A Comparison of Quiz Grading Criteria, Exam Question Type and Novelty, and their Effects on Exam Scores

Samantha Dalfen

The Graduate Center, City University of New York

How does access to this work benefit you? Let us know!

Follow this and additional works at: https://academicworks.cuny.edu/gc_etds

Part of the Applied Behavior Analysis Commons

Recommended Citation

https://academicworks.cuny.edu/gc_etds/2845

This Dissertation is brought to you by CUNY Academic Works. It has been accepted for inclusion in All Dissertations, Theses, and Capstone Projects by an authorized administrator of CUNY Academic Works. For more information, please contact deposit@gc.cuny.edu.
A COMPARISON OF QUIZ GRADING CRITERIA, EXAM QUESTION TYPE AND NOVELTY, AND THEIR EFFECTS ON EXAM SCORES

by

SAMANTHA DALFEN

A dissertation submitted to the Graduate Faculty in Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

2018
A Comparison of Quiz Grading Criteria, Exam Question Type and Novelty, and their Effects on Exam Scores

by

Samantha Dalfen

This manuscript has been read and accepted for the Graduate Faculty in Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

__________________________________________
Date                                      Daniel Fienup
                                          Chair of Examining Committee

__________________________________________
Date                                      Richard Bodnar
                                          Executive Officer

Supervisory Committee:

Alexandra Logue

Emily Jones

Kenneth Reeve

Jason Vladescu

THE CITY UNIVERSITY OF NEW YORK
ABSTRACT

A Comparison of Quiz Grading Criteria, Exam Question Type and Novelty, and their Effects on Exam scores

by

Samantha Dalfen

Advisor: Daniel Fienup

Students in higher education perform better on exams when they complete frequent quizzes on the assigned reading material; but little research has investigated the different ways quizzes can be administered and how these variations affect quiz and exam performance. One variable that influences how quizzes are administered is the grading criterion. The standard practice grading criterion consists of receiving a score for the quiz based on the number of correctly answered questions. A passing criterion consists of requiring a student to obtain a certain score to earn full credit for the quiz. Previous research has found that students, particularly those who are at risk for failing, do significantly better on exams when there is a low-passing criterion as compared to a higher-passing criterion. Currently, there is no research that compares the effects of quizzes with a standard practice criterion and quizzes with a passing criterion on exam scores. The present study sought to compare a low-passing criterion and a standard practice criterion for quizzes and their effects on exam scores. Furthermore, we manipulated the type of exam question and whether the question has been replicated from a previous quiz across both a low-passing criterion and standard practice. This study replicated previous research demonstrating that students performed better on low-passing criterion exam questions. Additionally, students performed better on comprehension questions and replicated questions. Future research should
conduct a parametric analysis of passing criteria to determine the optimal criteria for exam performance.

*Keywords:* Grading criterion, passing criteria, question type, question novelty.
ACKNOWLEDGEMENTS

There are several people without whom this document would not have been possible. First, I owe a tremendous thank you to Dr. Daniel Fienup, who took a chance on me by letting me assist him on various research projects for one summer when I was an undergraduate student at Binghamton University. I could not have completed this journey without your guidance, support, and mentorship. Thank you to Dr. Logue and Dr. Jones for their thoughtful feedback and support as internal committee members. Thank you also to my external committee members, Dr. Kenneth Reeve and Dr. Jason Vladescu, who provided important feedback on this manuscript. A big thank you to my husband David Omansky, who always provided me with support, encouragement, and the confidence to persevere through this endeavor. Thank you to my mother, Elaine Dalfen, father, Allan Dalfen, and sisters, Ariella and Jessica Dalfen, who were always by my side to help me celebrate the good times and encourage me during the difficult times. Finally, thank you to my friends who provided me with support, humor, and fantastic times throughout graduate school.
Table of Contents

Introduction ................................................................................................................................. 1

Grading Criteria ........................................................................................................................... 2

Outcome Discrepancies .............................................................................................................. 3

Passing Criteria as a Potential Solution ..................................................................................... 4

Exam Variables ............................................................................................................................ 4

Question Type and Novelty ......................................................................................................... 5

The Current Study ....................................................................................................................... 6

Method ......................................................................................................................................... 7

Participants and Setting .............................................................................................................. 7

Materials ...................................................................................................................................... 8

Quizzes ....................................................................................................................................... 8

Exams ....................................................................................................................................... 9

Dependent Variables .................................................................................................................. 10

Independent Variables .............................................................................................................. 11

Grading Criterion ....................................................................................................................... 11

Question Type .............................................................................................................................. 12

Question Novelty ......................................................................................................................... 12

Grading Reliability and Intervention Integrity ........................................................................... 12
Procedure.................................................................13

Standard Practice Condition...............................................14

Low-Passing Condition ....................................................14

Results .............................................................................14

Quiz Performance .............................................................14

Exam Performance .............................................................14

Results by Performance Level .............................................16

Discussion .........................................................................17

Contributions ....................................................................18

Grading Criteria Contribution .............................................18

Question Type and Novelty Contribution .............................19

Student Performance Level Contribution .............................21

Implications ........................................................................22

Limitations and Future Research .........................................23

References ..........................................................................34
LIST OF TABLES

Table 1. Breakdown of Exam Questions by Question Type and Novelty .......................... 26
Table 2. Quiz Results by Performance Group .................................................................. 27
Table 3. Exam Results by Question Type, Question Novelty, Grading Criteria, and Performance Group.................................................................................................................. 28
LIST OF FIGURES

Figure 1. Mean Quiz Scores .................................................................29

Figure 2. Percentage of Completed Quizzes ...........................................30

Figure 3. Mean Scores by Question Type and Novelty ..............................31

