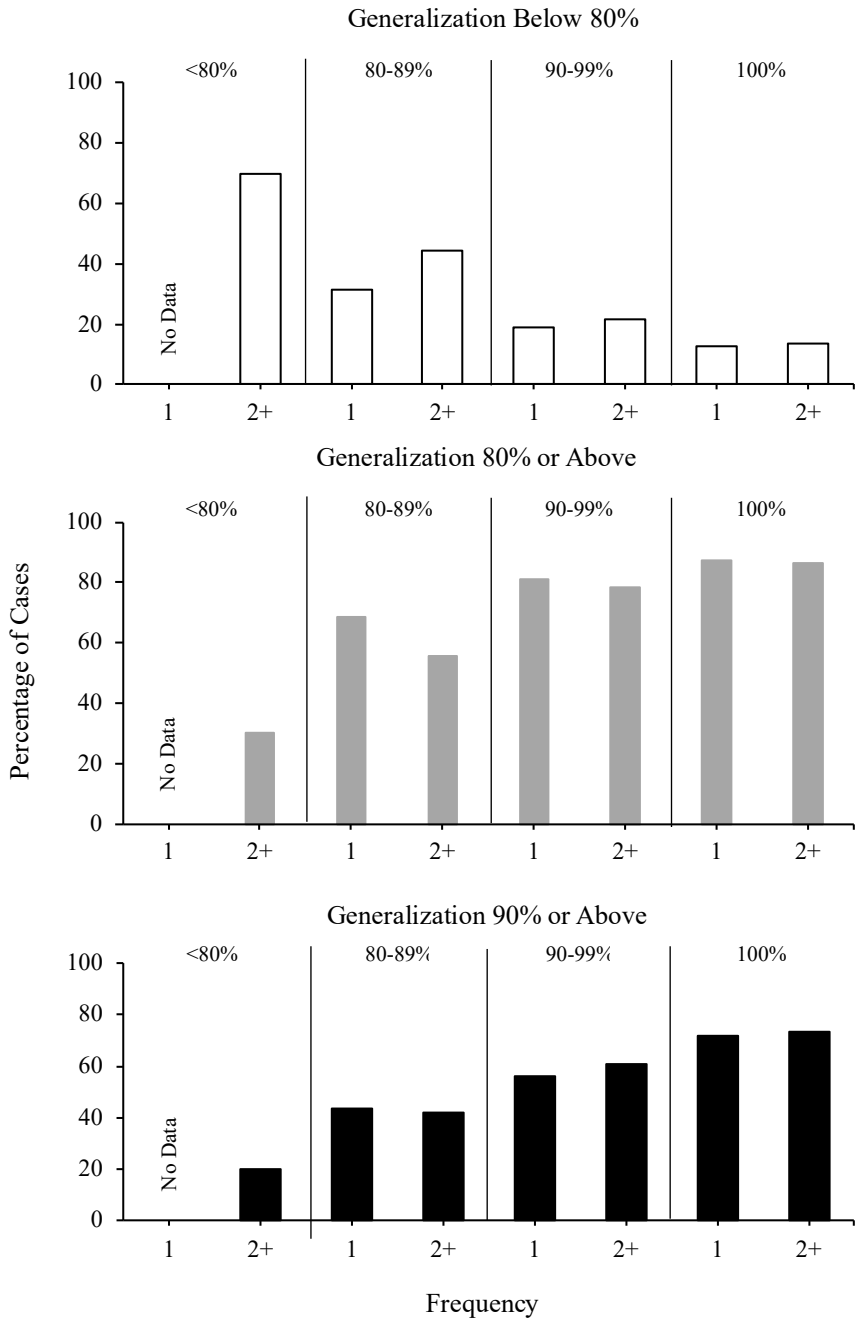
The Effects of Acquisition-Criterion Level on the Percentage of Cases Producing Generalization Results



Note. There were no data to report for an acquisition-criterion level under 80% that assessed for generalization.

Figure 6

The Effects of Acquisition-Criterion Level Across One Replication and Two or More Replications on the Percentage of Cases Producing Various Response Maintenance Results



Note. There were no data to report for an acquisition-criterion level under 80% that assessed for generalization.

