Establishing Performance Criteria for Skill Mastery

22

Sarah M. Richling, Daniel M. Fienup, and Kristina Wong

What Are Performance Criteria?

Applied behavior analysis (ABA), as a field, focuses on addressing socially significant behaviors through the implementation of effective behavior analytic interventions (Baer et al., 1987). The functional extension of effectiveness of treatment procedures within the therapeutic context to the natural environment is a central goal of ABA (Stokes & Baer, 1977). Within skill acquisition programming, the basic structure of instruction includes clear antecedents, opportunities for the learner to respond, and feedback on responses (Skinner, 1968). This may involve a child with autism responding to repeated opportunities to tact colors, college students responding to quizzes, or youth learning accurate sports behaviors. In each case, the respective instructor continues with teaching until the learner's behavior meets a predetermined criterion used to evaluate sufficient skill proficiency, often labeled a *mastery criterion* or *performance criterion.* This is systematically done such that behavior continues to occur, or maintains, in the presence of naturally occurring antecedents and consequences, outside of contrived analogue teaching conditions. This process has often been labeled as *programming for maintenance,* or more accurately, *response maintenance* (as discussed in the section "Terminology Considerations").

The use of instructor-determined performance criteria that serve as discriminative stimuli for terminating a teaching phase is a widely adopted practice and has a long history in the field of behavior analysis. In fact, the first article printed in the *Journal of Applied Behavior Analysis* includes a reference to the discontinuation of programmed treatment once a satisfactory rate of behavior was achieved (Hall et al., 1968, pp. 2–3). In 1997, Sayrs and Ghezzi noted the rapid growth in the reporting of mastery criteria in the *Journal of Applied Behavior Analysis* between 1968 and 1995. Around this same time in 1996, 53% of articles in the *Journal of the Experimental Analysis of Behavior* included a report on mastery criteria (Rehfeldt & Ghezzi, 1996). Early treatment manuals (e.g., Lovaas, 1981) also included recommendations for adopting performance criteria such as 9 out of 10 consecutive trials correct as indication to move onto the next step of training. More recently, clinical survey data (Love et al., 2009; Richling et al., 2019) suggest the wide adoption of evaluative performance criteria. These studies report that all survey respondents indicated utilizing criteria of various types, such as those based on a percentage of trials correct or a consecutive number of trials

S. M. Richling (✉)
Department of Psychological Sciences, Auburn University, Auburn, AL, USA
e-mail: smr0043@auburn.edu

D. M. Fienup · K. Wong
Teachers College, Columbia University, New York, NY, USA

J. L. Matson (ed.), *Handbook of Applied Behavior Analysis*, Autism and Child Psychopathology Series, https://doi.org/10.1007/978-3-031-19964-6_22

correct. As such, it is apparent that the use of performance criteria has a long-standing history and has become ubiquitous within the field of ABA.

Contemporary Use of Performance Criteria

Recently, researchers and clinicians have referred to predetermined performance goals as *mastery criteria.* Many performance criteria rules appear to come from clinical manuals or supervisors and seem to be accepted as universal rules (e.g., 80% correct responding or above for three consecutive sessions); however, these rules have little scientific support. The selection of performance criteria is undoubtedly nuanced and should be tailored to each unique behavioral target and each unique client. That is, the selected goal should be directly tied to how exactly this particular behavior is expected to occur, by this particular individual, in a particular natural context(s), at a particular time(s), in a particular way. For example, it may be the learner is expected to pass a written exam with a grade of B or better, requiring 80% accurate response on test items within a certain period of time. As such, the acceptable level of responding would need to be 80% accuracy or higher and occur at a certain rate of responding, to account for the timed test conditions. Performance on both of these features (i.e., accuracy and speed) of the response would need to be measured and observed in order to determine whether an intervention has been effective. In another case, it may be that an individual on a behavior analyst's caseload is being taught to cross the road safely. In this case, it would be important for the behavior analyst to require 100% accuracy with respect to safe behaviors while also ensuring programming is conducted such that behavior is likely to maintain 100% accuracy for a long duration of time post-training. There is less clinical need to require such stringent criteria for other topographies of behavior such as tacting animals, for example.

It is important we also recognize that for the above examples, it is possible a higher level of performance criteria must be established during teaching sessions in order to achieve desired levels of behavior which are expected to occur at a later time, accounting for behavioral deterioration over time. That is, there may be decreases in the accuracy and/or speed after a certain period has elapsed since the previous teaching session. Thus, the behavior analyst must also determine and assess what constitutes an effective teaching criterion that can reliably produce the desired response maintenance performance expectations at a later time during which the behavior actually needs to occur. In other words, we need to determine functional relations between performance criteria during teaching—when we reinforce correct responses and provide assistance as needed—and the resulting performances after teaching are done and the learner is expected to perform under more naturalistic conditions and rates of reinforcement. In behavior analysis, this initial teaching criterion has typically been referred to as the mastery criterion, or the criterion at which the learner must perform under teaching conditions before progressing to maintenance probe conditions. While these initial goals are important to indicate movement to the next phase of treatment, it is important to highlight here that the overall goal of demonstrating effectiveness does not end at the point of achieving a mastery criterion. Effectiveness is only demonstrated when the behavior occurs in the desired context following the termination of treatment. Only then might we consider the skill mastered. We discuss this and related issues with terminology in the following section.

Terminology Considerations

First, as pointed out by Cooper et al. (2007, p. 616) there is a need to distinguish between the terms *maintenance* and *response maintenance.* Response maintenance has been defined as *the degree to which* a behavior persists over time when all or part of the intervention variables responsible for training the behavior are no longer present (Freeland & Noell, 2002; Stokes & Baer, 1977). As such, it refers to a measurement of the occurrence of behavior and has also been

referred to as behavioral persistence or durability. Maintenance, however, does not refer to the behavior, but to the environmental stimulus conditions. Maintenance is utilized to describe a condition in which all or part of the treatment has been removed, albeit often with the intent of observing a potential degree of response maintenance. As such, response maintenance is best conceptualized as a dependent variable and maintenance as an independent variable. Thus, a primary goal of ABA interventions is the demonstration of a predetermined acceptable level of response maintenance, not merely the implementation of maintenance procedures.

Second, the term *mastery criterion* also warrants further discussion. As it stands, a mastery criterion has been loosely defined as "a specific guideline for performing a skill such that if the guideline is met, the skill is likely to be mastered" (Martin & Pear, 2007, p. 223) or as "performance requirements for practicing a skill such that if the criteria are met, the behavior has been learned" (Martin & Pear, 2007, p. 343). Fuller and Fienup (2018) highlight circumstances under which this term is used to describe performance criteria that do not meet this definition. For example, the authors state that once responding meets a predetermined level of accuracy, an instructor may move to a less restrictive prompt level, which does not suggest mastery, but rather behavior meeting an acceptable criterion given the current context. This highlights one of several problems with the current use of the term mastery criterion, namely the inclusion of the term *mastery*, itself.

