5-2023

Differing Mastery Criteria Effects on Maintenance of 1-Step Instruction Following in Children with Disabilities

Kyle Wagner
Utah State University

DIFFERING MASTERY CRITERIA EFFECTS ON MAINTENANCE OF 1-STEP INSTRUCTION FOLLOWING IN CHILDREN WITH DISABILITIES

by

Kyle Wagner

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Special Education

Approved:

_____ Ray Joslyn, Ph.D., BCBA-D, LBA Major Professor	_____ Sophia D'Agostino, Ph.D., BCBA-D Committee Member
_____ Trent Landon, Ph.D., CRC, LVRC Committee Member	_____ D. Richard Cutler, Ph.D. Vice Provost for Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2023

ABSTRACT

Differing Mastery Criteria Effects on Maintenance of 1-Step Instruction Following in Children with Disabilities

by

Kyle Wagner, Master of Science

Utah State University, 2023

Major Professor: Dr. Ray Joslyn
Department: Special Education and Rehabilitation Counseling

During EIBI, practitioners assign taught skills a specific mastery criterion. The most commonly used mastery criterion is 80% accuracy across three sessions. There is little to no evidence to support the use of an 80% criterion. Preliminary evidence suggests that 80% may not be sufficient to promote meaningful maintenance of skills taught. This study conducted an experiment to evaluate the effects of differing mastery criteria on maintenance of 1-step instructions in two children with developmental disabilities. They were each taught three, 1-step instruction skills with differing mastery criteria assigned to each skill. Results were idiosyncratic and no meaningful conclusions can be made in regard to differing mastery criteria effects on maintenance based on this study.

(39 pages)

PUBLIC ABSTRACT

Differing Mastery Criteria Effects on Maintenance of 1-Step Instruction Following in Children with Disabilities

Kyle Wagner

The majority of Early Intervention clinics throughout the United States are currently using a practice that is not considered evidence based. When teaching skills to children with developmental disabilities, clinicians assign an accuracy standard for those skills. The most common accuracy standard, called a "mastery criterion," is 80% across three teaching sessions. There is little to no evidence to support using this specific mastery criterion. The purpose of this study was to add more research regarding this topic, and how it affects how the skills taught maintain over time when teaching has ceased for a specific skill. Two children were taught different skills to different mastery criteria. The results of this study were not what was expected, and no meaningful conclusions can be made at this time regarding specific mastery criteria.

ACKNOWLEDGMENTS

I give special thanks to thank my wife, Sam, who is always there for me. Her patience with me as I have worked on my degree has been invaluable, and I hope I can repay her someday. Similarly, I'd like to thank my friends, family, and colleagues who have helped me get to this point, for what they have taught me, and their continued support and encouragement.

I would also like to thank Kylie Atkinson, Mikayla Hansen, Ausha Trujillo, and Jeni Buist. The first three for what they taught me about the field of behavior analysis, and all of them for their examples of being a good educator, team member, and most importantly, a person.

Kyle Wagner

CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Differing Mastery Criteria Effects on Maintenance of 1-Step Instruction Following in Children with Disabilities

Introduction

Early Intensive Behavioral Intervention (EIBI) clinics are learning centers for children with Autism Spectrum Disorder (ASD), usually between the ages of 3-5. These clinics use evidence-based principles of applied behavior analysis to teach daily living, communication, and academic skills to children in need of additional supports. Examples include, but are not limited to, receptive identification tasks, imitation/instruction following, fine and gross motor skills, receptive object labeling, and social skills. These kinds of skills are important for anyone to learn to function appropriately in society, and children with ASD often have trouble in these areas and may not learn these skills without explicit instruction. ASD is a neurodevelopmental disorder characterized by difficulties with social-communicative function and repetitive and restrictive behaviors and sensory activities (American Psychiatric Association [APA], 2013). To offset these deficits, children attending EIBI clinics learn key skills through a variety of evidence-based teaching strategies, including discrete-trial teaching (DTT), naturalistic teaching (NT), activity schedules (Brodhead et al., 2008), social skills training, and more (Love et al., 2019). There is substantial evidence to show that EIBI increases proficiency in these skills (Eldevik et al., 2009).

Smith et al. (2021) found that those who received 2 years of intervention in EIBI clinics gained increased standard scores in cognitive and adaptive functioning, and reductions in autism symptoms and maladaptive behavior. These benefits were maintained at 10 years after EIBI had concluded. Similarly, Noyes-Grosser et al. (2018) found that those who received EIBI showed reduced maladaptive behaviors and improved social communication skills. Lastly, Perry et al. (2019) found a "general pattern of stability" (p. 181) since the conclusion of EIBI up to 14 years later.

As children learn in EIBI clinics, the degree to which they retain their new abilities after the explicit teaching process ends is crucial to their development. There would be little long-term value in teaching a skill if the learner loses their ability to perform the newly learned skill shortly after acquisition. Response maintenance, which is defined as "the extent to which the learner continues to perform the target behavior after a portion or all of the intervention has been terminated" (Cooper et al., 2007, p.698), is a measure of the degree to which children retain the skills they have learned and is a central focus of EIBI. In a survey conducted by Love et. al (2009), 200/211 respondents (98%) stated that their instructional program includes procedures designed to increase maintenance of skills taught. There are several factors that influence response maintenance, including supports in the student's natural environment, the quality of the instruction they have received, and variables related to the programming of the instruction (Cooper et al., 2007). Despite there being an abundance of research in EIBI settings regarding the effects of behavioral interventions on skill acquisition and maintenance, one aspect that has very little research is how differing mastery criteria

affect maintenance of skills taught (Fienup & Carr, 2021). Mastery criteria refer to "the operationalized standard that behavior must reach before an instructor ceases or changes instruction" (Fienup & Carr, 2021). In other words, what fluency level (defined as the ability to accurately perform a task demand) must first be shown regarding skill performance before the skill is considered "mastered." Mastery criteria are often shown in the form of a percentage of correct trials, usually across one or more sessions or interventionists (Luiselli et al., 2009).

