Received: 3 June 2021 | Revised: 3 June 2021 | Accepted: 27 August 2021

DOI: 10.1002/bin.1827

REVIEW ARTICLE

WILEY

The use of performance criteria for determining "mastery" in discrete-trial instruction: A call for research

Daniel M. Fienup[1] | **James E. Carr**[2]

[1]Department of Health and Behavior Studies, Teachers College, Columbia University, New York, New York, USA

[2]Behavior Analyst Certification Board, Littleton, Colorado, USA

Correspondence
Daniel M. Fienup, Department of Health and Behavior Studies, Teachers College, Columbia University, 525 W 120th St., Box 223, New York, NY 10027, USA.
Email: fienup@tc.columbia.edu

Abstract

The *performance criterion* (also known as the mastery criterion) is the operationalized performance standard that behavior must reach before instruction ceases, changes, or becomes focused on other goals, such as stimulus generalization and response maintenance. Although performance criteria are widely used in skill-acquisition research and practice, there has been little experimental research on the topic. Thus, we provide a review of relevant research and offer suggestions for future research based on the different performance-criterion components.

KEYWORDS
discrete-trial instruction, generalization, maintenance, mastery criterion, performance criterion

1 | INTRODUCTION

Discrete-trial instruction (DTI) is a widespread teaching procedure, particularly for children with autism spectrum disorder (ASD). Although DTI has been used educationally and therapeutically for decades, it has received substantial research attention in recent years. For example, researchers have evaluated DTI in terms of how trials from multiple teaching programs are distributed within a session (e.g., Majdalany et al., 2014), how implementation errors affect learning outcomes (e.g., DiGennaro Reed et al., 2011), and how various data collection methods impact learning (e.g., Cummings & Carr, 2009), among other topics. One variable that has been curiously absent from the research literature is the *performance criterion* (also known as the mastery criterion[1]), the operationalized standard that behavior must reach before an instructor ceases or changes instruction, or focuses on other goals beyond acquisition such as stimulus generalization and response maintenance. This is a perplexing situation because performance criteria are almost universally included in applications of DTI (e.g., McDougale et al., 2020), but have only recently been the subject

of experimental investigations. Thus, the purpose of this article is to review recent research on performance criteria in DTI and offer suggestions for future research.

The performance criterion, a critical aspect of a behavioral objective (Mayer et al., 2012), is composed of three components (Richling et al., in press; see Figure 1). The first is the *criterion level*, which is the quantitative value that performance must meet. The criterion level is often expressed as a percentage-correct metric for session-based data or trial blocks (e.g., 90% correct; see McDougale et al., 2020 for a review) and could include other dimensions of behavior such as a fluency aim. The second element is the *criterion-level frequency value*, the number of consecutive observations during which the criterion level must be met (e.g., three consecutive sessions at or above 90% correct).[2] These first two components are generally designed to promote response maintenance. The third component is the use of *supplementary variables*, additional stimuli in whose presence the first two components must be met, such as in the presence of multiple therapists or across multiple days. Supplementary variables often appear to be designed to enhance stimulus generalization. Of course, performance criteria can be established in myriad ways, but the aforementioned description characterizes the general structure of most performance criteria used in DTI arrangements (Richling et al., 2019).

Various performance standards have been used in behavior-analytic research and practice over the decades. In the experimental analysis of behavior, steady-state criteria have been used to specify a limited range of behavior before changing experimental conditions (Rehfeldt & Ghezzi, 1996; Sidman, 1960). However, quantifying a steady state is perhaps best suited for free-operant behavior. By contrast, many teaching preparations involve restricted, or discrete, operants. In an early applied example, Keller's (1968) personalized system of instruction (PSI) included a series of curricular units, each of which was composed of assignments, readings, and a terminal "readiness" test or unit quiz. Keller implemented a unit-perfection requirement that required a student to score 100% during one readiness test to progress to the next unit. In a later example, Lovaas (1981) characterized the performance criterion in DTI by stating that an instructor "should move to the next step when the child can respond correctly *without prompting* on several consecutive trials. The child *responds to criterion* when he responds correctly on 9 out of 10, or 18 out of 20, consecutive trials" (p. 63). Both of these early examples of performance criteria included clear statements of the criterion level (i.e., 100% correct, 9 out of 10 correct trials). Over the subsequent decades, behavior analysts began using more elaborate performance criteria, which have become rather commonplace in both research and practice (McDougale et al., 2020).