Colloquially, mastery typically refers to the possession or demonstration of an exceptional skill or technique and one who can perform at this level may be referred to as a master of that skill or subject matter. Keeping this definition in mind, it is odd to refer to the minimally acceptable levels of performance as *mastery*. This is particularly curious when we consider this in some contextual scenarios. Oftentimes, a mastery criterion might be set at 80% correct (McDougale et al., 2019). If one were to go to work wearing 80% of their clothing, do we consider them a master dresser? Or if someone stops at 80% of stop signs while driving, would we label them a masterful driver? However, within behavior analysis, mastery typically refers to a level of performance indicating a behavior has been sufficiently learned. This gives rise to another question, what do we mean by "learned" (as well as "sufficient"). As behavior analysts, we may use the term *learned* to indicate observation of the behavior being evoked at acceptable levels in the presence of given contextual discriminative stimuli. However, *learned* does not necessarily mean the behavior has been acquired by the individual such that it will occur at the same levels ad infinitum.

To this point, as suggested by Fuller and Fienup (2018), there is a problematic underlying assumption inherent in mastery criteria. That is because behavior meeting this criterion functions as a discriminative stimulus for the teacher to engage in another behavior (e.g., decrease prompt level or introduce new targets), there is an implied expectation that behavior will maintain once the current instructional behaviors are terminated. The problem here, however, lies in the lack of literature supporting this assumption of maintenance following achieving specific performance criterion levels (Fuller & Fienup, 2018; Richling et al., 2019).

Now, let us look at the point during the training context at which the term mastery is typically utilized (i.e., after behavior is observed to occur at initial performance criteria levels under teaching conditions) and why this is problematic. Consider the possible response deterioration that is likely to occur following the removal of teaching procedures as described earlier. If deterioration in responding occurs, we may not label this mastery; it is just the first step in a series of teaching milestones toward a terminal goal. If the term mastery is used at this point in teaching, it may indicate to the therapist that their job is done, when that is far from the case.

Within precision teaching literature, specific attention is paid to performance standards and empirically associated learning outcomes (Kubina & Starlin, 2003). Fluency aims are conceptually similar to mastery criteria. Within this area of research, the assumption is that performance occurring within a certain frequency range

will be associated with retention and other extended learning outcomes (Kubina & Starlin, 2003). These learning outcomes include other behavioral measures such as retention across time, endurance of performance for a duration of time, performance in the presence of distractions, and application of a previously learned component in the context of learning a new composite skill (REAPS) (Binder, 1996; Haughton, 1984). These terms refer to various dimensions of performance (beyond a percentage of correct responses) that are expected to be observed before a skill is considered sufficiently learned. The importance placed on these other demonstrations of performance within the precision teaching literature highlights the need for adopting and clearly outlining the various dimensions of performance and milestones required for truly demonstrating mastery.

Mastery Redefined as a Collective of Multiple Performance Criteria

For the reasons outlined in the previous section, we have adopted the term *performance criteria* as an umbrella term encompassing individual criteria applied at separate learning goals. The term *mastery criterion* is reserved for describing the final milestone of having achieved all individual performance expectations. These individual performance criteria may include the initial acquisition criterion (e.g., 100% correct responding across three consecutive sessions), a fluency or rate criterion (e.g., 100% correct responding at a rate of 20 responses per minute), a response maintenance criterion (e.g., retained performance at 90% correct responding after a period of one month), and a generalization criterion (e.g., 90% correct performance in the natural environment, two additional novel settings, without the presence of the instructor, and in the presence of two novel individuals). If necessary, supplemental performance criteria may be adopted for the particular skills and expectations in the natural environment. For example, one may also require performance in the presence of distraction or for a duration of time without a decrease in rate. These features of performance might be expected, for example, for professional athletes or individuals taking long standardized tests. Throughout the remainder of this chapter, we will utilize the terms as described in this section. In the next section, we will address each of the abovementioned performance criteria in more detail.

Dimensions of Performance Across Which Criteria Can Be Applied

As suggested at the beginning of this chapter, conventional wisdom assumes a simple approach to the application of mastery criteria. The majority of practitioners and researchers within the field of ABA rely on percentages of correct responses as the standard for determining the mastery of any given skill (Richling et al., 2019). It is worth noting most of this work involves young children with developmental delays who are learning basic academic and social responses. Once an individual performs a task with 90–100% accuracy, it is typical to label the task as "mastered." What we will soon find, however, is that the application of mastery criteria is much more nuanced and complex.

Mastery encompasses several individual components for which a criterion should be uniquely established. As described in the previous section, we conceptualize mastery as a set of performances, comprised of acquisition, fluency, maintenance, and generalization (Fig. 22.1).

The initial stage of teaching requires the implementation of strategies to produce a response that was not previously in a learner's repertoire, also known as the acquisition stage. When a target behavior is in the acquisition stage, that is, the skill has yet to be performed successfully, acquisition criteria should be applied to determine when the intervention should be faded or terminated. The most commonly used dimension of acquisition criteria within ABA instruction is the level of accuracy during a session. Instructors administer a block, or set number, of teaching trials. Typically, practitioners and researchers report a percentage of correct responses across all trials within the session.

Fig. 22.1 The four pillars of mastery including acquisition, fluency, maintenance, and generalization for which instructors should establish a unique criterion for each

Higher percentages tend to lead to more durable responses as more time passes (Richling et al., 2019).

Another dimension of acquisition criteria identifies the frequency of observations at which the level of accuracy occurs. Practitioners and researchers typically establish a range of one to three consecutive sessions in which a predetermined accuracy level must be observed before they signal the termination of an intervention (Richling et al., 2019). To date, there have been no published studies systematically comparing the effectiveness of different frequencies of observation for producing subsequent response maintenance.

Selecting the right criterion for acquisition is nuanced. Many different variables should be considered during this process. Such variables include the type of novel skill that is targeted, the intervention procedure being used, and the skills the individual who is undergoing instruction possesses. For example, a 90% acquisition criterion may be adequate for an individual who is learning how to spell but certainly not adequate for an individual who is learning to stop at a stop light. An acquisition criterion of five consecutively correct responses to emit letter sounds may be appropriate for a student who demonstrates bidirectional naming (Miguel, 2016) but not for a student who does not demonstrate the ability to learn language incidentally. Bearing in mind all the nuances of skill acquisition, criterion selection needs to be a carefully thought out process.

The parameters of acquisition, which include the level of performance and frequency of observations should also be applied to the assessment of maintenance and generalization across settings and multiple instructors. For example, an instructor may establish a level-based criterion across two or more instructors to assess for the generalization across instructors. An instructor may also establish a level-based criterion across two or more settings to assess for the generalization across settings.