Chapter 5: General Discussion

Applied Behavior Analysis (ABA) is based on the principles of behavior and emphasizes the study of behaviors 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 how data were interpreted in the social sciences, by introducing the concept of the "average man" (Donnelly, 2015). Among behavior scientists, a criticism of psychology's treatment of data is that using normal distribution curves to identify an "average" that supposedly represents large groups actually neglects important individual differences. Instead, behavior scientists reject this concept and prioritize the study of individual behavior because the average response in a particular sample of individuals does not accurately represent the behavior of any single individual (Chiesa, 1994). As a result, behavior scientists aim to examine the controlling variables of an individual's behavior.

Generally, teaching procedures that target the acquisition of novel behavior involve the use of learn unit instruction or discrete trial instruction. In order to determine the acquisition of a skill, instructors establish a performance criterion for the learner to achieve. When a learner performs at the predetermined criterion, the instructor concludes that the skill is acquired or that the current phase of intervention may discontinue. The predetermined criterion is typically established for a set of multiple operants or skills during learn unit instruction or discrete trial instruction. The problem with this process is the criterion for acquisition is based on an aggregated level of responding across multiple operants, rather than on individual operants. When a criterion is aggregated and tied to multiple operants, this may hinder the skill acquisition process because one operant may be preventing the learner from achieving the aggregated

criterion. Furthermore, one or more operants may never be truly acquired based on this process of applying acquisition criterion. Richling et al. (2019) reported that the most widely utilized acquisition criterion among ABA practitioners is an 80% accuracy across three sessions. During a typical learn unit instruction session, which may consist of 20 learn units per session to teach four operants, the criterion requires a learner to respond accurately to 16 out of 20 responses to achieve acquisition. This criterion allows a learner to respond incorrectly four times during a session. It is possible that all four incorrect responses may congregate within one operant, thus allowing the entire set of operants to achieve acquisition with the learner responding to one operant correctly with only 20% accuracy. This problem highlights how the data analysis process in our field may be at odds with the foundational principle of individuality.

Experiment I and Experiment II bear on an apparent discrepancy within our field - the rejection of aggregating performances across individuals but apparent acceptance of aggregating performances within an individual. The studies specifically evaluate the traditional set-based criterion (SA, aggregating performances) compared to a novel operant-based criterion (OA, focusing in on individual behaviors) on sight-word acquisition. The results of Experiment I suggest that there are problems inherent to utilizing a set-based and aggregated-behavior approach to acquisition criteria. Specifically, SA slows learning by adding potentially unnecessary training during the skill acquisition process. When researchers engage in OA, all four participants acquired many more sight words compared to the SA condition within the same time frame. On average, each operant acquired under the SA condition required an average of over 10 learn units compared to operants acquired under the OA condition. These preliminary skill acquisition results suggest the advantages of an individual operant-based analysis of acquisition criteria. An additional dependent variable was also measured in Experiment I. I

assessed the accuracy of responses during 4-week response maintenance sessions. This evaluation of the persistence of the learned responses during extinction probes was crucial for the identification of the overall mastery of a skill (Richling et al., in press). Three participants responded accurately to a higher percentage of operants acquired under the SA condition compared to the OA condition and one participant responded accurately to 100% of the operants acquired under both conditions. Overall, the reliability of response maintenance results favored the SA condition for three out of the four participants.

One explanation for the differential response maintenance results is that some amount of overtraining may be necessary for the behavioral persistence of some populations. Overtraining, also known as overlearning, refers to the deliberate and repeated practice of a skill even after an objective is already achieved. Many studies have provided evidence for enhanced performance and retention after overtraining (Driskell & Willis, 1992; Hagman & Rose, 1983; Krueger, 1930). A meta-analysis of overtraining showed that the effectiveness of overlearning was moderated by the types of skills that were targeted, the response maintenance period, and the magnitude of overlearning (Driskell & Willis, 1992). Nevertheless, a potential caveat of overtraining may include additional costs and time that are required. It is imperative to consider these added costs before implementing overtraining and it may be necessary to evaluate the right amount of overtraining that produces persistent behavior in order to offset the costs. Furthermore, it may be necessary to consider the type of target skill being taught. For example, deliberate overtraining for academic skills that are taught as a part of a spiraled curriculum may not be necessary.