2 | RECENT RESEARCH ON PERFORMANCE CRITERIA

Over the past few years, several researchers have begun evaluating the use of performance criteria by practitioners and researchers, and experimentally evaluating performance criteria as independent variables for learning outcomes.

2.1 | Descriptive assessment of performance-criterion practices

Richling et al. (2019) conducted a survey of behavior analysts responsible for the delivery of treatment to individuals with developmental disabilities; 199 individuals responded to the survey. The majority of respondents (68%) reported using session-based performance criteria (i.e., percentage of correct trials), 28% reported using criteria based on correct consecutive trials, and 4% reported using performance criteria based on response rates. For respondents who reported using session-based performance criteria, 54% reported using an 80%-correct criterion level, 28% reported using a 90%-correct criterion level, and only 7% reported using a 100%-correct criterion level. For these same respondents, 8% reported requiring the criterion level to be met in a single session, 35% across multiple sessions, and 57% across multiple sessions along with other supplementary variables (e.g., in the presence of two or more therapists). Thus, the most common session-based performance criterion was an 80%-correct criterion level with a

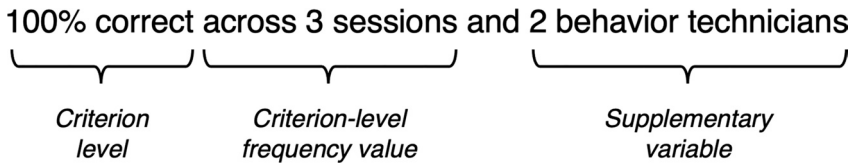

FIGURE 1 An illustration of the three components of the performance criterion

criterion-level frequency of 3 in the presence of other supplementary variables, which is consistent with a practitioner survey conducted a decade earlier (Love et al., 2009). When asked to report the origin of their performance-criterion practices, 44% of respondents attributed them to prior supervised experience, 28% to employer requirements, 16% to graduate training, and 10% to workshops. An interesting omission in these reports is reference to the empirical literature.

McDougale et al. (2020) expanded on Richling et al.'s (2019) survey by examining performance criteria reported in experimental literature that focused on skill acquisition.[3] The authors identified 157 articles in which performance criteria were used during skill acquisition. The majority of articles (54%) included performance criteria based on a percentage of correct trials. Sixty-one percent of these articles reported criterion levels between 90% and 100% correct, and 38% reported criterion levels between 80% and 89% correct. These findings suggest more stringent criterion levels are being used in research compared to those reported by practitioners (Richling et al., 2019). The most commonly reported criterion-level frequency value identified in the McDougale et al. review was the criterion level being met across two sessions (60% of articles), followed by one session (21%) and three sessions (21%). Finally, of the 157 articles (which included both free- and restricted-operant preparations), 59% did not require the criterion level to be met across supplementary variables.

The Richling et al. (2019) practitioner survey and the McDougale et al. (2020) literature review revealed that performance criteria are widely used by practitioners and researchers in skill-acquisition programs. However, these articles also identified substantial variability in performance-criterion practices. This is not surprising given the numerous possible permutations between criterion level, criterion-level frequency values, and supplementary variables. Unfortunately, the ubiquity of the performance criterion in research and practice vastly exceeds experimental literature on the topic, in which only a few relevant investigations have been published.

2.2 | The performance-criterion treated as a moderator variable in research

It is important to articulate the role of the performance criterion in a functional relation. Performance criteria are applied to a dependent measure (e.g., percentage accuracy of a new skill), possibly giving the impression that performance criteria affect the dependent variable (i.e., the skill being taught). In fact, the performance criterion actually serves to moderate the relation between an independent variable (e.g., the teaching procedure) and socially significant dependent variables (e.g., response maintenance; Fuller & Fienup, 2018). In research contexts, researchers have manipulated performance criteria as an independent variable while holding the intervention constant (e.g., Richling et al., 2019); however, the performance criterion is really a component of a skill acquisition teaching package and the performance criterion alone does not affect behavior change. The teaching procedures (antecedent and consequence manipulations) affect behavior and the research to date (reviewed below) suggests that continuing the teaching procedures until an individual's performance meets a stringent performance criterion positively affects the relationship between the intervention and outcomes. The research that has identified performance criteria as a moderator of DTI effects has been conducted in such a way that the researchers treat performance criteria as an independent variable. This section describes research on various performance-criterion iterations and their corresponding moderating effects on response maintenance.