Performance criterion may be applied to a whole session/set of operants or to individual operants. Thus, identifying the unit of analysis when determining performance criteria is important (Wong, Bajwa et al., 2022). Level of accuracy may be conceptualized as a session-based unit of analysis. That is, the emphasis is on the overall accuracy within a session, and the

criterion is applied to a set of operants rather than a single operant. This particular method of analyzing performance raises some important issues. ABA instruction within educational settings typically uses discrete trial instruction (DTI) that relies on teaching multiple operants or skills (a set of operants or skills) within one session. If a session contains 20 trials, there are usually four or five operants included in a teaching set. When performance criteria are applied to the session as a whole and the established criterion is less than 100% accuracy, errors centered on certain operants may be overlooked. For example, during tact instruction for four novel stimuli in a 20-trial session, a 90% correct criterion allows a student to respond incorrectly one or two times. Sometimes those two incorrect responses may fall on only one operant. Thus, the student responded correctly only three out of five times (60% accuracy) to one operant. The 90% correct criterion across one session hides this fact and assumes the student has acquired the entire set.

Another major issue with this method of analysis is that it affects the efficiency of instruction. Oftentimes, the raw data of skill acquisition programs show that students acquire a few operants in a set quickly, while needing additional sessions to acquire the remaining operants. Because the acquisition criterion is not met due to the pattern of errors for the remaining operants, the instructor delivers unnecessary instruction for the same set of operants until the set-based criterion is achieved. Thus, Wong, Bajwa et al (2022) proposed a unit of mastery analysis that is applied to individual operants rather than a set of operants. When the unit of analysis is at the individual operant level, the acquisition of discrete novel skills is not affected by other skills that are taught within the same set. Similarly, trial-based criteria that identify acquisition in terms of the number of correct *consecutive* responses can be used instead of set-based criteria. For example, an instructor may determine an adequate point to terminate a shoe-tying intervention when the child independently emits three consecutive correct shoe-tying responses. Approximately 28% of ABA practitioners and 18% of ABA researchers utilize trial-based criteria (McDougale et al., 2019; Richling et al., 2019).

The nuanced and diverse nature of selecting appropriate performance criteria is further complicated as we considered areas of practice outside of autism and developmental disabilities. The following section identifies the various areas in which performance criteria have historically been adopted. After reviewing this literature, we will return to a discussion of a standardized model for selecting performance criteria across a wide variety of practice areas.

Review of Literature Targeting Performance Criteria

Performance Criteria with Individuals with Autism Spectrum Disorder and Developmental Disabilities

ABA treatments are highly effective for teaching individuals with developmental disabilities, intellectual disabilities, and autism spectrum disorder (ASD) novel and socially significant skills. Individuals who receive ABA services are typically expected to achieve an established performance criterion for each skill they are taught. Love et al. (2009) surveyed 200 professional supervisors of Early Intensive Behavioral Intervention programs to identify different aspects of their teaching procedures. Over 60% of the respondents used a performance criterion that was either a certain percentage of accurate trials across multiple sessions or a certain percentage of trials across multiple therapists. Almost all the respondents (98%) included teaching procedures that promoted maintenance and generalization of the target skills.

Performance criteria are ubiquitous within ABA research. However, as ABA services have grown in scale, the standards to which skill is deemed learned or mastered varies across researchers and practitioners. Richling et al. (2019) conducted an online survey to gather information on common clinical practices as they relate to skill acquisition and mastery criteria. Approximately 200 BCBAs (Board Certified Behavior Analysts) and BCBA-Ds (Doctoral Level Board Certified Behavior Analysts) who serve individuals with ASD and intellectual disabili-

ties responded. Similar to the results of Love et al. (2009), 68% of the clinicians used a session-based mastery criterion that was a certain percentage of accurate trials and 57% of those clinicians applied that mastery criterion across multiple sessions with additional variables. Only 35% of clinicians reported that they utilized a percentage of correct trials across multiple sessions. A small minority of clinicians (28%) used a certain number of consecutively correct responses to determine mastery and only 4% of clinicians used an established rate of correct responses per unit of time to determine mastery. There were also varied responses regarding the percentages used to determine mastery. Of the respondents who indicated that they applied a certain percentage of correct trials across multiple sessions, 52% of them used an 80% criterion. A smaller percentage of clinicians (28%) used a 90% criterion, and 7% of the clinicians used a 100% criterion. No clinicians applied a mastery criterion that was less than 80%. Richling et al. (2019) also sought to gain insight into the primary information source clinicians based on their mastery criterion. The primary source for the selection of mastery criteria for 44% of the respondents was a personal supervised experience. That is, many of the clinicians applied a particular mastery criterion because their supervisor directed them to do so. The second highest percentage of respondents (20%) reported that employer policies and requirements dictated their selection of mastery criterion. Sixteen percent of the respondents reported that graduate school training determined the established mastery criterion. A smaller percentage of respondents (10% and less) referenced continuing education programs, regulatory requirements, and funding sources as the primary information source for the mastery criterion.

To extend upon the responses that were submitted by the BCBAs and BCBA-Ds, Richling et al. (2019) conducted two additional experiments to systematically evaluate the most commonly reported mastery criterion level (80% across three consecutive sessions) with a 60% mastery criterion across three consecutive sessions and a 100% mastery criterion across three sessions on response maintenance. Four children with developmental disabilities were taught receptive identification skills and expressive identification skills (tacting). The results of Experiments 2 and 3 demonstrated that the only mastery criterion that produced reliably durable maintenance results (>70% accuracy) was the 100% mastery criterion across three sessions. A fourth experiment included a 90% criterion across three consecutive sessions in the comparison with an 80% and 100% criterion across three consecutive sessions. The results showed that even a 90% mastery criterion failed to produce durable maintenance responses. The 100% mastery criterion was the only criterion that predicted maintenance responses at or above 70% accuracy during 1-week follow-up sessions.

A similar study conducted by Fuller and Fienup (2018) demonstrated slightly different results. The authors investigated the effects of three skill acquisition mastery criteria (50% accuracy across one session, 80% accuracy across one session, and 90% accuracy across one session) on response maintenance and skill acquisition rate for students learning vocal and written spelling responses. The authors found differentiated maintenance responses across all three acquisition criteria and 90% accuracy across one session reliably predicted higher accuracy in responses 3–4 weeks following the completion of the acquisition phase. The highest acquisition criterion produced the most durable maintenance responses similar to the results demonstrated by Richling et al. (2019). However, in contrast to the findings of Richling et al. (2019), Fuller and Fienup (2019) found that a 90% performance criterion across one session was stringent enough to predict durable maintenance responses. One exclamation for the durable maintenance responses produced by the 90% criterion in Fuller and Fienup (2018) is that more instructional trials were used during the acquisition phase (20 trials) compared to 10 trials in Richling et al. (2019). Another reason for the discrepancy may be due to another instructional design component in which additional targets were taught after the initial acquisition of a set in Richling et al. (2019). Further, it is possible that

the specific prompting procedure used in combination with particular mastery criteria may produce varying results. Longino et al. (2022) demonstrated that a 90% across three sessions criterion may be sufficient when employed in combination with a most-to-least prompting hierarchy rather than the least-to-most procedure adopted by Richling et al. (2019). Unlike Richling et al. (2019), Fuller & Fienup (2018), and Longino et al. (2022), Pitts and Hoerger (2021) reported that only small decreases in maintenance we observed following the employment of an 80% or above for three sessions criterion. However, these authors opted to provide reinforcement for correct responses during maintenance probes, which was not consistent with the aforementioned studies and may have resulted in the contrasting results. These various aspects of instructional design certainly warrant further research. New single-subject research in this area has been emerging to accomplish this goal over time. To help bridge the gap, Wong, Fienup et al. (2022) conducted a systematic analysis of the use of various forms of performance criteria on maintenance and found that even as specific procedural details varied, greater maintenance was observed with higher levels of a performance criterion.