Figure 4. Mean Question Novelty Scores ...............................................32

Figure 5. Percentage of Exam Questions Answered Correctly .................33
A Comparison of Quiz Grading Criteria, Exam Question Type and Novelty, and their Effects on Exam Scores

Large university courses tend to utilize lecture-based teaching, which depends heavily on students completing reading outside of class (Ventura & Onsman, 2009). Unfortunately, students report not completing reading assignments, especially students who are failing the course. Furthermore, instructors have limited resources to check if students read and understand the course material (Burchfield & Sappington, 2000). One solution is to implement frequent quizzes concerning the reading material. Several studies have demonstrated that providing frequent quizzes on the assigned reading material improved overall exam scores when compared to no quizzes (Geiger & Glasgow, 1976; Johnson & Kiviniemi, 2009). To explore the variables that influence the positive effects of frequent quizzes on exam scores, Dalfen, Fienup, and Sturmey (2018) examined a contingency for accuracy, defined as assigning a grade for each quiz based on the student’s response accuracy compared to providing frequent quizzes with no contingency for accuracy. The researchers found that simply providing students with frequent quizzes did not improve exam scores, and that a contingency was necessary for students to complete the quizzes with high accuracy and subsequently perform better on the exams. However, these effects were not seen equally across all students. The positive outcomes of frequent quizzes were least apparent in students who were failing the course. This may be because students who were failing the course tended to complete fewer quizzes and received relatively lower scores on quizzes (Dalfen et al., 2018).

Grading Criteria

Researchers have speculated that due to the effort required to earn course credit (the reinforcer), students at risk for failing a course may have received lower scores on the quizzes
and completed fewer quizzes than did the students not at risk for failing the course. A potential solution is to manipulate the requirement to earn full credit, or grading criterion, in a way that would decrease the response effort required to earn the course credit. However, there is little reliable evidence supporting a specific grading criterion for quizzes that increases both quiz and subsequent exam scores. Some researchers have found that having a high-performance grading criterion for quizzes produces better outcomes on both quizzes and exams (Carter, Carter, Couture, & Wright, 1976; Johnston & O’Neill, 1973; Semb, 1974) while other researchers have found that a lower performance grading criterion produces better outcomes on exams (Carlson & Minke, 1975; Dalfen, Fienup, & D’Ateno, in preparation). The discrepancy in outcomes may stem from the different ways that a researcher can implement grading criteria for quizzes.

There are different grading criteria for quizzes an instructor can use to promote quiz accuracy, completion, and exam scores. In **standard practice**, the grade a student receives on a quiz is equal to the percentage of the credit earned for that quiz, which is then averaged into a final grade. For example, if a student answered 6 questions correctly out of 10, he or she will receive a 60% on that quiz, which will be factored into the final grade. Another grading criterion is a **mastery criterion**. Typically, the instructor will set a criterion for the quizzes, and students must meet that criterion to earn full credit for that quiz and advance to the next quiz. If the student does not meet the criterion, he or she will not receive credit and will need to keep retaking the quiz until mastery is demonstrated on that quiz. Previous research has shown that a mastery criterion for quizzes resulted in improved exam scores when compared to quizzes graded with the standard practice method (Kulik, Kulik, & Bangert-Drowns, 1990). Another type of grading criterion is a **passing criterion**. A passing criterion is a binary criterion set for a quiz that students must meet to earn full credit. Any grade below the criterion results in no credit
earned, and any grade above the criterion is equivalent to full credit. Like standard practice, with a passing criterion the student does not get another opportunity to retake the quiz, and he or she can advance to the following quiz regardless of his or her score on the previous quiz.

**Outcome Discrepancies.** The three different grading criteria for quizzes vary along different dimensions and utilize different strategies to increase quiz and exam scores. Standard practice and passing criteria involve taking a quiz one time and do not let students retake the quiz, whereas mastery criterion arrangements allow for as many re-takes as one can complete in a semester. Kulik et al. (1990) conducted a meta-analysis on studies that compared mastery criteria to standard practice at college-level and found 91 studies. Sixty-seven of these studies reported a significant difference in exam scores between conditions, with students performing better on exams from the mastery criterion condition than the standard practice condition. Of these studies, 12 provided data on student performance by dividing the students into different groups based on their “aptitude.” These data were inconclusive; one study demonstrated stronger effects for high-performing students (Kulik & Kulik, 1976), and some studies demonstrated stronger effects for low-performing students (Cross & Semb, 1976; Fehlen, 1976; Fernald & DuNann, 1975; Maverech. 1986; Thompson, 1980). These results do not provide unequivocal evidence that mastery criteria benefit all students equally or explains why students tend to perform better on exams when mastery criteria are implemented for quizzes. Mastery criterion arrangements include extra practice, which could account for why mastery criteria sometimes produce better outcomes than standard practice; however, the extra practice also increases the effort necessary to receive the reinforcer that the instructor arranged for meeting the mastery criteria (i.e., points toward course grade). Additionally, some students may require multiple retakes before mastering a quiz, thus delaying their progress and preventing them from
completing the course. These requirements can cause students to receive a failing grade or withdraw from the course. Carlson and Minke (1975) discussed high withdrawal and failure rates in courses with mastery criteria, citing a 29% failure rate in one of Keller’s courses and a 25% withdrawal rate in a previous study conducted by Minke and Carlson (1972). Passing criteria, particularly a low-passing criterion, may solve the above stated problems by reducing the effort necessary to earn the reinforcer and giving students an opportunity to earn full credit for far less than a perfect score.