Researchers began evaluating the effects of different performance criteria on subsequent response outcomes (e.g., maintenance) almost a half century ago in the context of refining Keller's (1968) PSI method with college students. In this instructional approach, students take periodic unit quizzes on which they must meet a certain performance criterion before progressing to the subsequent unit. Johnston and O'Neill (1973) and Semb (1974) both found that end-of-course performance was generally correlated with criterion levels on unit quizzes. Semb also found that higher criterion levels produced higher levels of maintenance and generalization on cumulative exams. Subsequently, Carlson and Minke (1975) and Reiser et al. (1986; 1987) demonstrated the value of an 80% criterion level. Collectively, these studies demonstrated that various performance-criterion levels lead to different outcomes, ascribing an independent variable function to the performance criterion. Indeed, the post-acquisition performance of a response should be logically related to the metric used to determine when it has been acquired. After this early research on performance criteria in PSI, decades passed before the topic was again addressed in the experimental literature.

Four recent articles have reported experimental evaluations of different performance-criterion levels on outcomes within DTI preparations for individuals with developmental disabilities.[4] Fuller and Fienup (2018) evaluated multiple criterion levels and their effects on the maintenance of the spelling or sight-word performance of three children with ASD. The researchers taught skill sets until they reached 50%, 80%, or 90% correct for a single session. The 90%-criterion level was associated with approximately the same level of skill accuracy 3 to 4 weeks after instruction ceased. The 50% and 80% criterion levels were associated with highly variable and less accurate maintenance.

Richling et al. (2019) conducted three studies similar to Fuller and Fienup (2018). In the first two studies (Experiments 2 and 3 in the published article), four children with developmental disabilities were taught auditory-visual conditional discriminations and tacts under various performance-criterion levels: 60%, 80%, and 100% correct (across three consecutive sessions). Across four weekly follow-up assessments, the 100%-correct criterion resulted in the highest maintenance (approximately 80% correct or higher across participants). The 60%- and 80%-correct criterion levels were associated with less reliable and less accurate outcomes. In the third study (Experiment 4 in the published article), the same participants were taught tacts under 80%-, 90%-, and 100%-correct criterion levels across three consecutive sessions; additional minor procedural variations were also introduced. At a single one-week follow-up probe, skills taught under the 100%-correct criterion level maintained substantially higher than skills taught under the 90% or 80% criteria.

Two recent studies systematically replicated Richling et al.'s (2019) studies in two important ways. Pitts and Hoerger (2021) evaluated the 80%-, 90%-, and 100%-correct criterion levels (across three consecutive sessions) with a new behavior (sight-word recognition) for four participants with ASD. The authors found the highest maintenance in weekly follow-up probes in the 100%- and 90%-criterion conditions. Longino et al. (2021) evaluated the 80%-, 90%-, and 100%-correct criterion levels (across three consecutive sessions) with a new teaching tactic (most-to-least prompting) for teaching tacts to three children with ASD. The authors found the highest maintenance in weekly follow-up probes in the 100% and 90% conditions compared to the 80% condition.

Taken together, the Fuller and Fienup (2018), Longino et al. (2021), Pitts and Hoerger (2021), and Richling et al. (2019) studies demonstrated roughly parametric effects, in which skills taught under the highest performance criterion level maintained better than skills taught under lower criterion levels. However, within-participant parametric effects were not always demonstrated. That is, the second and third lowest criterion levels were not always associated with commensurate decrements in response maintenance. In addition, maintenance outcomes under a particular criterion level was not similarly effective across participants. For example, Fuller and Fienup (2018), Longino et al. (2021), and Pitts and Hoerger (2021) showed fairly high maintenance under a 90%-criterion level, whereas Richling et al. (Experiment 4, 2019) did not. Indeed, the latter study demonstrated reliably lower maintenance in the 90%-criterion-level condition. One interpretation of these discrepancies is that the most appropriate performance-criterion level might be idiosyncratic for each learner. Another interpretation is that the many variables that comprise performance criteria interact in different ways to achieve a particular outcome. For example, Fuller and Fienup, Longino et al., and Richling et al. taught using error correction, most-to-least prompting, and

least-to-most prompting, respectively. Differences in these variables across studies might be related to different outcomes.