McDougale et al. (2019) conducted a descriptive analysis to compare the performance criteria utilized by practitioners (Richling et al., 2019) with the performance criteria reported in articles published by behavior analyst researchers in three major journals between 2015 and 2017. Overall, the results showed many commonalities among the type of performance criteria utilized during skill acquisition interventions across both clinicians and researchers. The results show that the most utilized type of performance criterion was the session-based percentage of correct responses. There were differences in the level of accuracy. Among researchers, a 90% accuracy criterion was more widely used, and among clinicians, an 80% accuracy criterion was most widely used. With regard to the frequency of sessions observed at the established performance criterion, researchers favored a fewer number of consecutive sessions at 90% accuracy. As mentioned above, clinicians widely adopted an 80% accuracy across three consecutive sessions as mastery. The differences in performance criteria used between researchers and clinicians may be a result of different terminal goals of the researchers and the clinicians. Clinicians may operate within the constraints of the educational goals outlined in a learner's Individualized Education Plan and may have time limits to achieve the goals. In contrast, researchers may aim to achieve a greater difference in behavior change from baseline, and thus apply a higher, more stringent criterion for skill acquisition. Researchers may also have more flexibility and fewer time constraints compared to clinicians. An alarming finding from McDougale et al. (2019) is that greater than 50% of the research articles analyzed failed to include follow-up probe sessions to assess for maintenance of the skill.

Performance Criteria with School-Aged Children

In regular education settings, one common performance criterion is the use of fluency-based measures of performance. While many of the performance criteria discussed thus far relate to accuracy, fluency adds a time component. For example, one might define math fluency in terms of the number of math problems solved correctly within a minute or reading fluency as the number of words read accurately per minute. Indeed, whole systems of allocating educational services have been built on the notion of academic fluency benchmarks serving as indicators of (1) which children would benefit from universal educational services (tier 1), (2) which children require more intensive, small group instruction (tier 2), and (3) which children require highly individualized and possibly one-on-one instruction (tier 3). Called multi-tiered systems of support (MTSS; Jimerson et al., 2016), educators use student performance data—primarily measures of fluency—to make decisions about the appropriate educational support—whether the current instruction is effective or whether teaching tactics need to change.

MTSS begins with universal academic assessments of academic fluency, or curriculum-based measurement (CBM; Jimerson et al., 2016; Cummings & Petscher, 2016). The assessments include having children read, complete math problems, and write using materials from the school district's curriculum. Educators time the assessments and then calculate fluency. For example, a teacher or school psychologist may provide first graders with grade-level appropriate reading passages and ask the child to read the text aloud. The educator times the reading and marks which words were read incorrectly and then calculate words read correctly per minute (WRCPM) based on either the first minute of reading or based on reading the whole passage. The educator can then compare one child's reading fluency to peers and district norms to decide who should continue receiving current instruction (which should be empirically supported), who needs additional help, and who needs individualized services.

In one study, Ivarie (1986) utilized fluency-based measures to teach fourth-grade students concepts of Arabic and Roman numerals. The researchers manipulated the required fluency—either 70 correct responses per minute or 35 correct responses per minute—and observed that fourth graders who were taught to a higher fluency criterion maintained the skill longer and at a higher level than those whose criterion was set lower. These outcomes suggest faster fluency is associated with better educational outcomes. Additionally, they suggest that applying a more stringent teaching criterion produces better outcomes, which is similar to those effects found with a percentage correct criterion (Fuller & Fienup, 2018; Richling et al., 2019).

Another common performance measure in regular education settings is academic achievement—or scores on standardized assessments (Jimerson et al., 2016). Academic achievement tests (e.g., Woodcock-Johnson Tests of Achievement) involve an educator or school psychologist following a manual that includes academic antecedents related to reading, writing, and math to students, measuring responses to those antecedents, and providing no performance feedback. A test involves subtests which evaluate different aspects of an academic content area. For example, reading achievement often includes tests of letter identification, letter sounds, reading fluency, and reading comprehension. Achievement tests result in standard scores based on the child's grade and age. Standard scores are set such that the 50th percentile is a score of 100. The testing developer administers the test to many thousands of students at different educational levels and across different racial and economic groups to produce norms. Then, the educator can use software to evaluate how an individual student's academic achievement compares to other children in the same grade to make decisions about the type of instruction one requires to continue making academic gains.

While academic achievement tests are commonly used in practice for diagnosing learning disabilities, the use of academic achievement assessments for ongoing performance evaluation is limited (Jimerson et al., 2016). First, the tests are not designed to be administered frequently. Second, achievement tests are a general assessment across a number of academic areas that may not map onto specific educational goals that teachers are targeting. Thus, achievement assessments are only loosely related to performance on specific academic skills and the instruction going on in one's school. For these reasons, we suggest using CBM fluency measures and district norms to assess student performance on an ongoing basis in regular education settings. For more information on academic skills, refer to Chap. 55.

Performance Criteria with College Students

A number of studies have examined how altering performance criteria with college students affects student learning, generalization, and response maintenance. One of the first studies was conducted by Johnston and O'Neill (1973). The experiment was conducted within the context of Keller's Personalized System of Instruction (PSI; Keller, 1968), which includes weekly units composed of learning materials (e.g., readings) and

terminal quizzes. In his original conception, Keller (1968) required 100% accuracy on a terminal quiz in order to move from one unit to the next. Thus, PSI is "mastery" based and progression through a PSI course requires meeting criteria during a particular unit. Johnston and O'Neill (1973) examined the effects of different performance criteria assigned to the unit quizzes. Students experienced different criterion levels (low, medium, and high, defined specifically as a rate of correct responding on unit quizzes, with a minimum rate of correct and a maximum rate of incorrect). The researcher found, not surprisingly, that student performance changed as a function of the minimum criterion. That is, when the criterion was high, students performed better than when the criterion was low, revealing a positive linear relationship between criterion and performance.

After the publication of Johnston and O'Neill's (1973), two additional studies examined criterion effects, also within a PSI context. Semb (1974) extended this area of research by examining low and high criteria for short and long assignments. In Semb's study, participants completed four units, all with quizzes, and a cumulative "review" exam that covered content across the four units. There were three experimental conditions: 100% criterion applied to each unit quiz (short assignment, high criterion), 60% criterion applied to each unit quiz (short assignment, low criterion), and 100% criterion only applied to the cumulative exam (long assignment, high criterion). Semb found that students in the short assignment, high criterion condition performed at a much higher level than peers in other conditions, suggesting the strength of breaking learning into small chunks and requiring 100% performance criteria to move from one unit to the next. This four-unit structure extended across the semester, repeating itself a few times. In this study, short assignments were individual units and mastery of each unit was required to move on to the next. There were two variations of short assignments, one which required 100% performance on each unit quiz and the review exam in order to progress through the course, and another which required 60% performance on each unit quiz and the review exam to progress. Semb also reported on response generalization and maintenance as some questions from the unit quizzes were replicated on the cumulative exams or modified. Again, participants in the short assignment, high criterion condition fared the best on generalization and maintenance questions.