Preliminary research in this area has shown that differing mastery criteria may predict maintenance of behavior taught in the EIBI context (Fuller & Fienup 2018; Richling et al. 2019; Pitts & Hoerger 2021; Schneider, 2021). Fuller and Fienup (2018), Richling et al. (2019), and Pitts and Hoerger (2021) have all shown that the mastery criterion used by many clinics nationwide may not be sufficient to promote maintenance of skills. Richling et al. (2019) showed that using a 60% criterion showed just as good, if not better, maintenance of skills with most participants when compared to an 80% criterion. Although these criteria can certainly be obtained faster, they may not maintain as well as a more stringent criterion like 90%, or 100%. However, more stringent criteria usually take longer to teach (Richling et al., 2019). The obvious tradeoff is faster "mastery" for lower maintenance, compared to longer teaching periods for better mastery and maintenance.

The established mastery criterion for tasks/skills at many clinics nationwide is 80% across one or more consecutive sessions when teaching students (Richling et al., 2019). That is, a skill is not considered mastered until a student can show correct

responding 80% of the time for one or more consecutive teaching sessions. Approximately 54% of clinics across the United States use 80% across one or more sessions as their set mastery criteria (Richling et al., 2019). If more than half of the clinics nationwide use this mastery criterion then this leads one to wonder what evidence is there to support using an 80% across one or more sessions mastery criterion as the standard, or is there is any evidence at all? Researchers and practitioners might also inquire as to how the response maintenance is affected by differing mastery criteria. The majority of the literature on mastery criteria has focused on those who are "neuro-typical" college aged persons, (Carlson & Minke, 1975; Fienup & Broadsky, 2017; Johnson & O'Neil, 1973; Keller, 1968; Pitts & Hoerger p. 523, 2021; Semb, 1974), and has neglected the teaching of instruction following skills to younger populations, leaving a large gap in the literature regarding those who are "neurodiverse", are of a younger age group (i.e. 3-5 years old) and how differing mastery criteria and maintenance relate to instruction following skill acquisition.

Literature Review

To date, there are very few studies to have examined why an 80% mastery criterion is so common and whether it is sufficient to produce desirable response maintenance results. Using PsychInfo via EBSCOhost, the search terms "mastery criteria and maintenance" yielded 52 results. After limiting results to empirical studies published in journals there were 37 options. Upon further examination of each available article,

only 3 had relevance to the effects of mastery criteria on maintenance: Fuller & Fienup (2018), Richling et al. (2019), and Pitts & Hoerger (2021).

Richling et al's. (2019) study is the most comprehensive study on differing mastery criteria effects to date. Richling et al. (2019) conducted 4 experiments within their study. For the first experiment they conducted a survey of 199 Board Certified Behavior Analysts/-Doctorates (BCBA/BCBA-D) working in clinics across the United States. Of the 199 BCBA/BCBA-Ds, 44% of these specified that they used their mastery criterion as a result of previous supervision. Twenty percent indicated employer policies/requirements determined their mastery criterion. Sixteen percent cited graduate school, while 10% said continuing education (workshops etc.) were their reasoning. Nine percent said regulated requirements such as IEPs determined their mastery criterion. Finally, 2% cited funding sources as their reasoning. Of these, only 26% have cited reasoning that may have derived from empirical evidence (graduate school and continuing education/workshops). If asked, I would fall under the group of "employer requirements" (Richling et al., 2019) for where my "knowledge" of what mastery criteria are acceptable comes from. Fifty-four percent of respondents indicated using an 80% mastery criterion across one or more sessions. Twenty-eight percent said they use a 90% criterion, and 7% use a 100% criterion. Six percent use between an 81% and 89% criterion while 5% use between a 91% and 99% criterion. Lastly, 1% reported using below 80% as their mastery criterion.

For their second experiment they tested 60%, 80%, and 100% mastery criterion across three sessions. Four children, ages 6-9, with varying disabilities participated in

each experiment. Prior to each 10-trial session a Multiple Stimulus Without Replacement (MSWO) preference assessment (DeLeon & Iwata, 1996) was conducted. Targets included printed pictures of animals, plants, and food items used for receptive identification tasks, an auditory-visual discrimination. Each child was assigned three stimulus sets with each stimulus set being counterbalanced by using the same three stimulus sets across all participants, each participant having a different mastery criterion (60%, 80%, or 100%) assigned to each stimulus set. The experimenter followed common DTT procedures, placing three stimuli in front of the learner and giving a direction (e.g., "touch dog") and recording the responses to get a percentage of the number of correct responses for a 10-trial session. Weekly follow-up probes (one 10-trial session) were conducted each week for four weeks once mastery had been obtained. Results from this experiment indicated that all four participants scored at or above 80% mastery for skills taught to the 100% criterion. For the 80% criterion, two participants maintained near or slightly below 80%, and the other two participants had immediate drops in response accuracy, going as low as 40%. Lastly, skills taught to the 60% criterion dropped far below the expected 80% accuracy for all but one participant who acquired higher response accuracy compared to the 80% criterion. Overall, the 100% criterion is the only one that maintained above 80% for each participant.