2.3 | A call for additional research

It is clear that performance criteria play an important role in skill-acquisition procedures. In a sense, the performance criterion is the *initial* goal of skill acquisition. It is therefore surprising that such little attention has been paid to this topic in the experimental literature, especially given its prevalence in practice and in skill-acquisition research. As the Richling et al. (2019) survey identified, practitioners predominantly derive their performance criteria from a variety of sources other than the experimental literature, perhaps due to the lack of relevant studies. At this time, it seems appropriate to consider the elements of performance criteria as lore, at least in DTI procedures. Additional research is needed to elevate the status of the performance criterion from lore to evidence based. The remainder of this article will briefly describe a number of research topics worth evaluating in this area. The focus of the following section is on variables that can be manipulated, but it is worth noting that is important to assess the effects of these manipulations on socially significant educational outcomes such as response maintenance and generalization.

2.3.1 | Primary variables

An obvious starting point for research on performance criteria is the evaluation of their core elements: criterion level, criterion-level frequency values, and supplementary variables. Fuller and Fienup (2018), Longino et al. (2021), Pitts and Hoerger (2021), and Richling et al. (2019) represent a starting point in the evaluation of criterion levels. Because a criterion level cannot be evaluated without a criterion-level frequency value, and vice versa, we suggest that one of these variables be held constant when evaluating the other. For example, one could evaluate the effects of 80%, 90%, and 100% accuracy (criterion levels) across three consecutive sessions (the criterion-level frequency value), with the latter variable being held constant across conditions, similar to existing studies. Conversely, one could evaluate various criterion-level frequency values (e.g., one, two, or three consecutive sessions) by holding the criterion level (e.g., 100% accuracy) constant. Schneider et al. (in press) conducted a preliminary analysis of 90% accuracy across one or three sessions. Across three participants and five comparisons, the researchers found that maintenance was unaffected by the manipulation; however, children completed teaching phases in much less time when the criterion-level frequency was set to one as opposed to three consecutive sessions. In addition, this line of research is amenable to evaluating interaction effects. For example, one could evaluate whether (a) a lower criterion level with a higher criterion-level frequency value produces comparable maintenance to (b) a higher criterion level with a lower criterion-level frequency value. Isolating the effects of criterion levels and criterion-level frequencies should ultimately—in a systematic replication line—be studied across a variety of participants, target responses, and teaching procedures to evaluate the boundaries of effects.

We recommend that after researchers better understand criterion-level and criterion-level frequency effects, they begin to evaluate the effects of the numerous supplementary variables on their targeted outcomes. This would ideally involve the criterion level and criterion-level frequency value being held constant while different supplementary variables are evaluated (e.g., across two therapists vs. three therapists, across settings, across instructional materials). In addition, the potential value of a particular supplementary variable needs to be specified and its effects assessed. For example, the value of incorporating the demonstration of a newly acquired skill in the presence of multiple therapists would be to enhance stimulus generalization. Thus, studies on this topic should include, at a minimum, baseline and posttreatment assessments of stimulus generalization, with possible assessments during teaching.

It should be evident by now that given the numerous iterations of each aspect of the performance criterion and the possibility of their interaction, research on the topic should proceed in an orderly manner. Otherwise, the complexity of the resulting permutations could hinder confident conclusions about collective research findings.

2.3.2 | Secondary variables

In addition to the primary variables described above, there are a number of secondary variables that also warrant investigation. First, with session-based instructional arrangements, the number of trials in the session could be related to performance-criterion effects due to the formula used to calculate accuracy. For example, a criterion level of 80%-correct in a session of five trials compared to a session of 10 trials results in a different number of errors being allowed in determining whether instruction should cease. The number of permissible errors could theoretically affect maintenance. With a performance criterion of 100% accuracy across multiple sessions, the number of trials per session affects the total number of consecutive correct responses required, which could also affect response maintenance. Thus, researchers should systematically investigate the effects of the performance criterion while holding the number of targets constant and varying the number of trials per session to investigate the interaction.