Carlson and Minke (1975) further extended this area by examining different criterion levels, specifically 80% and 90% criterion levels. The authors observed that students repeatedly re-took unit quizzes following failure and this sometimes led to withdrawal from courses. Carlson and Minke compared 80% and 90% criterion levels to an ascending criterion that began with a low criterion (60%) and the criterion *ascended* every few units until the criterion was 90% near the end of the semester. Overall, the researchers found that students in the 80% criterion condition scored the highest grades in the class and passed a higher number of quizzes. Students in the 90% criterion condition did well, but less well than students in the 80% and ascending criterion conditions in terms of how many units the students completed. This study questioned the specific criterion requirements for college students completing PSI, but nonetheless demonstrated the need for relatively high-performance criteria.

More recently, this phenomenon was examined with a new type of performance: derived relations. Derived relations (see Sidman, 1994; Rehfeldt, 2011; Brodsky & Fienup, 2018), or inference making, begins with teaching overlapping conditional relations that result in multiple types of inferences, such as bi-directional relations (symmetry, if A goes with B, then B goes with A) and novel associations (equivalence, if A goes with B and A goes with C, then B and C go together). Fienup and Brodsky (2017) conducted an evaluation of this paradigm and studied how performance criteria during training affected the emergence of symmetry and equivalence relations. College students learned neuroanatomy classes that included the names of brain structures (A stimuli, e.g., Amygdala), a picture of the structure (B stimuli), a statement about the function of that structure (C stimuli), and a statement about the result of damage to that structure (D

stimuli). Teaching involved conditionally relating the A stimuli to the B, C, and D stimuli in consecutive phases. There were three performance criterion conditions. In the first condition, during each conditional relation, there were blocks of 12 trials and the criterion was 100% during a single block of trials. In the second condition, trials were repeatedly administered until a participant responded correctly to 12 consecutive trials. Both of these conditions constituted "stringent" criterion conditions. The third condition was the less stringent condition and required a participant to respond correctly to six consecutive trials. Fienup and Brodsky evaluated the performance criteria by examining tests of symmetry and equivalence and found that only stringent criteria reliably produced inferences, regardless of whether the criterion was evaluated in blocks or consecutive trials.

Collectively and across different measures, the research suggests that college students learn more and retain the information longer when high levels of performance criteria are applied to skill acquisition. This has been found across fluency (Johnston & O'Neill, 1973) and percentage correct (e.g., Semb, 1974) measures of performance. This includes a broad array of outcomes, such as initial performance (Johnston & O'Neill, 1973; Semb, 1974), generalization (Semb, 1974), response maintenance (Semb, 1974), and inferences (Fienup & Brodsky, 2017).

Performance Criteria in Sports

The evaluation of skill acquisition is fundamental in behavioral analytic research in sports performance (see Chap. 47). Evidence-based practices in behavioral sport psychology began in the late 1960s and early 1970s with the implementations of reinforcement contingencies (Rushall & Pettinger, 1969), self-monitoring tactics (Rushall & Siedentop, 1972), and behavioral assessments (McKenzie & Rushall, 1974) in sport settings. Since then, the body of research on behavioral interventions within the athletic industry remains relatively small. The results of the research that exist suggest that behavior analytic procedures are beneficial in improving performance in a variety of different sports such as football, gymnastics, tennis, figure skating, soccer, and golf (Barker et al., 2020).

The interventions used in sport-related performances rarely implement singular components. Instead, several strategies or components are typically combined into a treatment package. As the body of research continues to grow, it is important to evaluate each individual component, and the performance criterion is an important one. Martin and Thomson (2011) outline several stages of mastery based on the instructional hierarchy model within behavioral sport psychology. Under this model, an individual begins at the acquisition phase, in which the target skill is learned and performed in response to key discriminative stimuli. As soon as an individual acquires the target skill, the next stage of mastery is focused on fluency. Speed and accuracy are essential during this stage (Binder, 1996; Martin & Thomson, 2011). That is, the individual performs complex behavioral chains so accurately and fluently that an observer may characterize the performance as effortless and automatic. The acquisition and fluency of an acquired skill under practice conditions must extend to more naturalistic settings during the maintenance stage of sports mastery. Target behaviors are under different discriminative stimuli that resemble game-like conditions. This eventually extends to the generalization and adaptation of the skill, in which the individual performs the target behaviors under completely novel conditions and is capable of responding to complex and changing situations.

The complex nature of mastery in sports performance suggests the need for precise criteria to address acquisition, fluency, maintenance, and generalization. Behavioral researchers who implement interventions for enhancing sports performance typically apply performance criteria in three forms, percentage of accurate responses, number of accurate responses in succession, and rate of accurate responses. However, it is worth to note that reports of performance criteria are often missing from the published studies that were reviewed.

Level-based performance criteria combined with a particular frequency of observations component were applied and reported in a variety of different behavioral interventions including behavioral coaching packages, goal setting, oral feedback, and public posting (Brobst & Ward, 2002; Stokes et al., 2010; Tai & Miltenberger, 2017; Ward & Carnes, 2002). The instructors all established a percentage of 90% or 100% acquisition criteria. The rationale for the particular level of performance criteria that was established varied between studies. Some instructors justified their level of performance criteria to be adequate based on precedents set by existing literature on the same sport and on their personal expertise of the sport (Brobst & Ward, 2002), while other instructors allowed their participants to establish their own personal performance criteria (Ward & Carnes, 2002).

Another dimension of performance criteria utilized in behavioral sports research is the number of correct consecutive responses. An intervention package called teaching with acoustic guidance (TAGTeach) was implemented with an adult novice golfer who learned a series of target skill sets that comprise the full golf swing (Fogel et al., 2010). Each skill set consisted of small component skills. During the intervention, the introduction of each component skill was contingent on the participant's emission of six independently correct responses to the previous skill in the chain. Assessment of maintenance responding was conducted following the sixth session of the intervention. The researchers also assessed for the generalization of skills to a different golf club. Similarly, a chaining-mastery procedure was implemented with little league baseball players (Simek & O'Brien, 1988). Each task of the chain had a predetermined criterion of a number of consecutive correct responses or a certain number of correct responses out of the total number of opportunities given.

Fluency criteria have also been applied to interventions within behavioral sports research. Pocock et al. (2010) targeted two roller skating skills by implementing a precision teaching methodology (Lindsley, 1971). Because precision teaching emphasizes fluent behavior, the researchers applied a criterion that targeted the rate of responding. The criterion was established based on the behaviors of a model exemplar who was not included in the study.