In the context of the sight word instruction implemented in Experiment I, it was important to consider the benefits of additional overtraining trials. I increased the number of

criterion-level replications (100% x 1 v. 100% x 2) a learner must achieve at the target acquisition-criterion level to examine whether this increased stringency would lead to more accurate response maintenance results. The acquisition criterion for each condition was 100% accuracy across two sessions. Not only did all four participants of Experiment II acquire a greater number of sight words under OA (similar skill acquisition effects as Experiment I), but also all four participants also responded accurately to a higher or comparable percentage of sight words during 4-week response maintenance assessment sessions. These results suggest that increasing the criterion-level replication value from one session to two sessions led to higher behavioral persistence of the acquired operants.

The goal of Experiment I and Experiment II was to contribute to the literature on acquisition criteria. Acquisition criteria, or more commonly known as "mastery criteria" has been an integral part of ABA practices since the early 1970s, and the number of ABA practitioners who utilize an acquisition criterion has steadily increased since then (Rehfeldt & Ghezzi, 1996; Sayrs & Ghezzi, 1997). However, through the decades, only a handful of studies have evaluated acquisition criteria as an independent variable among college students (Carlson & Minke, 1975; Johnson & O'Neill, 1973; Semb, 1974) and more recently, among children with developmental disabilities (Fuller & Fienup, 2018; Longino et al., in press; Pitts & Horger, 2021; Richling et al., 2019). The limitation of these studies, including the aforementioned experimental analyses of OA and SA, is the small sample size. The total number of participants studied in Fuller and Fienup (2018), Longino et al. (in press), Pitts and Horger (2021), Richling et al. (2019), and Experiment I and Experiment II is less than 25. These studies demonstrate promising results regarding different components of acquisition criteria; however, the small sample size calls into question the ability for the results to generalize to different groups of individuals, target

skills, and teaching interventions. The need to examine the external validity of these results led to the descriptive analysis of acquisition criteria in Experiment III.

The Experiment III descriptive analysis evaluated the effects of percentage-based acquisition criterion levels in addition to different criterion-level replication values on response maintenance results and generalization results. In this analysis, 212 published articles and over 500 cases (individual participants completing a condition) were reviewed. The results of the sub-analyses of acquisition criteria and the effects on response maintenance support the existing literature (Fuller & Fienup, 2018; Pitts & Hoerger, 2021; Richling et al., 2019; Semb, 1974). The results of the sub-analyses on the effects of generalization extends the existing literature because there is only one study in this area (Semb, 1974). Overall, interventions that utilize higher criterion levels lead to better response maintenance and generalization while interventions that utilize lower criterion levels lead to inferior responding. Interestingly, these associations are evident despite a wide range of participant age, participant abilities, types of behaviors targeted, and types of interventions used.

The results of the descriptive analysis have important practical implications for ABA practitioners. According to the descriptive analysis, the most widely utilized acquisition-criterion level among both ABA practitioners and ABA researchers is between 80% and 89% accuracy. However, just over 60% of interventions that utilize this criterion range produce response maintenance results over 80% accuracy. Approximately 50% of interventions that utilize this criterion range produce maintenance results over 90% accuracy. Similar effects are shown for generalization results. These data provide evidence that an 80%-89% acquisition-criterion level may not be beneficial for the learner. Instructors must carefully consider what level of accuracy during response maintenance and generalization assessments is adequate for the learners and

adjust the acquisition criterion in their teaching interventions accordingly. For instance, an instructor may determine that an 80% accuracy during response maintenance is adequate for the particular behavior the instructor is teaching. If this is the case, almost 100% of all cases in published journals support a criterion-level between 90%-99% combined with a frequency of either one replication or two replications for producing response maintenance accuracy at or above 80% accuracy. It would not be cost effective for the instructor to utilize an acquisition-criterion level of 80% due to its ineffectiveness in producing high maintenance. Alternatively, the instructor may utilize an acquisition criterion of 90% across one replication rather than 90% across two replications. Both criteria (90% x 1 and 90% x 2) produce similar results, and the instructor would save additional costs associated with the 90% across two replication criterion. With the results reported in the descriptive analysis, instructors have information to determine the acquisition criteria that will produce beneficial results while reducing unnecessary overtraining. Furthermore, in both analyses of response maintenance and generalization, there is a decrease in the effects of criterion levels when moving from 80% or higher results to 90% or higher results. The generally lower level of cases producing 90% or higher results suggest that criterion-level alone may not be sufficient for producing high levels of response maintenance and generalization accuracy. Thus, instructors must consider how other variables in the teaching process may contribute to maintenance and generalization results.