Experiment 3 was designed "to systematically replicate Experiment 2" (Richling et al. 2019, p. 709). The main difference was the use of a non-choice-based task, a potential confound of experiment one. Vocal tacting, "a verbal operant in which a speaker names things and actions that the speaker has direct contact with through any of

the sense modes" (Cooper et al. 2007, p. 530), made up the tasks for this experiment. Three stimulus sets of three items were used again, counterbalanced in the same way as experiment two with the same mastery criteria being used. The targets were again pictures of animals, food, plants, and abstract or geometric shapes. The experimenter showed a picture from the stimulus set and said, "What is this?" and recorded correct responses. Weekly follow-up probes for 4 weeks were conducted after the mastery criteria had been achieved. Results indicated the 100% criterion showed at or above 70% for all participants. The 80% and 60% criteria varied widely, with one participant reaching up to 100% accuracy with the 80% criterion, while one showed 0% accuracy. In summary, this experiment provided extra evidence for using a 100% criterion, and against using an 80% or lower criterion.

Experiment four was identical to experiment three except for the use of a 90% criterion in place of the 60% criterion, and a few procedural variations, including number of sessions per day, and only one maintenance probe. Vocal tacting was again taught, but the stimuli differed from previous experiments (states, black and white symbols, and land formations). Once a stimulus set was mastered a non-experimental stimulus set was introduced so that the same number of targets were always being taught. Results indicated the 100% criterion maintained at 70% or higher response accuracy. The 90% criterion resulted in drops to 0% accuracy for three of the four participants, while the fourth showed 40%. The 80% criterion showed drops to 0% accuracy for three of the four participants, and 20% for the fourth, once again extending and replicating the

increasingly warranted use of a 100% mastery criterion while also giving evidence (albeit early) against even a 90% criterion, which may not be sufficient.

While Richling et al. (2019) contributes much to the current literature, there are still gaps. For example, the type of skill being taught (auditory-visual discrimination) is not the only kind of skill to be taught in an EIBI clinic. Instruction following skills are an important part of one's educational learning repertoire and should be considered when examining the relationship between mastery criteria and maintenance. Furthermore, according to their first experiment only about 1% of practitioners surveyed use a mastery criterion below 80% (Richling et al., 2019), yet they used 60% as a target for two of the three experiments they conducted.

Pitts & Hoerger (2021) conducted a study that was essentially a replication of the work done by Richling et al. (2019), with a few variations. Before each session a Paired Choice Preference Assessment (DeLeon & Iwata, 1996) was conducted. Three sets of three stimuli were used as targets, and common DTT procedures were used. Each stimulus set and mastery criterion was counterbalanced across individuals. 80%, 90%, and 100% mastery criteria were used. They required the first response in the 80% and 90% conditions to be correct for fear that if the first trial was incorrect and required prompting that it "wouldn't necessarily indicate true skill acquisition, and could have over inflated progress" (Pitts & Hoerger, 2021, p. 526; Richling et al., 2019). Maintenance probes were conducted each week for four weeks after mastery had been achieved. Sessions were 10 trials each. Results indicated that, for all participants, when set to a 90% or 100% criterion the targets maintained at or above the levels observed during teaching

sessions, while the 80% criterion showed variable or downward trends. Although many of the results are similar or the same as those found by Richling et al. (2019). One major difference present in the results obtained by Pitts & Hoerger (2021) is that the 90% criterion maintained relatively well, warranting further research on this particular piece of evidence. Further, instruction following skills were again neglected while auditory-visual discrimination tasks were the target, leaving room for more examination of instruction following skills and mastery criteria.

Fuller and Fienup (2018) were the forerunners for the body of research examining mastery criteria effects on maintenance in children with disabilities. They examined 50%, 80%, and 90% mastery criteria while using sight words as their target responses, either requiring the learner to read or spell the word when presented. Three participants participated, and weekly follow-up probes following mastery were conducted for 4 weeks. Results indicated that for all participants the 90% criterion produced the most consistent and highest maintenance results, while the 80% and 50% criterion varied widely. Again, the literature here shows that there is still a void surrounding mastery criteria effects on maintenance of instruction following skills. The same could be said regarding how few clinicians use anything lower than an 80% mastery criterion (Richling et al., 2019), bringing their decision to test a 50% mastery criterion under question.

While the data collected by Richling et al. (2019), Pitts and Hoerger (2021), and Fuller and Fienup (2018) have shown that an 80% mastery criterion may not be sufficient to produce quality outcomes, there are still gaps in the literature surrounding mastery criteria effects on maintenance. The aforementioned authors used receptive identification,

auditory visual conditional discrimination tasks, and literacy skills as their teaching targets. Therefore, no one to date has conducted a study that has examined the relationship between differing mastery criteria and its effect on the maintenance of instruction following skills. These skills are important for those with intellectual and/or developmental disabilities (like ASD) to learn so they can better function in society. Without learning how to at least follow simple 1-step instructions a child's ability to learn in a typical classroom setting will be diminished. Rehabilitation cannot fully take place if these skills are ignored. In order to fully service the children we work with, we need to consider all aspects of their learning and how we can improve the chances of the taught skills maintaining for years to come.

The standard for what constitutes an evidence-based practice as proposed by Horner et al. (2005) that "(a) a minimum of 5 single-subject studies… (b) …conducted by at least three different researchers across at least three different geographical locations, and (c) the five or more studies include a total of at least 20 participants" (p. 176) has not yet been met for this subject. As such, the body of research for mastery criteria effects on maintenance requires additional research.

Research Questions

The gap in the literature related to optimizing mastery criteria may lead clinicians to use criteria that are not evidence based due to the lack of evidence surrounding this topic. It can also be a pitfall for those who are not as familiar with behavior analytic practices, yet require a specific criterion be used (government agencies/policies, IEP

teams etc.). If a mastery criterion is too low it may not lead to meaningful maintenance of the skills being taught, requiring reteaching in the future. This becomes a time-consuming practice leading to missed opportunities to teach new skills and could even result in children finishing their time in the clinic and not being able to recall how to use skills that were taught to them. The time that children spend in an EIBI clinic (2-3 years) is relatively short compared to the amount of time spent in a public education system where they will likely receive little, if any, training on functional skills. This means that we must make the most of the time we have with the kids in their early years as it is critical for their future opportunities and true rehabilitation.