**Passing Criteria as a Potential Solution**

Dalfen et. al. (in preparation) compared the effects of a high vs. low-passing criterion on exam scores across high, mid, and low-performing students. In the 60-passing condition, students needed to answer 6 out of 10 questions correctly per quiz to earn full credit on that quiz. In the 90-passing condition, students needed to answer 9 out of 10 questions correctly per quiz to earn full credit on that quiz. The instructor also administered exams that tested the reading material from each condition. Results indicated that all students performed better on exam questions related to readings quizzed using the low-passing criterion. The low-performing students benefitted the most from the low-passing criterion; these students scored almost 7% better on examinations following implementation of the low-passing criterion for quizzes as compared to a nearly 3% increase with the highest-performing students. While the effects of passing criteria have generally been under-researched, this one study demonstrated promise that reducing the effort necessary to receive the reinforcer can be particularly effective with the lowest performing students. Nevertheless, this grading criterion has yet to be compared to either standard practice or the mastery criterion methodologies.
Exam Variables

As discussed above, there is a need for research that compares different grading criteria for quizzes and their effects on exam scores to discover which grading criterion produces better student exam scores. In addition, there are elements of the exams themselves that may affect how students perform on those exams. For example, the type of question asked on an exam and whether the students have a learning history with the question both influence the likelihood that the student will answer the question correctly (Semb, 1974). With respect to different types of exam questions, one way to classify these questions is using the Bloom Taxonomy of thinking, a multi-tiered model of classifying thinking in which the tiers increase in complexity (Forehand, 2010). The bottom tier of thinking is labeled “knowledge” and is defined as retrieving facts. The next tier, “comprehension”, is defined as applying knowledge to novel situations through examples, interpretation, or summarization. Based on this taxonomy, there are multiple ways one can format exam questions to ensure that students can do more than simply memorize facts from the study material. Since its inception, many teachers have used this model as a method to help develop examinations and to confirm that their exams are balanced across the different tiers of the taxonomy. Currently, it remains unclear if different grading criteria: (1) improve student’s retention for previously tested quiz items, (2) improve generalization to novel exam questions, (3) improve performance on facts based on the reading material, or (4) improve performance on comprehension-based questions.

Question Type and Novelty. There has been one article that has examined the effects of different mastery criteria on student performance across different question types (Semb, 1974). Semb described fact questions as “study questions” and comprehension questions as “probe questions”. Study questions required students to recall specific facts from their study material.
An example of a study question described in Semb (1974) was “name the three specific freedoms stressed by Neill” (p.63). Probe questions provided a discriminative stimulus ($S^D$) that was different from the studied information. By describing an example or situation, probe questions would require students to apply the factual information from their study material to the new example or situation. The original $S^D$ of the term or definition would not be present in the test question. An example of a probe question described by Semb (1974) was “If you train a child to discriminate between a horizontal line and a vertical line, then gradually change the vertical line until it is diagonal, what behavioral procedure are you using?” (p. 63). Furthermore, study questions were directly replicated from study materials given to the students while probe questions were novel questions, testing for both response maintenance and stimulus generalization. While students performed worse on the probe questions than the study questions, it is unclear whether the lower performance was due to the questions being novel, or because the student’s responding could not generalize responding from the original $S^D$ (study material) to the novel $S^D$ (exam question). Together these results constitute a promising demonstration that manipulating the type of exam question does affect student performance, but there has yet to be any research conducted on these variables with different grading criteria.

**The Current Study**

Therefore, the current study compared a low-passing criterion to standard quiz practice to determine which procedure produced: (1) better quiz scores, defined as the number of questions answered correctly for each quiz in each condition; (2) better quiz completion, defined as clicking the “complete” icon at the end of each quiz; and (3) better exam scores, defined as the number of questions answered correctly on exams. Additionally, the current study evaluated the effects of exam question type and novelty on accuracy of responding on exams. For this study,
response maintenance was defined as accurately responding to a question on an exam that was a direct replication of a question from a previous quiz. Stimulus generalization was defined as accurately responding to a novel question on the exam that relates in some way to previous questions from quizzes. “Fact” questions were defined as testing maintenance of responding by either providing the $S^D$ (a term) and having students select the correct conditional discrimination (definition) amongst a choice of four, or the symmetrical response (the $S^D$ is the definition while the conditional discrimination is the term). “Comprehension” questions were defined as testing comprehension of facts by describing a situation or example as the $S^D$ and having students select the correct conditional discrimination (term) amongst a choice of four.

The experimenter used a 2x2x2 factorial design across 105 students from a course on learning to determine the effects of the grading criteria for quizzes, the type of question on the exam, and the question novelty of the exam questions on exam scores. There were 12 chapters of material from the assigned textbook in the course, and each chapter had an assigned quiz. The quiz was either graded using a low-passing grading criterion or using standard practice grading. There were three exam units. Each unit tested on four chapters of material. Each unit consisted of two chapters that had an assigned quiz employing the low-passing criterion and two chapters that had an assigned quiz employing standard practice grading. Quizzes counted for a total of 20% of the students’ final grade. Students were divided into groups based on their self-reported Grade Point Averages (GPA) to determine whether they were high (any GPA from 3.61 to 4.0), mid (any GPA from 3.10 to 3.60), or low-performing (any GPA from 2.0 to 3.09) students to determine which grading method aided which type of student the most.

Method

Participants and Setting
One section of an introductory learning and behavior analysis course participated in this study. One hundred and five undergraduate students were enrolled in the course. To be included in this study, a student had to complete all the course exams on the days the exams were scheduled. The experimenter removed 22 students from data analysis for either withdrawing from the course, not completing one or more exams, missing one or more exams, or taking a makeup exam. A total of 83 students were included in data analysis. Seventy-five of those students self-reported his or her grade points average (GPA). A wide range of students from freshmen to seniors were enrolled in this course. Students ranged in age from 18 years and older. This experiment was designed to be part of the course requirements and did not require additional participation above and beyond the course requirements.