Overall, these results are important for two main reasons. First, the data provide ABA practitioners a scientific basis on which they can formulate important decisions regarding acquisition criteria. Second, the data provide support for the external validity of the existing research on the topic of acquisition criteria. The handful of recent studies examining the effects of different acquisition criteria on response maintenance accuracy produced promising findings,

although the overall sample size that were studied were perhaps too small to draw broad conclusions regarding acquisition criteria effects. The descriptive analysis in Experiment III supported the use of more stringent criterion levels and that these acquisition criteria led to higher maintenance results and higher degrees of generalization results regardless of participant age, gender, grade-level, or diagnosis. However, the lack of sufficient participant information across all skill acquisition studies prevented the current descriptive analysis from examining how individual characteristics may affect the outcomes of different acquisition criteria. Research in ABA should consider including more detailed participant information including demographic information, test scores, and levels of verbal behavior.

It may be particularly important to examine the differential effects of acquisition criteria with individuals of various degrees of verbal behavior functioning especially when considering the educational implications of the descriptive analysis findings. According to the Verbal Behavior Developmental Theory (Greer & Ross, 2008; Greer & Speckman, 2009), children progress through a developmental trajectory and attain key verbal cusps. Verbal behavior cusps allow children to contact new environmental contingencies and learn skills in new ways, often at an accelerated rate (Greer & Ross, 2008; Rosales-Ruiz & Baer, 1997). As children acquire more verbal behavior cusps, such as Bidirectional Naming (BiN), they learn in more efficient ways and do not need extensive direct instruction (Greer et al., 2011; Greer & Du, 2015; Hranchuk et al., 2019). Likewise, children with BiN may not require stringent levels of acquisition criterion in to acquire novel skills. Future research should consider how acquisition criterion may change as a function of an individual's level of verbal behavior functioning of students.

An accurate characterization of a mastered skill goes beyond the initial acquisition of the skill. It also requires the long-term persistence of the skill (maintenance), and the generalization

of the skill to novel individuals, stimuli, and settings (stimulus generalization). Simply identifying a skill as mastered when a learner emits accurate responses during instructional sessions does not cover the range of expected educational outcomes that should be entailed in such a definitive term as "mastery." Both response maintenance and generalization should be considered in the identification of a mastered skill (Richling et al., in press). However, there is not enough research in our field that adequately evaluates the complete mastery of a novel skill. Perhaps one of the most important findings of the descriptive analysis is the lack of studies reporting the assessment of response maintenance and stimulus generalization. Only half of the published studies targeting skill acquisition report assessments of response maintenance. Even fewer studies report the assessment of any form of generalization. Yet, these same skill acquisition studies mention the term "mastery" 1,841 times in total. The field will benefit from more empirical studies on acquisition criteria and on the components associated with mastery of novel skills. Moreover, with only 36% of the articles published in *JABA, BIN,* and *BAP* between 2017-2019 targeting the acquisition of novel skills, the field of ABA could benefit from an increased focus on improving socially significant behaviors in the realm of skill acquisition more broadly.

References

Adams, L. (2011). *Learning a new skill is easier said than done.* Gordon Training International. https://www.gordontraining.com/free-workplace-articles/learning-a-new-skill-is-easiersaid-than-done/

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

Ardoin, S. P., Witt, J. C., Connell, J. E., & Koenig, J. L. (2005). Application of a three-tiered response to intervention model for instructional planning, decision making, and the identification of children in need of services. *Journal of Psychoeducational Assessment, 23*(4), 362-380. https://doi.org/10.1177/073428290502300405

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. https:// doi.org/ 10.1901/jaba.1968.1-91

Baer, D. M., Wolf, M. M., & Risley, T. R. (1987). Some still current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, *20*(4), 313–327. https://doi.org/10.1901/jaba.1987.20-313