Another consideration is even though a 100% mastery criterion may show better maintenance, there are more things to consider alongside maintenance, such as time and effort. Perhaps using a 100% criterion is too stringent and doesn't leave enough time for new skill teaching. What if the 90% criterion is a better compromise between results and efficiency? Because of the short amount of time these kids spend in EIBI clinics it is crucial to be as efficient as possible in teaching the students varying types of skills. Do mastery criteria have a different effect on how instruction following skills maintain compared to other types of skills, like receptive identification and literacy skills? This study aims to extend the current literature, providing extra evidence for abolishing a seemingly traditional practice that is not evidence-based.

Method

Participants and Setting

Participants in this study included two children with ASD. Both participants, Winston and Reginald (given pseudonyms to protect their identities) were 3 years old. Although diagnoses were not required, both children's caregivers reported that they have ASD. Both participants were recruited through email notices to clients in on-campus clinics and families in a local school district, both having the intended population (3-5 years old). Each participant received DTT instruction for 3-5 sessions per day, 3-5 times per week, with sessions being 10 trials. The participants received this instruction either in their home, their preschool, or in their school. All sessions were conducted by the researcher.

For Winston, baseline and teaching sessions were conducted at his school in a small, approximately 4x6-foot room. The room consisted of a desk and two chairs. All but one of his maintenance trials were conducted at the kitchen table of his home. For Reginald, all sessions were conducted at his preschool in an approximately 9x9-foot room. The room had many toys (serving as the preschool's play room), and one table with a few chairs.

Target Responses and Data Collection

One defining feature of this study that separates it from the work done by Richling et al. (2019), Pitts and Hoerger (2021), and Fuller and Fienup (2018) is the targeted responses for each participant. While the above-mentioned authors used receptive identification, auditory visual conditional discrimination tasks, and literacy skills as their teaching targets, this study tested the differing mastery criteria effects on the maintenance of 1-step instruction following. The following targets were selected

because they are generally easy to perform regarding motor capability, commonly used in classroom and/or other social settings, and require no additional materials. Target responses for both participants were “touch head,” “wave”, and “clap hands.” “Touch head” is defined as the participant raising one or both hands and making physical contact with the top, back, or sides of their own head when given the instruction “touch head.” “Waving” is defined as the participant raising one hand and moving it from side to side when the researcher says “wave.” The discriminative stimuli for this target was originally waving and saying “hello” to the child, expecting them to wave back, but was changed to simplify the expectation. “Clap hands” is defined as the participant bringing both hands together one or more times with the palms facing each other with enough physical force to produce an audible sound when given the instruction “clap hands.”

Data were collected on correct independent, correct prompted, and incorrect responses during discrete trials. A correct independent trial was defined as a correct response within three seconds of the instruction being given without any prompting other than the instruction. A correct prompted trial was defined as any trial where the researcher used any prompt in the prompting hierarchy (model, partial physical, full physical) and the participant does the trial along with the prompt. Incorrect trials were defined as any trial in which the learner emits a response that does not accurately match the instruction given (e.g., the participant claps their hands when given the instruction “touch head,” or refuses to allow physical prompts). If no response was made, then the trial was scored as an incorrect trial after 3 seconds of no responding. Data were collected on data sheets denoting the date, the step to follow, and the discriminative stimulus for

that instruction. A "+" was scored if the researcher gave the instruction and the participant completed the trial independently and correctly. A "P+" was scored if the trial was completed by the participant with a prompt. A "-" was scored for incorrect trials, trials where the participant refused physical prompts, or trials in which the participant did not respond for 3 seconds. The number of correct trials divided by the number of total trials was turned into a percentage to determine whether mastery had been met according to the mastery criterion applied to that discrete trial.

Interobserver Agreement

Participants were given a video release form to sign, giving consent to being recorded. This allowed for recordings to be coded and scored for the purpose of determining treatment integrity and interobserver agreement (IOA). Treatment integrity was examined after the sessions were all completed by way of the video recordings. Researchers were given a treatment integrity checklist outlining each step for the researcher to take during any given session. Researchers were graded on each opportunity to complete a step of the procedure, receiving either a "+" for engaging in the appropriate step after an opportunity or a "-" for missing an opportunity to engage in that step. The total number of "+s" and "-s" was divided by the number of "+s" and then multiplied by 100 to obtain a percentage. Treatment integrity was calculated for 27% and 25% of sessions for Winston and Reginald, respectively. Mean treatment integrity was 97% for Winston (range 91-100%). For Reginald, mean treatment integrity was 96% (range 93-100%). IOA was also calculated by using the video recorded sessions. The researcher's data collected on discrete trials was compared to a second observer's data. The total

number of data points counted by both researchers (in this case total number of discrete trials observed by both researchers during a given session) was divided by the total number of agreements and then multiplied by 100 to obtain a percentage score. The mean IOA for Winston was 96% (range 90-100%). For Reginald, the mean IOA was 94% (range 80-100%).

Design and Procedures

Experimental Design

This study conducted a partial replication of Richling et al. (2019). A multiple-baseline design across participants with an embedded multi-element design was used. This allowed for within-subject treatment comparisons, and between-subject replications. The experiment consisted of three conditions: baseline, teaching, and weekly follow-up probes/maintenance checks.