The exams took place in a large lecture room that seated up to 150 students. The lecture room had rows of connecting tables with corresponding chairs. The front of the lecture room had a small podium with a computer that the experimenter used in presenting lectures. The experimenter delivered all lectures accompanied by a PowerPoint slideshow, which was projected on the blackboard. Students completed online quizzes outside of class in self-determined locations.

**Materials**

All students were required to purchase the textbook *Learning and behavior (7th ed. or 6th ed.)*, Chance, 2013).

**Quizzes.** The experimenter constructed 12 quizzes, one for each of the 12 textbook chapters covered in the course. Each quiz consisted of 10 multiple choice questions, with four answer options per question marked A-D. Five questions were “fact” questions, defined as providing the S^D (the term) and having students select the correct conditional discrimination
(definition) amongst a choice of four, or the symmetrical response (the $S^D$ is the definition while the conditional discrimination is in question format). The other five questions were “comprehension” questions, defined as describing a situation or example as the $S^D$ and having students select the correct conditional discrimination (term) amongst a choice of four. All 12 quizzes were reviewed by three other instructors who have taught the course to determine that all quizzes were of equal difficulty and readability. Furthermore, this study was piloted in other sections of the course to confirm feasibility of the procedure. Of the 12 quizzes, 6 were assigned to a low-passing criterion condition and 6 were assigned to a standard practice condition.

Students accessed the quizzes through Blackboard©. Students needed a desktop or laptop computer with an internet connection to complete the online quizzes. However, students were only able to use a PC; students were not able to access the quizzes on a Mac©, smart phone, or tablet. The experimenter used a PC to score all online quizzes. Blackboard© introduced each quiz in the low-passing condition with a note saying “This is quiz (#) and it is due on (date) by (time). You must answer 6 out of 10 questions correctly to receive full credit. This quiz will not be timed and feedback will be provided on all answers. Good luck!” Blackboard© introduced each quiz in the standard practice condition with a note saying “This is Quiz (#) and it is due on (date) by (time). This quiz score will be averaged into your total score. Answering 6 out of 10 questions correctly would result in a score of 60% averaged into your final grade for the course. This quiz will not be timed and feedback will be given on all answers. Good luck!” Students had to click submit at the end of each quiz for the score to be tabulated. Blackboard then displayed the student’s quiz score along with which questions they answered correctly, and which questions they answered incorrectly.
**Exams.** The experimenter constructed three exams on standard double-sided computer paper. Each exam covered 4 chapters of material from the textbook and included 40 questions composed of 10 questions from each chapter. Table 1 displays the following breakdown of the exams. Twenty of the 40 questions (two chapters) were drawn from quizzes that used the low-passing criterion, while the other 20 (two chapters) questions were drawn from quizzes that used the standard practice criterion. In addition, of the 40 questions on each exam, 20 were “fact” questions, and 20 were “comprehension” questions. Further, of the 40 questions, each exam had a total of 16 questions that were direct replications from the quizzes, while 24 questions were novel. Two questions from each chapter were replicated “fact” questions from each quiz while two questions from each chapter were replicated “comprehension” questions from each quiz. Therefore, 40% of the questions on the exams were questions that were identical to questions tested on the quizzes while 60% of the questions on the exams were novel questions. There were 3 exams with 40 questions each, for a total of 120 exam questions over the course of the semester. Students recorded their answers on a 40-answer Scantron© with A through E choice options. Students were required to use a number 2 pencil to fill in their answers on the Scantron© sheet. The experimenter independently graded each exam. The first exam included an extra question at the bottom of the exam stating “please report your overall college GPA. If you do not have a college GPA, please report your overall high school GPA.” There was a blank space where the student reported his or her GPA.

**Dependent Variables**

There were three dependent variables in this study: quiz accuracy, quiz completion, and exam accuracy. The first dependent variable was the percentage correct score on each quiz. The experimenter measured percentage correct by dividing the student’s number of correct answers
on the quiz by the total number of questions on that quiz, and then multiplying by 100. The second dependent variable was the percentage of students who completed each quiz, defined as clicking “submit” at the end of the quiz. The experimenter measured percentage of completed quizzes by dividing the number of students who completed the quiz by the total number of students, and then multiplying by 100. A student who did not complete a quiz would receive a grade of 0 for that quiz. The third dependent variable was the percentage of correct exam question answers. The experimenter measured percentage correct by dividing the student’s number of correct answers on the exam by the total number of questions on that exam, and then multiplying by 100. The experimenter entered all quiz and exam scores into an Excel© spreadsheet. The experimenter tabulated all mean scores for quizzes and exams on the excel spreadsheet. The experimenter did not enter the data into Excel© until after the end of the semester once grades were posted.

**Independent Variables**

This study employed a 2x2x2 factorial design to evaluate the effects of quiz grading criteria, exam question type, and exam question novelty on exam scores. Each independent variable had two levels, (two grading criteria, two question types, and an exam question could be novel or not). There were three exams that each covered four chapters of material, two chapters that had a low-passing quiz grading criterion and two chapters that had a standard practice quiz grading criterion. The course grading breakdown was as follows: exams were 80% of the final grade, and quizzes were 20% of the final grade.

**Grading criterion.** The first independent variable was the grading criterion for quizzes. The two levels were a **low-passing criterion** and **standard practice**. These two conditions alternated across the 12 quizzes (ABAB); A represents the standard practice condition while B
represents the low-passing criterion condition. During the low-passing condition, students were required to answer 6 out of 10 questions correctly to receive full credit for that quiz. For example, if a student answered 6 out of 10 questions correctly, that student received a 100% score that was averaged into their final grade. Any score less than 6 out of 10 was the equivalent of 0% and was averaged into their final quiz score. During the standard practice condition, students received a score based on the number of questions that were answered correctly, and that score was averaged into their final grade. For example, if a student answered 6 out of 10 questions correctly, the student received a 60% score that was averaged into his or her final quiz grade. Each condition consisted of six quizzes.