Bransford, J. D., & Schwartz, D. (2009). It takes expertise to make expertise: Some thoughts about how and why. In K. A. Ericsson (Ed.), *Development of Professional Expertise: Toward Measurement of Expert Performance and Design of Optimal Learning Environments* (pp. 432–448). New York: Cambridge University Press. https://aaalab.stanford.edu/assets/papers/2006/Takes_Expertise_to_Make_Expertise.pdf

Brigham, T. A., & Sherman, J. A. (1968). An experimental analysis of verbal imitation in preschool children. *Journal of Applied Behavior Analysis, 1*(2), 151-158. https://doi.org/10.1901/jaba.1968.1-151

Brodsky, J., & Fienup, D. M. (2018). Sidman goes to college: A meta-analysis of equivalence based instruction in higher education. Perspectives on Behavior Science, 41 (1), 95-119. https://doi.org/10.1007/s40614-018-0150-0

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*, https://doi.org/10.1177/0741932521996071.

Carlson, J., & Minke, K. (1975). Fixed and ascending criteria for unit mastery learning. *Journal of Educational Psychology, 67*(1)*,* 96–101. http://dx.doi.org/10.1037/h0078676.

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

Cooper, J. O., Heron, T. E., & Heward, W. L. (2020). *Applied behavior analysis*, 3rd edition. Hoboken, NJ: Pearson.

Cregger, R., & Metzler, M. (1992). PSI for a college physical education basic instructional program. *Educational Technology, 32,* 51-56. https://eric.ed.gov/?id=EJ450464

Craig, A. R., Nevin, J. A., & Odum, A. L. (2014). Resistance to Change. *The Wiley Blackwell Handbook of Operant and Classical Conditioning*, 249. https://doi.org/10.1002/9781118468135.ch11

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

Doughty, S. S., Chase, P. N., & O'Shields, E. M. (2004). Effects of rate building on fluent performance: a review and commentary. *The Behavior analyst*, *27*(1), 7–23. https://doi.org/10.1007/BF03392086

Dreyfus, S. E., & Dreyfus, H. L. (1980). *A five-stage model of the mental activities involved in directed skill acquisition*. California Univ Berkeley Operations Research Center. https://apps.dtic.mil/dtic/tr/fulltext/u2/a084551.pdf

Dreyfus, S. E. (2004). The five-stage model of adult skill acquisition. *Bulletin of Science, Technology & Society*, *24*(3), 177-181. https://doi.org/10.1177/0270467604264992

Driskell, J. E., Willis, R. P., & Copper, C. (1992). Effect of overlearning on retention. *Journal of Applied Psychology, 77*(5), 615–622. https://doi.org/10.1037/0021-9010.77.5.615

Ericsson, K. A. (2006). *The Influence of Experience and Deliberate Practice on the Development of Superior Expert Performance.* In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (683–703). Cambridge University Press. https://doi.org/10.1017/CBO9780511816796.038

Ericsson, K. A., Nandagopal, K., & Roring, R. W. (2009). Toward a science of exceptional achievement. *Annals of the New York Academy of Sciences*, *1172*(1), 199. https://doi.org/10.1196/annals.1393.001

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

Fuchs, L. S., & Deno, S. L. (1991). Effects of curriculum within curriculum-based measurement. *Exceptional Children, 58*(3), 232-243. https://doi.org/10.1177/001440299105800306

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.

Good III, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5*(3), 257-288.

Gladwell, M. (2008). *Outliers: The story of success*. Little, Brown.

Greer, R. D. (1996). The education crisis. In M. A. Mattaini & B. A. Thyer (Eds.), *Finding Solutions to Social Problems: Behavioral Strategies for Change* (p. 113–146). American Psychological Association. https://doi.org/10.1037/10217-005

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

Greer, R.D., Corwin, A., & Buttigieg, S. (2011). The effects of the verbal developmental capability of naming on how children can be taught. *Acta de Investigación Psicológica*, *1*(1), 23-54.