Second, it is becoming increasingly common for skill-acquisition procedures to focus on teaching skill sets (i.e., multiple targets) rather than one skill at a time (e.g., Kodak et al., 2020). With a skill set, one could theoretically apply the performance criterion to the entire set or to each individual target skill within it, perhaps with target replacement within the set after it individually met the performance criterion. Although the former would be a more efficient approach, a criterion level of less than 100% correct could result in one or more skills within the set being deemed "acquired" prematurely. That is, high performance of the other skills would contribute toward the criterion level being met even if one was still somewhat inaccurate. In two recent studies, Wong and Fienup (in press) and Wong et al. (2021) demonstrated this effect in the context of sight-word instruction for seven children with ASD. Acquisition was faster and maintenance was similar or relatively stronger when the performance criterion was applied to individual targets within a set (with target replacement) compared to an entire set without replacement. Future research should evaluate the boundaries of this effect by examining individual and set applications of criteria with a wide variety of educational teaching targets.

Third, skill-acquisition procedures are actually treatment packages comprised of multiple independent variables (e.g., prompts and fading, reinforcer type and schedule, task interspersal, intertrial-interval duration). The differences in these variables can result in more or less effective instruction. Less effective instruction should probably require a more stringent performance criterion, but a highly effective instructional approach might permit a less stringent performance criterion. In addition, a sufficiently stringent performance criterion might help compensate for variably effective instructional methods, which would be relevant for settings in which instructional integrity is difficult to maintain at consistently high levels. Research on the effects of such tradeoffs would be quite useful for practitioners. Of course, any deviation from accurate performance can weaken stimulus control, so a more stringent performance criterion might not offset the effects of less rigorous instruction.

Finally, as with all skill-acquisition studies, participant characteristics (e.g., attending skills, discrimination abilities) and setting characteristics (e.g., level of staff training, number of environmental distractions), and different instructional procedures (e.g., least-to-most vs. errorless prompting) likely interact with the observed findings, sometimes to such an extent that they constitute a systematic replication variable. This phenomenon is already evident in the performance-criterion literature. Whereas Fuller and Fienup (2018) and Longino et al. (2021) demonstrated that a criterion level of 90% correct resulted in strong maintenance after several weeks for their six participants, Richling et al. (2019; Experiment 4) found that a 90%-correct criterion level resulted in negligible maintenance for their four participants after only one week. There were numerous differences between studies including participant diagnosis, the skills being taught, prior experience with behavioral instruction methods, and the use of instructional components (error correction, most-to-least prompting, least-to-most prompting, respectively). It is likely that some interaction of these variables with the performance criterion resulted in the discrepant findings across studies. In addition, Pitts and Hoerger (2021) and Richling et al. (Experiment 4) conducted fairly similar studies with similar participant diagnoses and found different maintenance effects for the 90%-criterion level condition. This suggests some additional unknown variable was responsible for the discrepant findings. Thus, it is important that researchers evaluating performance criteria, at a minimum, fully describe all relevant characteristics for replication purposes (for a broader discussion,

see Jones et al., 2020). When evidence of an interaction occurs, that variable could then be explicitly evaluated. As the line of research progresses, if sufficient variability in findings occurs across studies, a potential solution would be to develop an assessment based on the relevant literature that could be used to help select performance-criterion characteristics for a single learner. This approach has become increasingly popular in the literature as a way to individualize instruction (e.g., Kodak & Halbur, 2021).

3 | CONCLUSION

Performance criteria are a necessary feature of skill-acquisition procedures and are widely used in practice and research. Unfortunately, their specific arrangements and their effects are probably best characterized as lore since such little experimental research exists on the topic. We hope this review and suggestions for future research are helpful in guiding this line of research such that our eventual evidence base is better matched to the ubiquitous use of performance criteria. Finally, the limited number of empirical investigations in this research line thus far have come from a correspondingly limited number of research groups and there is a need to expand the number of research groups investigating the phenomena.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

There are no associated data for this manuscript.