**Question type.** The second independent variable was the question type that was asked on the exam. The two levels were fact and comprehension questions. Fact questions were defined as testing memorization by either providing the $S^D$ (a term) and having students select the correct conditional discrimination (definition) amongst a choice of four, or the symmetrical response (the $S^D$ is the definition while the conditional discrimination is the term). Comprehension questions were defined as testing comprehension of facts by using a situation or example as the $S^D$ and having students select the correct conditional discrimination (term) amongst a choice of four. There was a total of 20 fact questions and 20 comprehension questions per exam.

**Question novelty.** The third independent variable was exam question novelty. The two levels were maintenance and generalization questions. Response maintenance was defined as accurately responding to a question on an exam that was a direct replication of a question from a previous quiz. Stimulus generalization was defined as accurately responding to a novel question on the exam that relates in some way to previous questions from quizzes but was never actually
present on a quiz. There were 24 questions per exam that were generalization questions, and 16 questions per exam that were maintenance questions.

**Grading Reliability and Intervention Integrity**

The experimenter and another independent observer conducted inter-observer agreement (IOA) assessments for 30% of all exam scores to determine if they were coded reliably. The instructor of the course de-identified the scantrons and assigned each student an identifying code. The experimenter and the independent observer scored the students’ exam answer keys by question type and novelty and marked how many out of each category the student answered correctly. For example, for exam 1, there were 2 questions on chapter 1 that were replicated fact questions. The experimenter and the independent observer both marked how many of these questions the student answered correctly. Agreements were instances of both observers recording the same score for a question type. IOA was calculated by dividing the number of agreements by the total number of questions and multiplying by 100. IOA results indicated that the experimenter and the observer agreed for 99.6% of responses.

A different independent observer examined all 12 quizzes on Blackboard© to observe if the procedural steps were implemented as planned. The steps were (1) The experimenter posted the quiz on blackboard© and students could access the quiz, (2) There was a note for the students to read prior to starting each quiz stating which grading criterion the quiz would use, (3) The note stated the time and date for when the quiz was due, and (4) The note stated that students would receive feedback on all his or her answers. The independent observer reported that all above steps were completed with 100% integrity across all 12 quizzes.

**Procedure**
During the first class, the instructor explained the Blackboard© quiz system, including that each quiz was designed to test facts and comprehension from a specific reading assignment. The instructor posted each quiz immediately after the relevant class to ensure students had time to complete the quiz. The instructor told students that quizzes were due once a week before the next class began. Quizzes alternated between using a low-passing criterion and standard practice. During the first class, the instructor described the course, the Blackboard© quiz system, and stated that quizzes were graded either using a low-passing criterion or standard practice. There was a total of six quizzes using each of the two grading criteria. Each quiz had 10 questions, 5 that were “fact” questions, and 5 questions that were “comprehension” questions. Each exam had 40 questions that tested from four chapters from the assigned textbook, two chapters had a low-passing criterion quiz and two chapters had a standard practice quiz.

**Standard Practice.** Students received a score based on the number of questions answered correctly. These scores averaged into the students’ final grades. For example, if a student answered 7 out of 10 questions correctly, he/she received a 70% that was averaged into his/her quiz grade.

**Low-Passing Condition.** Students had to answer at least 6 out of 10 multiple choice questions correctly to receive full credit on the quiz. If the student answered five or less questions correctly, the student received 0% for the quiz.

**Results**

**Quiz Performance**

Figures 1 and 2 represent student performance related to quizzes. Figure 1 displays the students average quiz score when a quiz had a low-passing criterion or the standard practice criterion. For low-passing criterion quizzes, the average quiz score across all six quizzes in the
condition was 67.2% (SEM=1.9). For standard practice criterion quizzes, the average quiz score across all six quizzes in the condition was 63.5% (SEM=2.3). The experimenter conducted a two-tailed paired samples t-test and found that students scored significantly higher on low-passing criterion quizzes, \( t(82) = -2.186, p=0.03 \). Figure 2 displays the percentage of students who completed the quizzes for each grading criterion. For the low-passing criterion, students completed an average of 87.1% (SEM=1.9) of quizzes. For the standard practice criterion, students completed an average of 83.5% of quizzes (SEM=2.4). The experimenter conducted a two-tailed paired samples t-test and found no significant difference in quiz completion as a function of grading criterion, \( t(82) = -1.698, p>.05 \).

