

Are Multiple-Choice Questions Suitable for a Final Examination in a STEM course?

Garrick Aden-Buie¹, Autar Kaw¹, Ali Yalcin¹, Ram Pendyala²

¹University of South Florida (USF), ²Arizona State University (ASU)

I. Background

How many times has a student come to see you to look at their graded final examination? The break is over, a new semester has started, and most students have moved on. There may be a conscientious student who wants to learn from the mistakes made in the examination or a hopeful student who is looking to improve their grade through re-grading. But mostly, a final examination serves to measure overall achievement of the learning objectives of the course and rarely as a learning instrument for the student.

Multiple-choice (MC) examinations are certainly more easily graded, but for a number of reasons instructors are reluctant to use MC-format examinations in computationally-intensive courses. Constructed response (CR) examinations, on the other hand, require significantly more time and effort to grade.

So we ask the question: Can a final examination of a hybrid format - multiple-choice test that allows constructed response for partial credit suitably measure the learning objectives of the course? Such a format could save grader effort and thereby increase time available for other learning objectives.

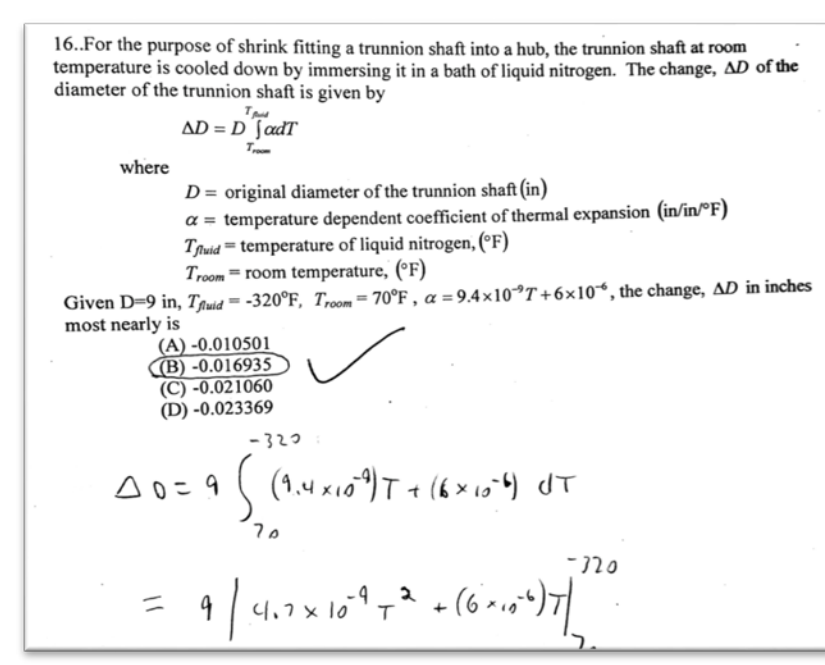
II. Objectives

We hypothesize that a hybrid format - multiple-choice question test that allows constructed response for **partial credit** (MC+PC) - can provide an equally appropriate measure of student mastery in a final examination when compared with other standard test formats: **constructed response** (CR) and **multiple-choice** (MC). If all examinations measure the learning objectives similarly, what are the relative savings in the grading effort?

III. Research Design

A final examination with 24 questions with identical stems was given to students in three consecutive semesters for a course in Numerical Methods. The three final exams differed only in terms of item format, as described below.

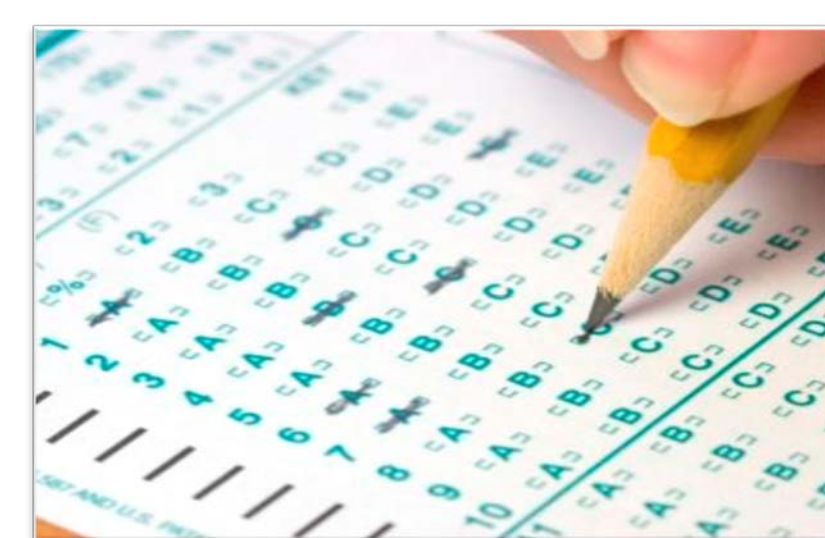
MC+PC Semester: For each question, students were asked to choose an answer from four choices as well as give a constructed response. If the choice was correct, the student received 4 points, otherwise the provided constructed response was graded on a scale of 0-3. Multiple-choice only scores were recorded for this semester as well (MC-PC Semester).