Greer, R.D., Du, L. (2015) Experience and the Onset of the Capability to Learn Names Incidentally by Exclusion. *The Psychological Record,* 65**,** 355–373. https://doi.org/10.1007/s40732-014-0111-2

Greer, R. D., & Longano, J. (2010). A rose by naming: How we may learn how to do it. *The Analysis of Verbal Behavior*, *26*(1), 73-106. https://doi.org/10.1007/BF03393085

Greer, R. D., & Ross, D. E. (2004). Verbal behavior analysis: A program of research in the induction and expansion of complex verbal behavior. *Journal of Early and Intensive Behavior Intervention*, *1*(2), 141. https://doi.org/ 10.1037/H0100286

Greer, R. D., & Ross, D. E. (2008). *Verbal behavior analysis: introducing and expanding new verbal capabilities in children with language delays.* Boston, MA: Pearson Education, Inc.

Greer, R. D., & Speckman, J. (2009). The integration of speaker and listener responses: A theory of verbal development. *The Psychological Record, 59*(3), 449–488.

Gutierrez Jr, A., Hale, M. N., O'Brien, H. A., Fischer, A. J., Durocher, J. S., & Alessandri, M. (2009). Evaluating the effectiveness of two commonly used discrete trial procedures for teaching receptive discrimination to young children with autism spectrum disorders. Research in Autism Spectrum Disorders, 3(3), 630-638. https://doi.org/10.1016/j.rasd.2008.12.005

Hagman, J. D., & Rose, A. M. (1983). Retention of military tasks: A review. *Human Factors, 25*(2), 199–213.

Hall, R. V., Lund, D., & Jackson, D. (1968). Effects of teacher attention on study behavior. *Journal of Applied Behavior Analysis, 1*(1), 1–12. https://doi.org/10.1901/jaba.1968.1-1

Hannon, J. C., Holt, B. J., & Hatten, J. D. (2008). Personalized system of instruction model: Teaching health related fitness content in high school physical education. *Journal of Curriculum and Instruction, 2*(2), 20- 33

Hoffman, R. R. (1998). How can expertise be defined? Implications of research from cognitive psychology. In *Exploring Expertise*, (pp. 81-100). Palgrave Macmillan, London. https://doi.org/10.1007/978-1-349-13693-3_4

Honken, N. (2013). *Dreyfus Five-Stage Model of Adult Skills Acquisition Applied to Engineering Lifelong Learning* [Conference session]. 120th ASEE Annual Conference & Exposition, Atlanta, GA, United States. https://doi.org/10.18260/1-2--19457.

Hranchuk, K., Greer, R. D., & Longano, J. (2019). Instructional demonstrations are more efficient than consequences alone for children with naming. *The Analysis of Verbal Behavior*, *35*(1), 1-20. https://doi.org/10.1007/s40616-018-0095-0

Jimerson, S. R., Burns, M. K., & VanDerHeyden, A. M. (2016). *Handbook of response to intervention: The science and practice of multi-tiered systems of support* (2nd Edition). Springer.

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

Kelly, L., & Holloway, J. (2015). An investigation of the effectiveness of Behavioral Momentum on the acquisition and fluency outcomes of tacts in three children with Autism Spectrum Disorder. *Research in Autism Spectrum Disorders*, *9*, 182-192. https://doi.org/10.1016/j.rasd.2014.10.007

Krueger, W. F. C. (1930). Further studies in overlearning. *Journal of Experimental Psychology, 13*(2), 152–163. https://doi.org/10.1037/h0075484

Lerman, D. C., & Iwata, B. A. (1996). Developing a technology for the use of operant extinction in clinical settings: An examination of basic and applied research. *Journal of Applied Behavior Analysis, 29*(3), 345-382. https://doi.org/10.1901/jaba.1996.29-345

Lindsley, O. R. (1971) Precision teaching in perspective: An interview with Ogden R. Lindsley. *Teaching Exceptional Children, 3*(3), 114-119. https://doi.org/10.1177/004005997100300303

Lindsley, O. R. (1991). Precision teaching's unique legacy from BF Skinner. *Journal of Behavioral Education*, *1*(2), 253-266. https://doi.org/10.1007/BF00957007

Longino, E. B., McDougale, C. M., Richling, S. M., & Palmier, J. (in press). Evaluation of a mastery criteria and maintenance using a most-to-least prompting hierarchy. *Behavior Analysis in Practice.*

McDougale, C., Richling, S. M., Longino, E. B., & O'Rourke, S. A. (2019). 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