**Exam Performance**

The following results refer to student performance on exam questions. Figure 3 displays the average percentage of correctly answered exam questions as a function of grading criterion, question type, and question novelty. The experimenter conducted a three-way repeated measures ANOVA to determine the effect of quiz grading criterion, exam question type, and exam question novelty. Results indicated that there was a main effect for quiz grading criterion, with students performing better on exam questions from the low-passing criterion condition \((M=76.3\%, \ SEM=1.1)\) than from the standard practice condition \((M=74.4\%, \ SEM=1.1)\), \( F(1, 82)=6.141, p=.015 \). Results indicated that there was a significant main effect for exam question type, with students performing better on comprehension questions \((M=76.6\%, \ SEM=1.1)\) than fact questions \((M=74.1\%, \ SEM=1.1)\), \( F(1, 82)=9.629, p=.003 \). Results indicated that there was also a significant main effect for exam question novelty, with students performing better on maintenance questions \((M=78.9\%, \ SEM=1.2)\) than novel questions \((M=71.9\%, \ SEM=1.0)\), \( F(1, 82)=67.251, p=.0001 \).
In addition to three significant main effects, there was one significant interaction. Students performed significantly worse on novel questions from the standard practice condition when compared to all other conditions, $F(1, 82)=8.685$, $p=.004$. The respective data are displayed in Figure 4, which displays the percentage of correct exam questions across both grading criteria and question novelty. Students on average answered 73.9% (SEM=1.0) of the novel question from the low-passing condition correctly, 69.8% (SEM=1.3) of the novel questions from the standard practice condition, 78.7% (SEM=1.4) of the replicated questions from the low-passing condition, and 79.1% (SEM=1.3) of the replicated questions from the standard practice condition. Results did not indicate any additional significant interactions between grading criteria and question type, $F(1, 82)=3.192$, $p>.05$, question type and question novelty, $F(1, 82)=2.609$, $p>.05$, nor a three-way interaction, $F(1, 82)=1.414$, $p>.05$.

**Results by Performance Level**

The data reported to this point represent those from the 83 students who completed all three exams. What follows is an analysis of performance based on categorization from self-reported GPAs. Thus, the following outcomes are based on the 75 students who self-reported their GPA. The experimenter divided students into three groups based on the students self-reported GPAs. Twenty-six students who reported a 3.61-4.0 GPA were placed in the “higher-performing group,” 26 students who reported a 3.1 to 3.6 GPA were placed in the “mid-performing group,” and 23 students who reported a 2.0 to 3.09 GPA were placed in the “lower-performing group”.

Table 2 displays both average quiz completion and quiz accuracy per performance group for both grading criterion conditions. The experimenter conducted a 2x3 factorial ANOVA for performance level and quiz condition on quiz completion frequency and found no significant
difference between performance group and quiz completion, $F(2, 72)=.449, p>.05$, nor a significant interaction between performance group, grading criteria, and quiz completion, $F(2, 72)=2.319, p>.05$. The experimenter conducted a 2x3 Factorial ANOVA for performance level and quiz condition for quiz accuracy and found a significant difference in quiz scores for performance group, $F(1, 72)=7.449, p=.001$, with tukey post-hoc test revealing that the high-performing group performed significantly better on quizzes with an average score of 76.1% (SEM=2.9) than both the mid-performing group with an average score of 63.4% (SEM=2.9; $p=.003$) and low-performing group with an average score of 60.9% (SEM=3.1) ($p=.003$). There was no significant interaction between the quiz condition and performance group, $F(2, 72)=3.069, p>.05$.

Figure 5 displays the average percentage of correct exam questions across the low-passing and the standard practice conditions by performance group. The experimenter conducted a 2x2x2x3 Factorial ANOVA for performance level and grading criteria, question type, and question novelty for exam accuracy and found a significant difference in exam scores for performance group, $F(1, 72)=18.640, p=.0001$, with tukey post-hoc test revealing that the high-performing group performed significantly better on exams with an average score of 83.3% (SEM=1.5) than both the mid-performing group with an average score of 73.4% (SEM=1.5) ($p=.0001$) and low-performing group with an average score of 70.4% (SEM=1.6) ($p=.0001$). There was no significant difference between the mid and low-performing groups ($p>.05$). There were no other significant interactions between performance group and either grading criteria, $F(2, 72)=1.755, p>.05$, question type, $F(2, 72)=.473, p>.05$, or question novelty, $F(2, 72)=.605, p>.05$ (see Table 3 for data pertaining scoring by group on each question type, novelty, and grading criteria).
Discussion

Previous research comparing grading criteria for quizzes has focused on mastery criteria compared to standard practice. Recent research on passing criteria has shown them to be an effective method for improving exam scores across all student performance levels (Dalfen et al., in preparation). The current study evaluated the effects of a passing criterion and standard practice on quiz accuracy, quiz completion, and exam accuracy. This study found that students performed better on both quizzes from the low-passing criterion and related exam questions. Furthermore, this study evaluated whether exam question type and question novelty influenced exam performance, and whether the grading criterion interacted with these variables. The current study found that students performed better on comprehension compared to fact questions, maintenance compared to generalization questions, and that the standard practice criterion negatively impacted scores on generalization exam questions.

Contributions

The current study contributes to the current literature in three distinct ways. First, the current study compared the effects of a low-passing criterion and standard practice, a comparison that has not previously been made in the grading criteria research literature. The present study found new evidence indicating that a low-passing criterion for quiz accuracy facilitates better quiz and exam scores when compared with standard practice. Second, the current study evaluated the effects of fact, comprehension, replication, and novel exam questions on accuracy of responding on exams. This study has differed from previous studies by finding that students may perform better on comprehension questions when compared to fact questions and that the grading criterion chosen for quizzes may impact how students perform on novel questions on exams. Third, the experimenter divided students into three groups based on their reported GPAs
to determine if any of the above variables influenced quiz accuracy, quiz completion, and exam accuracy by performance level. The current study did not find any significant difference in how students with different GPAs perform on exams in terms of grading criteria, question type, or question novelty; however, the results support that a low-passing criterion is beneficial regardless of performance level.