Stimulus Preference Assessments

A Multiple Stimulus Without Replacement (MSWO, DeLeon & Iwata, 1996) preference assessment was conducted prior to each session for Reginald. MSWOs involve placing an array of items in front of the participant and giving the instruction "pick one." Whichever stimulus is picked (whether it be food or a tangible item such as a toy) is given to the participant to interact with. After 15 s with a food item, or 30 s with a tangible item, the item is removed and a new array without any previously chosen item is presented. The pick order is tracked, and the first three items picked are used as reinforcement for baseline and teaching sessions. Although a MSWO was conducted before each session with Reginald, he possessed the skills to tell me what item he wanted

to work for before each set of trials, so items other than his top three (according to the MSWO) were often used as reinforcers.

For Winston, after conducting several MSWOs, no more were conducted for a few reasons. Winston often showed problem behavior during the MSWOs when preferred items were taken away and a new trial without that preferred item would begin, setting the stage for a session full of non-compliance, distress, and, in some cases, physical aggression towards the researcher or his caretaker, who was present for all sessions. Further, it became evident that no items offered to Winston, other than his personal tablet, was rewarding enough to him to engage in the task demands being placed upon him. It was the only item he would pick when it was an option and was by far the most motivating reinforcer.

Baseline

During baseline sessions, participants were given 15s or 30s of access to food items and/or tangibles respectively, based on their preference assessment results, noncontingent of behavior on a 30s fixed interval schedule. This was done to help decrease the likelihood of noncompliant behavior occurring during teaching sessions (Richling et al. 2019). This also helps the learner pair the experimenter with reinforcement, thus increasing the likelihood of the experimenter serving as a discriminative stimulus (Sd), signaling the availability of reinforcement for the learner. Winston had two sessions of baseline while Reginald had four sessions of baseline.

Teaching

Teaching sessions were similar to baseline, with a few procedural variations. First, each participant was given praise on a fixed-ratio (FR) 1 schedule for independent correct trials. For Winston, he was given access to his preferred reinforcers on a variable-ratio (VR) 1 schedule for the first three full sessions. During the first three full sessions he was given reinforcer access after prompted trials on some occasions. Prompted trials no longer earned reinforcement delivery after the first three full sessions. His reinforcement schedule was thinned to a VR3 within three more sessions. Similarly, he was later often offered a small edible alongside his tablet after he manded for an edible during a session. Cooper et al. (2007) defines a mand as a type of verbal behavior where a learner demands, states, or implies a want for a specific reinforcer. This was something he had not yet done at that point. Winston was given 30 s with his tablet, unless he had finished a target session (10 trials of a single target). He was originally given the tablet for two minutes after finishing a target session, which was then thinned to one minute a few sessions later. Reginald was given edible reinforcers for 15 s, and tangible reinforcers for 30 s on a VR3 schedule for independent trials.

Incorrect trials (either independent or from lack of responding after 3s) resulted in the use of a least-to-most prompting procedure, going from gestural to partial-physical, and then full-physical prompts. Prompted responses resulted in less salient praise (i.e. saying “good job” with less enthusiasm than if the trial had been independent). These teaching sessions were at least 10 trials long while never ending with a response that was not an independent correct response (this was true of all sessions except for one, where Winston’s caretaker had to leave due to time constraints). Teaching sessions lasted until

the participant attained the designated mastery criteria (80%, 90%, or 100%) for three consecutive sessions. Weekly follow-up probes then began while the other targeted skills were being taught. Winston's targets for the 80%, 90% and 100% criterion were "touch head," "wave," and "clap hands" respectively. Reginald's targets were "clap hands," "touch head," and "wave" for 80%, 90%, and 100% mastery criteria respectively.

Weekly Follow-Up Probes and Maintenance

Weekly follow-up probes consisted of a single 10-trial session. They occurred at one, two, three, and four-week intervals after attaining the mastery criteria of each individual skill. This was true for all follow-up sessions but the first one for Winston, due to school being canceled, which resulted in the follow-up being nine days apart, rather than seven from the time he initially mastered the skill.

Results

Figure 1 displays the results of the teaching sessions and weekly follow-ups for Winston and Reginald. Table 1 shows how the targets were counterbalanced to allow for within-subject treatment comparisons and between-subject replications. Winston's targets were "touch head," "wave," and "clap hands" for the 80%, 90%, and 100% criteria respectively. Reginald's targets were "clap hands," "touch head," and "wave" for the 80%, 90%, and 100% criteria respectively. Winston obtained the mastery criteria for each target within 14 teaching sessions. Of note, his designated 100% criterion ("clap hands") was acquired the fastest of all targets, obtaining mastery six sessions earlier than the next mastered target. Conversely, the 90% criterion target ("wave"), was the hardest target for

him master, taking all 14 sessions to reach the mastery criterion. For the weekly follow-ups, Winston's 100% criterion target maintained at 100% for the first two follow-ups, and then fell to 90% for the third and fourth follow-ups. Winston's 90% criterion target ("wave") saw an immediate decrease to 30% for the first session, and then fell to 0% for all remaining maintenance checks. Lastly, the 80% criterion target ("touch head") maintained at 100% for the first two sessions, and then fell to 20% and 0% for the final two sessions respectively.

Figure 2 shows the results of teaching sessions and weekly follow-ups for Reginald. It took Reginald four teaching sessions to obtain the mastery criterion for all targets. His first target, "clap hands," was the lowest scoring across all targets. During the weekly follow-ups, Reginald maintained a 100% mastery for all targets across all four follow-up sessions.

Discussion

This study attempted to replicate Richling et al. (2019) but has failed to bring any further light to the question of whether an 80% mastery criterion is sufficient for producing meaningful maintenance results of skills taught. Our results were highly variable, making it hard to make any claims about the efficacy of differing mastery criteria effects on maintenance of 1-step instruction following. Looking at Reginald's data, it seemed to not matter what target he was being taught. He met the 100% criterion for all targets and at the same rate, only requiring one prompt for each new target. Then, for his follow-up sessions he continued to show 100% accuracy. This makes it next to

impossible to make any claims about whether any criterion is better than another. Similarly, because mastery was met at the same rate across targets, we cannot speak to the efficiency of any one criterion. The likely reason this occurred is due to Reginald's current skill set. It was obvious from baseline sessions that these targets would not be very hard for him, but he exceled at a rate that the researchers did not anticipate. It is likely that these targets were too easy for him, despite his caregiver denoting via the pre-experiment screener that he could not do any of these targets consistently.