CR Semester: For each question, students were asked to give a constructed response. Each question was graded on a scale of 0-4.



MC Semester: For each question, students were asked to choose an answer from four choices. If the choice was correct, the student received 4 points; if the choice was wrong, the student received 0 points.



We collected demographic and performance data for each treatment: age, gender, transfer status, pre-requisite GPA, and average grade in semester tests of the course.

IV. Results/Evaluation

Table 1 presents student demographics in addition to the mean GPA in prerequisite courses and the mean of the in-course exam grades for each treatment. The in-course examination grades are a strong external measure of mastery of the topics in the course as they collectively cover all of the topics in the syllabus.

Treatment	N	Age			Gender			Transfer Status			Prerequisite GPA		Average In-Course Exam Grade
		Mean	Male	Female	Started at USF	Community College	Other	Mean	SD	Mean			
MC+PC	65	22.52	97%	3%	65%	27%	11%	3.22	0.52	77.19%			
CR	75	23.15	87%	13%	63%	45%	8%	3.04	0.54	74.57%			
MC	57	23.39	91%	9%	48%	42%	19%	2.98	0.53	75.03%			
Total/Mean	197	23.02	91%	9%	59%	38%	12%	3.08	0.53	75.60%			

Table 1. Student demographics and academic performance.

Multiple linear regression models were used to evaluate the three examination formats under two scoring methods, with partial credit (CR and MC+PC, $R^2_{adj} = 0.582$, $p < 0.001$) and with correct/wrong only (MC-PC and MC, $R^2_{adj} = 0.452$, $p < 0.001$), taking into account the student's age, gender, transfer status, pre-requisite course GPA and in-course exam grades.

For both scoring formats, the students' in-course examination grade average was a statistically significant predictor of performance on the final examination ($p < 0.001$), while performance in the prerequisite courses was statistically significant for the partial credit scoring method ($p = 0.007$) and nearly significant in the dichotomous scoring method ($p = 0.052$). Significant effects were not observed for the demographic factors, thus a strong bias was not demonstrated for age, gender or transfer status.

Student performance on the final examination is presented in Table 2 by unadjusted score and according to the final examination format. The MC-PC format represents the hybrid MC+PC with the partial credit removed.

To test for consistency in the ranking of students by the four final examination formats, student performance on the final examination was compared to prior performance on the three in-class examinations. A Spearman's rank test (Table 2) indicates statistically significant correlation between performance on in-class examinations and performance on the final examination for all of the final examination formats. This is a strong indicator that the final examination, in all of the studied formats, provides a good evaluation of the student's mastery of the subjects presented in the course.

A Spearman's rank correlation test was also applied to the final examination grades of the MC+PC semester with and without the partial credit option applied. The correlation coefficient is near unity ($cor = 0.968$), indicating that student ranking is largely unaffected by the partial credit option ($p < 0.001$).

Scoring	Format	N	Final Exam Score			Correlation with Rank by In-Course Exam Average	
			Mean	Median	SD	Spearman's Coefficient	p value
Partial Credit	CR	75	58.1	58	13.7	0.673	< 0.001
	MC+PC	65	69.6	71	13.3	0.619	< 0.001
Dichotomous	MC-PC	65	59.9	60	14.8	0.626	< 0.001
	MC	57	59.3	60	15.4	0.676	< 0.001

Table 2. Student performance on the final examination for the three formats and two scoring strategies, including Spearman's Rank Test for comparison of ranking by average in-course examination grade with score on the final examination.

In terms of the effort required to grade student responses, manual scoring was required for the CR and MC+PC treatments. Figure 1 shows the percentage of items requiring scoring for the two treatments. MC+PC grading required manual scoring for only 41% of all responses. As a conservative estimate, this is approximately a 30% reduction over the CR items requiring detailed grading.