**Grading Criteria Contribution.** The first novel contribution of this study is the comparison of two different grading criteria that have not been previously compared, and their effects on quiz completion, quiz scores, and exam scores. The current study found that regardless of performance group, students completed low-passing and standard quizzes at similar rates and scored higher on low-passing quizzes. The high level of quiz completion in both the low-passing and standard practice condition was consistent with past studies that used a contingency for quiz completion (Dalfen et al., 2018; Geiger & Glasgow, 1976; Johnson & Kiviniemi, 2009). The current study replicated previous research conducted with passing criteria (Dalfen et al., in preparation), which demonstrated that a low-passing criterion for quizzes produced better exam results than a high-passing criterion for quizzes. This may be due to the hypothesis described in the introduction that lowering the effort to earn a larger magnitude reinforcer (full credit) increased the likelihood of engaging in the quizzes. According to Skinner (1968), teachers arrange contingencies to increase desired behaviors via reinforcement and decrease undesired behaviors via extinction or punishment. In a college classroom setting, the primary consequence that can be used to arrange contingencies by the instructor is points added toward or subtracted from a student’s final grade. During the low-passing condition, students contacted a higher quality reinforcer for studying for their quizzes in the form of receiving a perfect score for a relatively low effort response. Given that students scored an average mean of 67% in the low-
passing condition (in which 60% was required to receive full credit), and 63% in the standard practice condition, they received a 100% of the available reinforcement in the low-passing condition, but only received 63% of the available reinforcement in the standard practice condition. This high-quality reinforcer in the low-passing condition could, in theory, increase studying behavior, which may explain why students performed better on questions from the chapters that had low-passing criterion quiz assignments: Students were more likely to study those chapters. In contrast, students received a lower-quality reinforcer in the form of receiving a lower score for engaging in the same studying behavior for the standard practice condition, thus decreasing the likelihood students studied those chapters for the exams.

**Question Type and Novelty Contribution.** The second novel contribution of the present research to the literature is the evaluation of question type and question novelty on exam scores, which have previously not been studied in situations with different quiz grading criteria. Students performed significantly better on comprehension questions than they did on fact questions on exams, which did not replicate Semb (1974)’s findings. Semb (1974) compared fact questions that were replicated from the study material with comprehension questions that were novel, while the current study separately examined the two variables of question type and novelty. By separating the variables of question type and novelty, each variable’s effects on exam performance could be analyzed. These results, which showed that students performed better on comprehension questions when compared to fact questions, challenge Bloom’s Taxonomy of thinking (Forehand, 2010), which described testing factual knowledge as the easiest form of questions students can answer, while testing comprehension requires more effort from the student. A possible explanation of the current results is that the comprehension questions utilized examples from current events that may have been more similar to student’s daily lives than were
the fact questions. Comprehension questions included an everyday example that needed to be paired with the correct behavior analytic terminology, whereas the fact questions included the behavior analytic terminology paired with the correct behavior analytic definition. This possible explanation of the results should be considered when creating exam and quiz material. Instructors should consider matching the distribution of comprehension and fact questions with the educational goals of the course because students may respond differently to different question types. The current study found that students performed better on maintenance questions over generalization questions. From a behavior analytic perspective, maintenance questions involve presenting the $S^D$ (the question) which the students have already had a learning history with on the quiz, while generalization questions require stimulus generalization from the original $S^D$ (quiz question) to a novel $S^D$ (exam question). Therefore, it is not surprising that students always received higher scores on maintenance when compared to generalization questions, regardless of question type. This study also identified an interaction between the grading criterion and question novelty. The interaction indicated that students performed significantly worse on generalization questions from the standard practice condition. This implies that students were more likely to generalize correct responding to novel questions in the low-passing criterion. This interaction may be due to the quality of reinforcer received in the low-passing versus standard practice conditions addressed above. Students received a higher-quality reinforcer for taking the low-passing criterion quizzes, and therefore studied more, and were thus more prepared for novel questions from the low-passing criterion on the exams.

**Student Performance Level Contribution.** The third novel contribution of the present study to the research literature concerns the examination of possible interactions of students’ self-reported GPAs with quiz and exam performance under various conditions. The current study
did not find any significant difference between performance group and how students scored on exams by grading criterion, question type, or question novelty. Furthermore, all performance groups completed a similar number of quizzes. While high-performers always consistently performed better on both quizzes and exams, there were no other differences between groups. The effects that were found across all grading criteria, question type, and question novelty. Regardless of performance level, students performed better on comprehension questions, on replicated questions, and questions from the low-passing condition on both exams and quizzes. These results did not replicate the results of Dalfen et al. (in preparation), which found that the low-performing group performed far better on exam questions from the low-passing criterion than the high-passing criterion when compared to other performance groups. However, Dalfen et al. (in preparation) compared two different passing criteria while the current study compared two different grading criteria. It is possible that when compared to a high-passing criterion, the low-passing criterion is more effective for low-performers, but there is less of a difference between performance groups when a low-passing criterion is compared to standard practice.

**Implications**

The results of the current study lead to several implications for higher education instructors. First, the current study provides evidence that a low-passing criterion is an effective alternative to setting a mastery criterion for producing better quiz and exam scores when compared to standard practice. Mastery criteria are cumbersome to implement for instructors, specifically in large lecture courses. It can be difficult for instructors to track the progress of each student across the assigned quizzes to see which students are making sufficient progress and which students are falling behind. Furthermore, Carlson and Minke (1975) discussed the negative effects for students, due to many failing to progress through the course and receive
passing grades. The current study provides evidence for a successful alternative for setting a grading criterion that is less effort for the instructor to implement and does not prevent students from advancing in the course in a timely manner. Second, instructors should consider implementing a low-passing criterion instead of standard practice for quizzes to promote generalization from maintenance exam questions derived from the quizzes to novel exam questions. While students performed similarly on maintenance questions regardless of grading criteria, students scored a mean of a C grade on novel questions derived from the low-passing criterion compared to a mean of a C- grade on novel questions derived from the standard practice condition. These results demonstrate an educational significance that higher-education instructors may want to replicate in their own classrooms. Third, the current study demonstrates that students scored better on comprehension questions compared to fact questions, and that instructors may want to construct exam questions that utilize everyday examples to promote better student responding on exam questions.