Meindl, J. N., Ivy, J. W., Miller, N., Neef, N. A., & Williamson, R. L. (2013). An examination of stimulus control in fluency-based strategies: SAFMEDS and generalization. *Journal of Behavioral Education*, *22*(3), 229-252. https://doi.org/10.1007/s10864-013-9172-6

Merriam-Webster, (n.d.). *Merriam-Webster.com dictionary.* Retrieved December 20, 2020, from https://www.merriam-webster.com/

Miguel, C. F. (2016). Common and intraverbal bidirectional naming. *The Analysis of Verbal Behavior*, *32*(2), 125-138. https://doi.org/10.1007/s40616-016-0066-2

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

Nevin, J. A., Mandell, C., & Atak, J. R. (1983). The analysis of behavioral momentum. *Journal of the Experimental Analysis of Behavior*, *39*(1), 49-59. https://doi.org/10.1901/jeab.1983.39-49

Nevin, J. A. (1992). An integrative model for the study of behavioral momentum. *Journal of the Experimental Analysis of Behavior, 57*(3), 301-316. https://doi.org/10.1901/jeab.1992.57301

Nevin, J. A. (1996). The momentum of compliance. *Journal of Applied Behavior Analysis*, *29*(4), 535-547. https://doi.org/10.1901/jaba.1996.29-535

Nevin, J. A., & Shahan, T. A. (2011). Behavioral momentum theory: equations and applications. *Journal of Applied Behavior Analysis*, *44*(4), 877–895. https://doi.org/10.1901/jaba.2011.44-877

Pitts, L., & Hoerger, M. L. (2021). Mastery criteria and the maintenance of skills in children with developmental disabilities. *Behavioral Interventions,* 1-10. https://doi.org/10.1002/bin.1778

Pritchard, T., Penix, K., Colquitt, G., & McCollum, S. (2012). Effects of a weight training personalized system of instruction course on fitness levels and knowledge. *Physical Educator, 69*(4), 342-359.

Rehfeldt, R. A., Ghezzi, P. M. (1996). The steady-state strategy in human operant research: How

stable are we? *Experimental Analysis of Human Behavior Bulletin, 14*(2), 23-25.

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

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.

Rosales-Ruiz J, Baer DM. Behavioral cusps: a developmental and pragmatic concept for behavior analysis. *Journal of Applied Behavior Analysis*. *30*(3):533-544. https://doi.org/10.1901/jaba.1997.30-533.

Schendel, J. D., & Hagman, J. D. (1982). On sustaining procedural skills over a prolonged retention interval. *Journal of Applied Psychology*, *67*(5), 605–610. https://doi.org/10.1037/0021-9010.67.5.605

Schneider, J., & Hutt, E. (2014). Making the grade: A history of the A–F marking scheme. *Journal of Curriculum Studies*, *46*(2), 201-224. https://doi.org/10.1080/00220272.2013.790480

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

Shapiro, E. S. (1996). *Academic skills problems.* New York: Guilford.

Simon, C., Bernardy, J. L., & Cowie, S. (2020). On the "Strength" of Behavior. *Perspectives on Behavior Science*, *43*(4), 677-696. https://doi.org/10.1007/s40614-020-00269-5

Stokes, T. F., & Baer, D. M. (1977). An implicit technology of generalization. *Journal of Applied Behavior Analysis, 10*(2), 349–367. https://doi.org/10.1901/jaba.1977.10-349

Thorndike, E. L. (1913). Educational psychology, Vol 1: The original nature of man. https://doi.org/10.1037/13763-000

Uhl, C. N., & Young, A. G. (1967). Resistance to extinction as a function of incentive, percentage of reinforcement, and number of nonreinforced trials. *Journal of Experimental Psychology, 73*(4, Pt.1), 556–564. https://doi.org/10.1037/h0024389

Wang, J. M., & Zorek, J. A. (2016). Deliberate practice as a theoretical framework for interprofessional experiential education. *Frontiers in Pharmacology*, *7*, 188. https://doi.org/10.3389/fphar.2016.00188

Zencius, A. H., Davis, P. K., & Cuvo, A. J. (1990). A personalized system of instruction for teaching checking account skills to adults with mild disabilities. *Journal of Applied Behavior Analysis*, *23*(2), 245–252. https://doi.org/10.1901/jaba.1990.23-245