A limitation of performance criteria in behavioral sports research is that some movements are fluid and require precise body movements and positioning (e.g., gymnastics). A standard criterion that is typically used in say, academics may not be as viable with sports because near flawless performance (90–100% accuracy) may be difficult to achieve for even the most elite athletes. Establishing a performance criterion of 90% or 100% accuracy may also be problematic because participants have reported feeling emotionally distressed when their performance criteria were not achieved (Brobst & Ward, 2002). It is important to consider alternative means of signaling the termination of intervention, including the establishment of more modest levels of performance criteria dependent on the participant's skill levels or criterion that is based on a percentage of improvement from previous performances.

Performance Criteria in Organizational Behavior Management (OBM)

Organizational behavior management (OBM) is an approach that applies behavioral principles to increase the effectiveness and efficiency of workers in organizational settings within a wide range of disciplines such as government, industry, business, and human service. There is an emphasis on the implementation of practical interventions to change behavior. Like treatments in ABA research, OBM interventions have predetermined performance criteria to signal the termination of an experimental condition. Some widely used strategies in the human service sector include checklists, providing feedback, trainings or workshops, applying self-monitoring techniques, goal setting, and rewards (VanStelle et al., 2012). These strategies have the aim of improving the accuracy of treatment implementation (procedural integrity) and staff performance.

Many OBM studies related to human service published between 2010 and 2016 in JABA and BAP utilize a percentage of correct responses to signal the termination of an intervention or treatment (Gravina et al., 2018). The following studies applied an 80%, 90%, or 100% accuracy criterion across single or multiple sessions.

Casey and McWilliam (2011) implemented a checklist-based training procedure to help teachers and staff decrease student transition times within a classroom setting. The training was stopped if the staff members performed at least 80% of the checklist task most of the time for three consecutive sessions. In this study, the experimenters also conducted maintenance probes following the end of the training. Ditzian et al. (2015) also applied an 80% accuracy criterion for their feedback-based intervention to improve proper door closing of therapy rooms. The experimenters determined that 80% accuracy across two consecutive sessions was appropriate to stop the intervention. Graff and Karsten (2012) implemented an instructional package that included enhanced written instructions and written instructions with data sheet to increase the accuracy implementation of stimulus preference assessments for simulated consumers. The performance criterion was 90% accuracy across two consecutive sessions. The experimenters also conducted generalization probes with real consumers. Lambert et al. (2013) trained staff at a community residential facility to conduct trial-based functional analyses. In order for the training to conclude, the staff members were required to implement all trial types with 100% accuracy. Nabeyama and Sturmey (2010) also applied a 100% correct response to all target actions during a behavioral skills training program for staff. Additionally, the experimenters established an additional criterion for intervention enhancement if the staff members that included increased opportunities for training and rehearsing. The instructor also included modeling correct responses. These extra components were implemented if the staff members performed less than six target components correctly within the first two sessions of the intervention.

For interventions that target treatment fidelity, the research in OBM shows that a criterion of 80% or more is necessary before the training should be concluded. However, it is important to evaluate which percentage level is appropriate for different types of target skills. For example, interventions that target client safety should have a performance criterion of no less than 100% accuracy because client well-being and safety are at stake. In addition to the utilization of a percentage of accuracy for performance criterion, the number of sessions where performance is observed at a certain level is another aspect of the criterion. Typically, the number of observations range from one session to three consecutive sessions with an accurate performance at a certain percentage. Studies in OBM research also assess for response maintenance or generalization of the target skill across settings or people (Casey & McWilliam, 2011; Graff & Karsten, 2012; Nabeyama & Sturmey, 2010; Nigro-Bruzzi & Sturmey, 2010; Parsons et al., 2012). Response maintenance and generalization are crucial in the discussion of mastery.

A Model for Establishing Performance Criteria

A model for selecting performance criteria requires nuance and consideration of the learning context, educational goals, and type of learner. In some cases, the literature supports a specific model and in other cases, additional research is needed before clear, research-based suggestions can be made. When working with school-aged children in regular education settings, performance can be assessed using frequent curriculum-based assessments of fluency (e.g., reading, math, writing). Comparing an individual child's fluency to district or national norms should indicate to the teacher the child's current proficiency given the instruction and instructional modifications can be made as necessary (Cummings & Petscher, 2016). With college students, there is compelling evidence that performance criteria drive performance and, thus, performance criteria should be set as high as experts believe is necessary. In any given

college class, this could involve setting performance criteria at a value such as 80% accuracy or 100% accuracy and providing additional instruction until the performance criterion is met.

Skill Acquisition for Learners with Disabilities

Much of ABA is conducted with young children with disabilities who require intensive, deliberate, and individualized instruction. At this point, there is too little research to suggest specific performance criteria guidelines with the exception that acquisition criteria should be set high (e.g., a high level of accuracy). Beyond establishing high levels of performance, the specific component of performance criteria should be individualized, just as the specific learning objectives are individualized. Indeed, this same approach has been argued for selecting other instructional components such as prompts (Seaver & Bourret, 2014; Cengher et al., 2016; Cengher et al., 2018; Schnell et al., 2020) and error correction procedures (Carroll et al., 2018). The impetus for individualized assessments is the fact that when comparing different instructional components, it is often the case the effects are idiosyncratic: no specific component that works best for all learners, but often there are components that are reliably better for a single learner. While the performance criterion data published thus far show consistency in terms of the need for high acquisition criteria (e.g., Richling et al., 2019), the research has been conducted with a narrow range of children and response types (e.g., tacts and sight words). A deeper understanding of learner characteristics and response characteristics will undoubtedly bring nuance to our understanding of how to tailor performance criteria. While we await such data, a framework for assessing learner-specific criteria is useful.

Determining Goals of Instruction

Prior to directly comparing performance criteria, one should begin by asking questions to help guide their own analysis. The first question is, "What are the goals of teaching?" Answers range from generalized responses that are not affected by context, teacher, or specific stimuli to durable responses that maintain over time after instruction has ceased. Answering this question sets up one's dependent variables. For example, if one is interested in promoting durable responses that maintain for at least a month, the appropriate dependent variable—or performance outcome variable from an analysis—to study. If one is interested in both response maintenance and generalization, then one should measure both.

Level of Performance for Specific Behaviors

After determining the desired effect of one's teaching, one should ask "What are acceptable levels of behavior?" There is not necessarily an agreed upon standard for what is an acceptable level of behavior and consensus may vary as a function of the skill being taught. For example, when teaching a child to tact colors, 100% performance may not be necessary; however, if teaching a child to look both ways before crossing a street, any performance level below 100% may be wholly unacceptable due to dangerous outcomes. Another manner of developing appropriate performance criteria is through social validation (Van Houten, 1979). In this approach, intervention targets are developed based on normative sample data or by observing competent individuals. For example, if one is trying to teach toy play to infants at risk for delayed motor development, social validation may involve defining what constitutes toy play for an infant and collecting data on the duration of toy play with several typically developing infants to derive an intervention goal for the infant you are working with. Social validation focuses on appropriate levels of behavior, rather than trying to fit behavior on a scale of percentage correct out of 10 opportunities.