**Limitations and Future Research**

There are a few limitations to this study that could be remediated in future research. First, there are multiple differences between fact and comprehension questions that are unrelated to the question type. For example, comprehension questions tend to have a longer word count and less jargon used. It is possible that one of these variables influenced student exam performance and not the question type itself. Future research should control for these variables to determine why students perform better on one question type over the other. Second, the experimenters were not able to observe students’ studying behaviors outside of class. Therefore, we cannot with certainty determine whether the different conditions influenced studying behavior, only the scores on the exams. Future research should replicate this study in a lab setting to control for outside variables.
and to better observe studying behaviors and their relationships to exam scores. Conducting this type of study in a laboratory could confirm whether receiving a higher quality reinforcer for quizzes increases the frequency or duration of study behavior. Third, the current study did not obtain actual GPA scores for students and therefore the performance group divisions are based on the students’ self-reported GPA scores. These scores may not be accurate, and some students chose not to report their scores entirely. A greater number of students with low GPAs may have abstained from reporting their GPAs, potentially skewing the sample of students self-reported their GPAs There is currently no research determining whether self-reported GPA or actual GPA scores is more accurate to the student’s actual performance respective to that class. Future research should obtain GPA scores from the university’s registrar before the start of the study to compare both self-reported and actual GPAs respective to performance groups within the course. Fourth, the current study only evaluated these variables within one Introduction to Learning class. Therefore, future research should replicate this study in other courses, with different material, and with different instructors to assess the degree to which these results generalize. Fifth, the current study only evaluated one low-passing criterion with standard practice. While this study discovered further evidence to support that a low-passing criterion is more beneficial to students in comparison to standard practice, there has yet to be any research that compares mastery criteria, passing criteria, and standard quiz practice. Future research should include a parametric analysis for different passing and mastery criteria to determine which criterion from each grading method produces better quiz and exam performance. An important extension of the current study would be to compare the best passing, mastery criterion, and standard practice grading methods to determine which grading method for quizzes produces better student performance for both quiz and exam scores. The current study found very promising results that
further support that a low-passing criterion for quizzes does improve exam scores when compared to standard practice, and that other variables such as question type and novelty can also influence performance.
Table 1

*Breakdown of Exam Questions by Question Type and Novelty*

Exam Example: 40 questions covering 4 chapters

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Standard Practice Quiz Grading</th>
<th>Low-Passing Quiz Grading</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Criterion: 20 questions</td>
<td>Criterion: 20 questions</td>
</tr>
<tr>
<td></td>
<td>Chapter 1</td>
<td>Chapter 3</td>
</tr>
<tr>
<td>Fact Questions</td>
<td>5 questions</td>
<td>5 questions</td>
</tr>
<tr>
<td>Maintenance/Generalization Distribution</td>
<td>2 Maintenance, 3 Generalization</td>
<td>2 Maintenance, 3 Generalization</td>
</tr>
<tr>
<td>Comprehension Questions</td>
<td>5 questions</td>
<td>5 questions</td>
</tr>
<tr>
<td>Maintenance/Generalization Distribution</td>
<td>2 Maintenance, 3 Generalization</td>
<td>2 Maintenance, 3 Generalization</td>
</tr>
</tbody>
</table>
### Table 2

**Quiz Results by Performance Group**

<table>
<thead>
<tr>
<th>Performance Group</th>
<th>Low-Passing Condition</th>
<th>Standard Practice Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Completion</td>
</tr>
<tr>
<td>High Performance</td>
<td>76%</td>
<td>(SEM= 3.4)</td>
</tr>
<tr>
<td></td>
<td>88%</td>
<td>(SEM= 3.0)</td>
</tr>
<tr>
<td>Mid Performance</td>
<td>68%</td>
<td>(SEM= 3.0)</td>
</tr>
<tr>
<td></td>
<td>89%</td>
<td>(SEM= 2.8)</td>
</tr>
<tr>
<td>Low Performance</td>
<td>61%</td>
<td>(SEM= 3.0)</td>
</tr>
<tr>
<td></td>
<td>87%</td>
<td>(SEM= 3.9)</td>
</tr>
</tbody>
</table>
Table 3

*Exam Results by Question Type, Question Novelty, Grading Criteria, and Performance Group*

<table>
<thead>
<tr>
<th>Performance Group</th>
<th>Low-Passing Condition</th>
<th>Standard Practice Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FR</td>
<td>FN</td>
</tr>
<tr>
<td>High Performance</td>
<td>86%</td>
<td>77%</td>
</tr>
<tr>
<td></td>
<td>(2.4)</td>
<td>(1.5)</td>
</tr>
<tr>
<td>Mid Performance</td>
<td>75%</td>
<td>73%</td>
</tr>
<tr>
<td></td>
<td>(3.3)</td>
<td>(2.3)</td>
</tr>
<tr>
<td>Low Performance</td>
<td>70%</td>
<td>68%</td>
</tr>
<tr>
<td></td>
<td>(2.8)</td>
<td>(1.5)</td>
</tr>
</tbody>
</table>

*Note.* The abbreviation “F” stands for fact, “C” stands for comprehension, “R” stands for replication, and “N” stands for novel. The standard error of the mean (SEM) is listed in the parentheses.
Figure 1. Mean quiz scores across both the low-passing and the standard practice condition.
Figure 2. Percentage of students who completed quizzes across both the low-passing and the standard practice condition.
Figure 3. Mean percentage of correct exam questions across the low-passing and the standard practice condition for both question type and novelty.
Figure 4. Mean percentage of correct exam questions across the low-passing and the standard practice condition for question novelty.
Figure 5. Mean percentage of correct exam questions across the low-passing and the standard practice condition by performance group.
References