Winston's data is similarly confusing. His 100% criterion target was mastered the fastest, implying that he has had a previous learning history with that target. He actually had 100% across four sessions because of a logistical error from the experimenter. His 80% and 90% criterion targets were closer to what was anticipated, taking 13 and 14 sessions respectively to master. Anecdotally, Winston's 90% criterion target ("wave") was the hardest for him master, and this was similarly shown during his weekly follow-up sessions. However, it is important to note that during the weekly follow-ups, which were not conducted in the same location as teaching sessions, he had significantly more non-compliance and refusal of work. This was true for the second follow-up session for the 90% criterion as well as the third and fourth sessions of the 80% criterion. The 100% criterion target during follow-up sessions maintained at 100% for weeks one and two, and 90% for weeks three and four, which was expected based on findings of previous research like Richling et al. (2019). With all of this information it is hard to make any conclusions about the efficacy and efficiency of each mastery criterion.

Even when comparing both participants' data next to each other, it is hard to gather any meaningful information from these two data sets. This is why single-subject designs rely on having 3 or more participants, and although many attempts were made at recruiting, these two participants were the only candidates for the current study. Further, this is a reason that single subject-design also relies on systematic replications. That being said, even with a third participant showing data closer to what was anticipated the results would still be hard to interpret/gather any meaningful data from. Therefore, any extrapolations of the data contained within the study should be interpreted with caution.

It is possible that the proper mastery criterion is something that should be considered on a case-by-case basis, or as relating to the type of target being taught (i.e., 1-step instructions, or auditory visual discrimination tasks). Further, with the data from this study we cannot make any claims about time efficiency across the most common mastery criteria (Richling et al., 2019). Both participants defied the idea that a 100% criterion may take too long to teach, and not leave as much time for other targets to be taught. Although this would normally be a good claim, with only two participants, and when comparing data from other research in this area, both participants attaining the 100% criterion so quickly is atypical of what has been seen in other data sets (Fuller & Fienup, 2018; Richling et al., 2019; Pitts & Hoerger, 2021), making it difficult to claim that a 100% criterion is just as time efficient as another criterion.

Limitations

This study had a number of limitations that should be discussed, and considered for future replications. First, the amount of time the researchers had to teach targets to

participants was relatively short. Because this study was done independently (i.e., not in a clinical setting), the amount of DTT instruction being given was likely not sufficient to fully build up the skills of the participants. Because so few targets were taught, this led to situations where non-compliant behavior became more reinforcing than actually working for rewards. In other words, because of the frequency of the sessions (3-5 per day, 3 days per week) the expectations and contingencies were not established as well as in a clinical setting where the child spends significant amounts of time in a daily routine.

For example, many of Winston's maintenance checks were impacted by him refusing to do any of the work. This was further reinforced during maintenance sessions because, unlike teaching sessions, the only goal of the maintenance session was to get 10 trials of a target and then be done. During teaching sessions, the session did not end until an independent response was made (one of the initial teaching sessions took 341 trials of single target). This means that because the researcher was no longer requiring an independent response before finishing a session or giving him prompts during maintenance sessions, Winston's motivation to work for a reward was diminished because he learned that not responding would get him the same reinforcement as responding, but with less work. In fact, the procedures demanded that after 3 s of no responding the trial be counted as incorrect. In theory, this means Winston could wait 30 s before the maintenance check ended and would then have free access to reinforcers we were previously withholding contingent on work completion. In a clinical setting this would be mitigated by returning from a maintenance check to normal teaching sessions,

establishing expectations better and reducing the likelihood of reinforcing non-compliant behavior.

Second, there is an inherent flaw with trying to teach a skill to a specific mastery criterion. This was highlighted by Reginald's results. He mastered all three targets with 100% accuracy across three sessions. This then makes it impossible to make any claims about any single mastery criterion's effectiveness because the researcher failed to teach a specific target to anything less than 100% accuracy across three sessions. In order to do so a researcher would have to purposefully do something to make a trial or two incorrect in order to bring down the score to the designated mastery criterion, which then defeats the purpose of teaching a skill, and similarly could create an experimental confound. Doing so would confuse the learner and could make it so that they are not sure what response to give when given a specific discriminative stimulus because they no longer know how you will react. Even Winston's 90% target was taught to 100% accuracy across 3 sessions. Again, this makes it hard to make any claims about the efficacy of a specific mastery criterion. Perhaps another researcher with more time and resources could look at a learner's data and find targets that have already been mastered, matching specific mastery criteria, and use that to inform how differing mastery criteria affect maintenance.

Similarly, the procedures for this study made it impossible for a participant to meet 90% across three sessions. Beginning a session with an incorrect response then means that, regardless of the results of the current session, it will take at least 3 additional sessions before the mastery criterion is met. This is because a prompt would be given

after the incorrect response. That means that 2/10 trials for that session have been done, leaving only eight trials for independent responses, thus making it impossible to ever meet 90% for three sessions. Because reaching 90% across three sessions was not possible, there is no way to claim any effectiveness for this specific criterion given the current procedures. This seemed to not be a problem for Richling et al. (2019), despite the fact their procedures were very close to the present study. However, their data shows that 90% across three sessions was met by most participants. Future researchers should consider this when designing their protocol.