ORCID

Daniel M. Fienup https://orcid.org/0000-0002-0593-2929
James E. Carr https://orcid.org/0000-0002-6445-2992

ENDNOTES

[1] We use the term "performance criterion" because it is does not connote high proficiency. We suggest that the term "mastery criterion" be reserved for instances in which behavior reaches not only an initial accuracy criterion, but also other relevant goals such as generalization and maintenance. See also Richling et al. (in press).

[2] For instructional arrangements in which trials are not aggregated into sessions or trial blocks, the criterion level and criterion-level frequency value can be expressed in terms of a specific number of consecutive correct trials.

[3] Although McDougale et al. (2020) did not limit their review to studies that used DTI, the fact that the majority of studies used trial-based performance criteria establishes the relevance of the McDougale et al. findings to the present review.

[4] Other recent studies have focused on other aspects and uses of performance criteria, including evaluating criterion-level rigor on derived relations (Fienup & Brodsky, 2017) and the effects of applying mastery criteria to individual versus sets of verbal operants (Wong et al., 2021).

REFERENCES

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

Cummings, A. R., & Carr, J. E. (2009). Evaluating progress in behavioral programs for children with autism spectrum disorders via continuous and discontinuous measurement. *Journal of Applied Behavior Analysis*, *42*(1), 57–71. https://doi.org/10.1901/jaba.2009.42-57

DiGennaro Reed, F. D., Reed, D. D., Baez, C. N., & Maguire, H. (2011). A parametric analysis of errors of commission during discrete-trial training. *Journal of Applied Behavior Analysis*, *44*(3), 611–615. https://doi.org/10.1901/jaba.2011.44-611

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

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/10.1007/s40617-017-0201-0

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

Jones, S. H., St. Peter, C. C., & Ruckle, M. M. (2020). Reporting of demographic variables in the. *Journal of Applied Behavior Analysis*, *53*(3), 1304–1315. https://doi.org/10.1002/jaba.722

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

Kodak, T., & Halbur, M. (2021). A tutorial for the design and use of assessment-based instruction in practice. *Behavior Analysis in Practice*, *14*(1), 166–180. https://doi.org/10.1007/s40617-020-00497-w

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

Longino, E. B., McDougale, C. M., Richling, S. M., & Palmier, J. (2021). Evaluation of a mastery criteria and maintenance using a most-to-least prompting hierarchy. *Behavior Analysis in Practice*. Advance online publication. 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

Majdalany, L. M., Wilder, D. A., Greif, A., Mathisen, D., & Saini, V. (2014). Comparing massed-trial instruction, distributed-trial instruction, and task interspersal to teach tacts to children with autism spectrum disorders. *Journal of Applied Behavior Analysis*, *47*(3), 657–662. https://doi.org/10.1002/jaba.149

Mayer, G. R., Sulzer-Azaroff, B., & Wallace, M. (2012). *Behavior analysis for lasting change* (2nd ed.). The Cambridge Center.

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

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

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.

Reiser, R. A., Driscoll, M. P., Farland, D. S., Vergara, A., & Tessmer, M. C. (1986). The effects or various mastery criteria on student performance and attitude in a mastery-oriented course. *Educational Communication and Technology*, *34*(1), 31–38.

Reiser, R. A., Driscoll, M. P., & Vergara, A. (1987). The effects of ascending, descending, and fixed criteria on student performance and attitude in a mastery-oriented course. *Educational Communication and Technology*, *35*(4), 195–202.

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

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

Schneider, A., Fienup, D. M., Gussin, R., & Moss, P. (in press). A preliminary analysis of mastery criterion frequency values: Effects on acquisition and maintenance. *Behavioral Interventions*.

Semb, G. (1974). The effects of mastery criteria and assignment length on college-student test performance. *Journal of Applied Behavior Analysis*, *7*(1), 61–69. https://doi.org/10.1901/jaba.1974.7-61

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*: Basic Books.

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

Wong, K. K., & Fienup, D. M. (in press). Units of analysis in acquisition-performance criteria for "mastery": A systematic replication. *Journal of Applied Behavior Analysis*.

How to cite this article: 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*, 1–8. https://doi.org/10.1002/bin.1827