An Experimental Approach to Establishing Individualized Performance Criteria (Approach 1 of 2)

Once a clinician has determined the goal and acceptable performance, she has a refined dependent variable—or outcomes variable. This is now

the benchmark by which to compare the effects of different performance criteria. All that is left to do are some minor experimental preparations followed by a comparison of teaching responses to predetermined performance criteria and examining which produces the intended outcomes. To experimentally establish performance criteria, one must teach independent, but equally different targets. Cariveau et al. (2020) provide guidance on this process, but the basics are controlling for effort and difficulty. For instance, if your client is learning sight words, one would select two sets of sight words that are from the same grade level and have the same number of syllables and letters. By equating targets, one is in a better position to attribute the effects of the performance criteria to the criteria you implemented, and not that the sets of stimuli are simply more or less difficult. In the same vein, one should teach using the same procedures (e.g., using or not using error correction, prompt fading, etc.), regardless of performance criterion. After one sets up two or more conditions that should produce the same learning, they can assign one performance criterion to one set of stimuli and one to another set of stimuli and begin teaching. After the client's behavior meets the acquisition performance criterion, the therapist now tests for other relevant performances, such as response generalization and response maintenance. Performance of the behavior under different conditions (generalization) and after teaching has been terminated (maintenance) now serves as the indicator of which acquisition performance criterion produces the intended effects. If one acquisition performance criterion produces the intended effects, but the other does not, then the answer of which is more appropriate is clear. In the case where both performance criteria produce the intended effect, then the therapist should look back at the acquisition data and if one condition produced the quicker acquisition, then that should be the performance criterion moving forward. If both performance criteria fail to produce the intended generalization and maintenance effects, then the therapist should look to strengthen the performance criteria (higher level of performance, higher frequency component, across more instructors) or examine whether there are more effective teaching tactics.

A Naturalistic Approach to Establishing Performance Criteria (Approach 2 of 2)

Some therapists may not have the resources to conduct individualized evaluations. In this case, a more naturalistic evaluation is appropriate, although this comes with less confidence in the outcomes. In this case, the therapist should establish an acquisition criterion that appears reasonable based on the goals of instruction (e.g., 90–100% accuracy across two consecutive observations). Next, the therapist should establish acceptable generalization and maintenance performance criteria. From this point, the therapist simply teaches new behavior as she normally does until the acquisition criterion is met and then tests to see if the generalization and maintenance criteria are also met. If the generalization and maintenance criteria are met, this provides preliminary evidence that the acquisition criterion is sufficient to produce all of the intended effects of instruction. If the criteria are not met; however, the therapist should change the acquisition criterion in specific ways to produce better outcomes. For example, if the maintenance criterion is not met, consider a higher level of performance for the acquisition criterion and consider applying the criterion across a greater number of sessions or across multiple days (e.g., first-session of the day). If the generalization criterion is not met, consider adding a component that the acquisition criterion must be met across two or more instructors or two or more sets of stimuli (see Chap. 15).

Future Directions and Concluding Remarks

This chapter outlined some of the historical and contemporary treatments of performance criteria, highlighted related terminological issues, outlined a variety of areas of research utilizing performance criteria, and provided a potential model for selecting performance criteria for individual clinical use. As mentioned in the introduction of

this chapter, many of the recommendations made here are speculative and constitute best practices based on scientific deduction and clinical recommendations. However, there is a need for further research evaluating performance criteria as independent variables which function in coordination with other training procedures and may directly impact response maintenance and other learning outcomes. Without a solid evidence base from which we can derive distinct rules regarding which performance criteria to use universally, we must be mindful to not fall victim to engaging in clinical lore practices. Instead, we can mitigate some of this risk by intentionally engaging in critical consideration of performance criteria on a case-by-case basis. In addition, we can supplement our confidence by engaging in individual assessment of the impact of specific performance criteria and directly measure-related learning outcomes using rigorous single-case designs (see Chap. 20) rather than adopting train-and-hope strategies.

References

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

Barker, J. B., Slater, M. J., Pugh, G., Mellalieu, S. D., McCarthy, P. J., Jones, M. V., & Moran, A. (2020). The effectiveness of psychological skills training and behavioral interventions in sport using single-case designs: A meta regression analysis of the peer-reviewed studies. *Psychology of Sport and Exercise, 51*, 101746. https://doi.org/10.1016/j.psychsport.2020.101746

Binder, C. (1996). Behavioral fluency: Evolution of a new paradigm. *The Behavior Analyst, 19*(2), 163–197. https://doi.org/10.1007/BF03393163

Brobst, B., & Ward, P. (2002). Effects of public posting, goal setting, and oral feedback on the skills of female soccer players. *Journal of Applied Behavior Analysis, 35*(3), 247–257. https://doi.org/10.1901/jaba.2002.35-247

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., Batchelder, S., Ball, S., & La Cruz Montilla, A. (2020). *Review of methods to equate target sets in the adapted alternating treatments design*. Advanced online publication. Behavior Modification. https://doi.org/10.1177/0145445520903049

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/https://doi.org/10.1037/h0078676.

Carroll, R. A., Owsiany, J., & Cheatham, J. M. (2018). Using an abbreviated assessment to identify effective error-correction procedures for individual learners during discrete-trial instruction. *Journal of Applied Behavior Analysis, 51*(3), 482–501. https://doi.org/10.1002/jaba.460

Casey, A. M., & McWilliam, R. A. (2011). The impact of checklist-based training on teachers' use of the zone defense schedule. *Journal of Applied Behavior Analysis, 44*(2), 397–401. https://doi.org/10.1901/jaba.2011.44-397

Cengher, M., Shamoun, K., Moss, P., Roll, D., Feliciano, G., & Fienup, D. M. (2016). The effects of two prompt-fading strategies on skill acquisition in children with autism spectrum disorder. *Behavior Analysis in Practice, 9*(2), 115–125. https://doi.org/10.1007/s40617-015-0096-6

Cengher, M., Budd, A., Farrell, 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*(5), 155–173. https://doi.org/10.1007/s10882-017-9575-8

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Pearson.

Cummings, K. D., & Petscher, Y. (2016). *The fluency construct*. Springer.

Ditzian, K., King, A., Tanz, J., & Wilder, D. (2015). N evaluation of the performance diagnostic checklist-human services to assess an employee performance problem in a center-based autism treatment facility. *Journal of Applied Behavior Analysis, 48*(1), 199–203. https://doi.org/10.1002/jaba.171

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

Fogel, V. A., Weil, T. M., & Burris, H. (2010). Evaluating the efficacy of TagTeach as a training strategy for teaching a golf swing. *Journal of Behavioral Health and Medicine, 1*(1), 25–41. https://doi.org/10.1037/h0100539

Freeland, J. T., & Noell, G. H. (2002). Programming for maintenance: An investigation of delayed intermittent reinforcement and common stimuli to create indiscriminable contingencies. *Journal of Behavioral Education, 11*(1), 5–18. https://doi.org/10.1023/A:1014329104102

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

Graff, R. B., & Karsten, A. M. (2012). Evaluation of a self-instruction package for conducting stimulus

preference assessments. *Journal of Applied Behavior Analysis, 45*(1), 69–82. https://doi.org/10.1901/jaba.2012.45-69

Gravina, N., Villacorta, J., Albert, K., Clark, R., Curry, S., & Wilder, D. (2018). A literature review of organizational behavior management interventions in human service settings from 1990 to 2016. *Journal of Organizational Behavior Management, 38*(2–3), 191–224. https://doi.org/10.1080/01608061.2018.1454872

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

Haughton, E. C. (1984). Standards: Refining measurement. *Journal of Precision Teaching, 4*(4), 96–99.