Third, part of the procedures initially matched that of Richling et al. (2019), having "non-experimental" targets to teach while maintenance sessions were being conducted once a target was mastered. After a few sessions of working with Winston it became evident that he struggled discriminating between targets when the only prompt was a vocal prompt. Thus, adding in more vocal prompts for targets that we were not concerned with measuring could have adversely affected his ability to learn the targeted skills. This mostly has to do with the time and resources available to the research team. To do so would have been unfair both to him and the research team because we would have clinical expectations for Winston while not being able to fully provide a clinical experience (time, resources, expertise) to him, which. Because of this, the process of teaching "non-experimental" targets was removed from the current study but should be implemented by future researchers with greater time and resources available to them given the benefits of doing so (Richling et al., 2019).

Lastly, there were some discrepancies regarding typical procedures for clinical settings using DTT that were not done in the current study. First, during teaching sessions each target was taught one full session at a time, rather than rotating through targets (i.e., "touch head, great job! "Clap hands, awesome work, here is your tablet!"). Instead, one target would be repeated several times with praise between each independent correct response. Second, during maintenance trials, reinforcement was still given on a VR3 schedule, meaning only a few variations between teaching and follow-up sessions existed. This should have been made clearer in the procedures by the researchers. Once the mistake was realized, the researchers decided it was best to continue how it had been done to maintain consistency.

Future Directions

Because this study was not able to replicate preliminary work done by Fuller and Fienup (2018), Richling et al. (2019), and Pitts and Hoerger (2021), there is still room for research in this area, particularly that of differing mastery criteria effects on maintenance. Because this study did not provide meaningful results, more data and replications are needed to formalize an evidence base for mastery criteria (Horner et al., 2005).

Future studies should replicate this study, using 1-step instructions to test differing mastery criteria, as opposed to using other types of tasks such as auditory-visual discrimination tasks, like the work done by Fuller and Fienup (2018), Richling et al. (2019), and Pitts and Hoerger (2021). As far as we are aware, to date this is the only study to have tested differing mastery criteria effects on 1-step instruction following

skills, warranting further examination. Researchers conducting these replications should carefully consider the methodological limitation of the current study.

Similarly, future research may also consider tailoring the targets to the skill level of each learner, using the same type of target, but not the same target, to account for instances where a learner may already have, unbeknownst to caregivers or researchers, a learning history with specific targets. Or to avoid instances where the targets are simply easier for one learner than another due to their current skill sets. One way to do this would be to use a more stringent screener when recruiting participants. Including other caregivers and stakeholders in the screener could help prevent this issue from occurring. Similarly, probing for the targets during baseline sessions would also be appropriate. For example, inviting classroom teachers to also participate in the screener could help the researcher knowing how to program their methods.

Finally, future research could use more than one type of target and teach more than the three targets for a comprehensive determination of differing mastery criteria effects on maintenance across different types of targets. For example, in a clinical setting it would be feasible to have six targets, each being a different type of task (auditory visual discrimination, gross motor skills, and fine motor skills). These additional targets would provide more information about the maintenance associated with each level of mastery criteria and would be useful in cases where some skills are mastered too quickly for a certain mastery criterion to be used.

The current study was the first to examine how differing mastery criteria affects the maintenance of 1-step instructions in children with disabilities. This evaluation

included only two participants and failed to gather any meaningful results support of the claim that an 80% mastery criterion may not be sufficient to produce meaningful maintenance of skills taught to children with disabilities. Additional research is needed to provide a strong enough evidence base to discourage the use of an 80%, and in some cases a 90% mastery criterion, the most commonly used mastery criteria among many EIBI clinics in the United States (Richling et al., 2019). Providing such data could allow behavior analysts to provide better treatment and "make programming decisions based on evidence rather than on lore" (Richling et al., 2019, p. 716).

References

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th edn.). Washington, DC: American Psychiatric Association.

Brodhead, M. T., Higbee, T. S., Pollard, J. S., Akers, J. S., & Gerencser, K. R. (2014). The use of linked activity schedules to teach children with autism to play hide-and-seek. *Journal of Applied Behavior Analysis*, *47*(3), 645–650. https://doi-org.dist.lib.usu.edu/10.1002/jaba.145

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

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). Applied behavior analysis.

DeLeon, I. G., & Iwata, B. A. (1996). Evaluation of a multiple-stimulus presentation format for assessing reinforcer preferences. Journal of Applied Behavior Analysis, 29(4), 519–533. https://doi-org.dist.lib.usu.edu/10.1901/jaba.1996.29-519

Eldevik, S., Hastings, R. P., Hughes, J. C., Jahr, E., Eikeseth, S., & Cross, S. (2009). Meta-analysis of early intensive behavioral intervention for children with autism. Journal of Clinical Child and Adolescent Psychology, 38(3), 439–450. https://doi-org.dist.lib.usu.edu/10.1080/15374410902851739

Fienup, D.M., & Broadsky, J. (2017). Effects of mastery criterion on the emergence of derived equivalence relations. Journal of Applied Behavior Analysis, 40, 843–848. https://doi.org/10.1002/jaba.4

Fienup, D. M., & Carr, J. E. (2021). The use of performance criteria for determining

"mastery" in discrete-trial instruction: A call for research. *Behavioral Interventions*, *36*(4), 756–763. https://doi-org.dist.lib.usu.edu/10.1002/bin.1827

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

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

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

Love, J. R., Carr, J. E., Almason, S. M., & Petursdottir, A. I. (2009). Early and intensive behavioral intervention for autism: A survey of clinical practices. Research in Autism Spectrum Disorders, 3(2), 421–428. https://doiorg.dist.lib.usu.edu/10.1016/j.rasd.2008.08.008

Perry, A., Koudys, J., Prichard, A., & Ho, H. (2019). Follow-up study of youth who received EIBI as young children. Behavior Modification, 43(2), 181–201. https://doi-org.dist.lib.usu.edu/10.1177/0145445517746916

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

Noyes-Grosser, D. M., Elbaum, B., Wu, Y., Siegenthaler, K. M., Cavalari, R. S., Gillis, J.