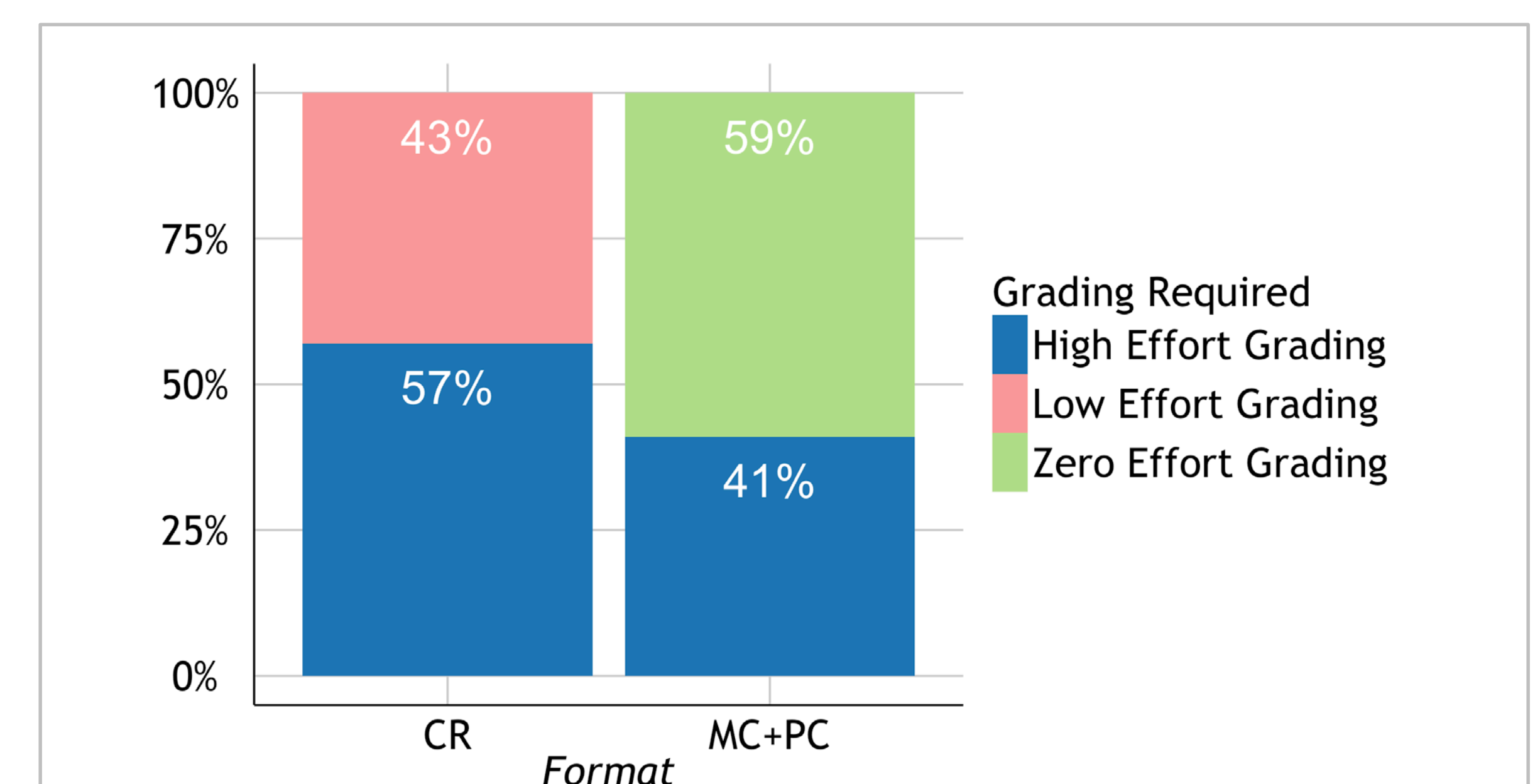


Figure 1. Administrative grading requirements. High effort grading required grading by hand. Low effort grading means correct answer determined by hand. Zero effort grading indicates correct answer verified by Scantron.

V. Conclusions

This study presents and compares the use of three examination formats - constructed response (CR), multiple-choice (MC) and a hybrid multiple-choice with partial credit, the last format under two scoring methods - partial credit (MC+PC) and dichotomous scoring (MC-PC).

- The combination of MC items with optional partial credit provides a middle ground between a CR- or MC-only examination. Grading demands decreased significantly for the MC+PC examination when compared to the CR examination, while the MC+PC examination demonstrated reliability equivalent to the CR-only examination.
- Performance on the final examination was found to be highly correlated with performance on the in-course examinations for each of the three treatments.
- Students' academic performance in prerequisite courses were significant predictors of performance on the final examination for both the partial credit and dichotomous scoring strategies; the effect of prerequisite course performance was slightly less than significant under dichotomous scoring.
- The hybrid MC+PC examination format provided a similar level of reliability when compared with the other formats under their respective scoring methods.

VI. Future Plans

We are adding one more treatment to the experiment. A final examination with 50-50 combination of multiple-choice and free-response questions is being tested currently. The MC questions are asked on the low levels of Bloom's taxonomy, while the free-response questions are asked on the high levels of Bloom's taxonomy. The MC questions are graded dichotomously while the free-response questions are graded using a rubric on a scale of 0-4. This combination still allows for quick grading while being able to ask higher-order thinking questions that may also be open-ended.