Ivarie, J. J. (1986). Effects of proficiency rates on later performance of a recall and writing behavior. *Remedial and Special Education, 7*(5), 25–30. https://doi.org/10.1177/074193258600700506

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 ed.). 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

Kubina, R. M., & Starlin, C. M. (2003). Reading with precision. *European Journal of Applied Behavior Analysis, 4*(1–2), 13–21. https://doi.org/10.1080/15021149.2003.11434212

Lambert, J. M., Blooms, S. E., Kunnavantana, S. S., Collins, S. D., & Clay, C. J. (2013). Training residential staff to conduct trial-based functional analyses. *Journal of Applied Behavior Analysis, 46*(1), 296–300. https://doi.org/10.1002/jaba.17

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

Longino, E., Richling, S. M., McDougale, C. B., & Palmier, J. M. (2022). The effects of mastery criteria on maintenance: A replication with most-to-least prompting. *Behavior Analysis in Practice, 15*(2), 397–405. https://doi.org/10.1007/s40617-021-00562-y

Lovaas, O. I. (1981). *Teaching developmentally disabled children: The me book*. University Park Press.

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://doi.org/10.1016/j.rasd.2008.08.008

Martin, G., & Pear, J. (2007). *Behavior modification: What it is and how to do it* (8th ed.). Pearson.

Martin, G. L., & Thomson, K. (2011). Overview of behavioral sport psychology. In J. K. Luiselli & D. D. Reed (Eds.), *Behavioral sport psychology: Evidence-based approaches to performance enhancement* (pp. 3–21). Springer.

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

McKenzie, T. L., & Rushall, B. S. (1974). Effects of self-recording on attendance and performance in a competitive swimming training environment. *Journal of Applied Behavior Analysis, 7*(2), 199–206. https://doi.org/10.1901/jaba.1974.7-199

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

Nabeyama, B., & Sturmey, P. (2010). Using behavioral skills training to promote safe and correct staff guarding and ambulation distance of students with multiple physical disabilities. *Journal of Applied Behavior Analysis, 43*(2), 341–345. https://doi.org/10.1901/jaba.2010.43-341

Nigro-Bruzzi, D., & Sturmey, P. (2010). The effects of behavioral skills training on mand training by staff and unprompted vocal mands by children. *Journal of Applied Behavior Analysis, 43*(4), 757–761. https://doi.org/10.1901/jaba.2010.43-757

Parsons, M. B., Rollyson, J. H., & Reid, D. H. (2012). Evidence-based staff training: A guide for practitioners. *Behavior Analysis and Practice, 5*(2), 2–11. https://doi.org/10.1007/BF03391819

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

Pocock, T. L., Foster, T. M., & McEwan, J. S. (2010). Precision teaching and fluency: The effects of charting and goal-setting on skaters' performance. *Journal of Behavioral Health and Medicine, 1*(2), 93. https://doi.org/10.1037/h0100544

Rehfeldt, R. A. (2011). Toward a technology of derived stimulus relations: An analysis of articles published in the *Journal of Applied Behavior Analysis*, 1992-2009. *Journal of Applied Behavior Analysis, 44*(1), 109–119. https://doi.org/10.1901/jaba.2011.44-109

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

Rushall, B. S., & Pettinger, J. (1969). An evaluation of the effects of various reinforcers used as, motivators in swimming. *Research Quarterly, 40*(3), 540–545. https://doi.org/10.1080/10671188.1969.10614875

Rushall, B. S., & Siedentop, D. (1972). *The development and control of behavior in sport and physical education*. Lea & Febiger.

Sayrs, D. M., & Ghezzi, P. M. (1997). The steady-state strategy in applied behavior analysis. *The Experimental Analysis of Human Behavior Bulletin, 15*(2), 29–30.

Schnell, L. K., Vladescu, J. C., Kisamore, A. N., DeBar, R. M., Kahng, S., & Marano, K. (2020). Assessment to identify learner-specific prompt and prompt-fading procedures for children with autism spectrum disorders. *Journal of Applied Behavior Analysis, 53*(2), 1111–1129. https://doi.org/10.1002/jaba.623

Seaver, J. L., & Bourret, J. C. (2014). An evaluation of response prompts for teaching behavior chains. *Journal of Applied Behavior Analysis, 47*(4), 777–792. https://doi.org/10.1002/jaba.159

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

Sidman, M. (1994). *Equivalence relations and behavior: A research story*. Authors Cooperative.

Simek, T. C., & O'Brien, R. M. (1988). A chaining-mastery, discrimination training program to teach Little Leaguers to hit a baseball. *Human Performance, 1*(1), 73–84. https://doi.org/10.1207/s15327043hup0101_4

Skinner, B. F. (1968). *The technology of teaching*. Appleton-Century-Crofts.

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

Stokes, J. V., Luiselli, J. K., Reed, D. D., & Fleming, R. K. (2010). Behavioral coaching to improve offensive line pass blocking skills of high school football athletes. *Journal of Applied Behavior Analysis, 43*(3), 463–472. https://doi.org/10.1901/jaba.2010.43-463

Tai, S. S. M., & Miltenberger, R. G. (2017). Evaluating behavioral skills training to teach safe tackling skills to youth football players. *Journal of Applied Behavior Analysis, 50*(4), 849–855. https://doi.org/10.1002/jaba.412

Van Houten, R. (1979). Social validation: The evolution of standards of competency for target behaviors. *Journal of Applied Behavior Analysis, 12*(4), 581–591. https://doi.org/10.1901/jaba.1979.12-581

VanStelle, S. E., Vicars, S. M., Harr, V., Miguel, C. F., Koerber, J. L., Kazbour, R., & Austin, J. (2012). An objective review and analysis: 1998-2009. *Journal of Organizational Behavior Management, 32*(2), 93–123. https://doi.org/10.1080/01608061.2012.675864

Ward, P., & Carnes, M. (2002). Effects of posting self-set goals on collegiate football players' skill execution during practice and games. *Journal of Applied Behavior Analysis, 35*(1), 1–12. https://doi.org/10.1901/jaba.2002.35-1

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

Wong, K. K., Fienup, D. M., Richling, S. M., Keen, A., & Mackay, K. (2022). Systematic review of acquisition mastery criteria and statistical analysis of associations with response maintenance and generalization. *Behavioral Interventions, 37*(4), 993–1012. https://doi.org/10.1002/bin.1885