M., & Romanczyk, R. G. (2018). Early intervention outcomes for toddlers with autism spectrum disorder and their families. Infants & Young Children, 31(3), 177–199. https://doi-org.dist.lib.usu.edu/10.1097/IYC.0000000000000121

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://doiorg.dist.lib.usu.edu/10.1002/jaba.580

Schneider, A., Fienup, D. M., Gussin, R., & Moss, P. (2021). A preliminary comparison of mastery criterion frequency values: Effects on acquisition and maintenance. *Behavioral Interventions*. https://doiorg.dist.lib.usu.edu/10.1002/bin.1834

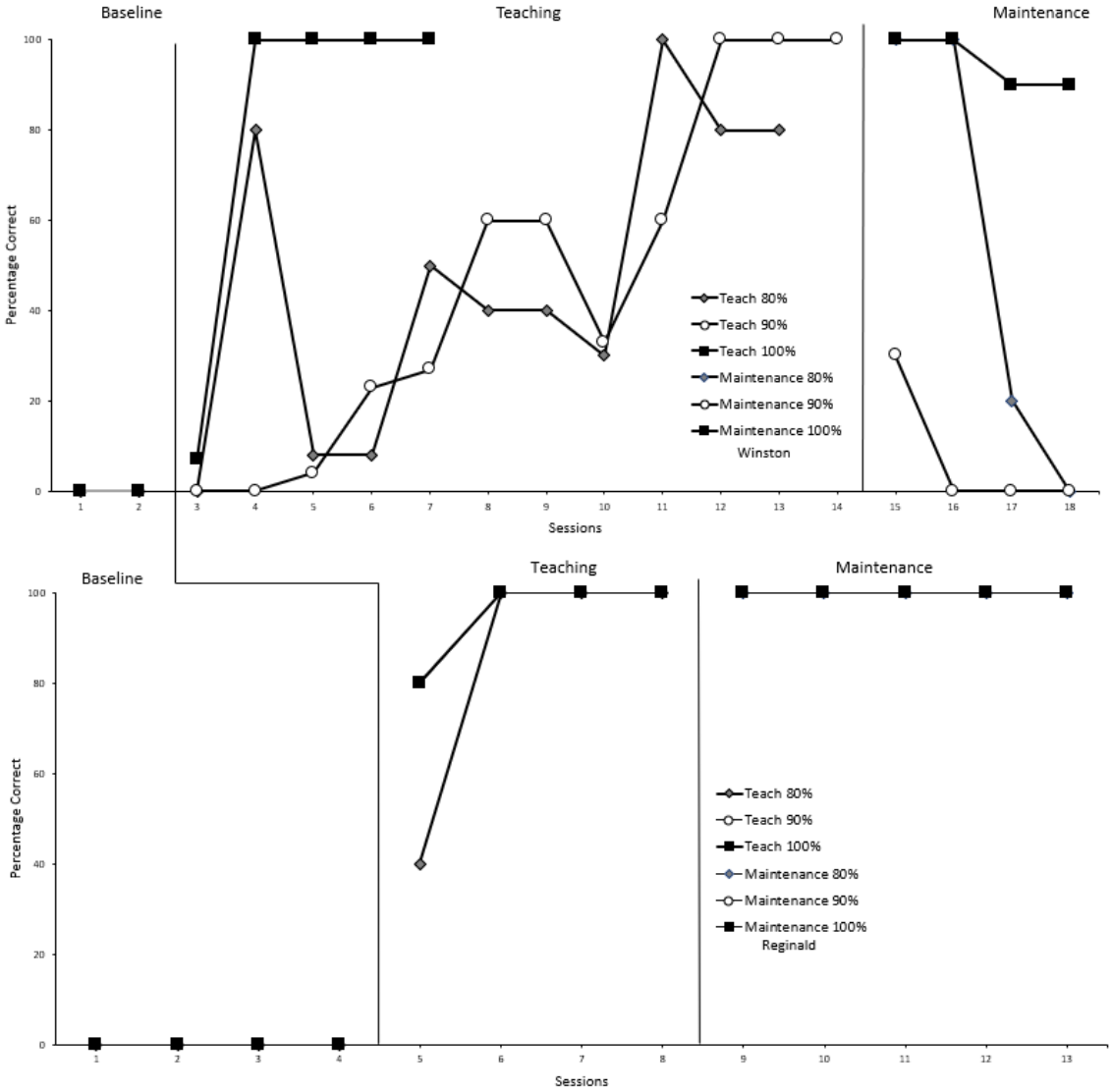
Semb, G. (1974). The effects of mastery criteria and assignment length of college-participant test performance. Journal of Applied Behavior Analysis, 7, 61–69. https://doi.org/10.1901/jaba.1974.7.6

Smith, D. P., Hayward, D. W., Gale, C. M., Eikeseth, S., & Klintwall, L. (2021). Treatment gains from Early and Intensive Behavioral Intervention (EIBI) are maintained 10 years later. *Behavior Modification*, *45*(4), 581–601. https://doi-org.dist.lib.usu.edu/10.1177/0145445519882895

Figures

Figure 1

Differing Mastery Criteria Effects on Maintenance



Note. Percentage of correct responses for the 80%, 90%, and 100% across three-sessions mastery criteria, and weekly follow-up series.

Table 1.

Targets and Criteria

Participant	*80%*	*90%*	*100%*
Winston	Touch Head	Wave	Clap Hands
Reginald	Clap Hands	Touch Head	Wave