International Journal of Engineering Education Vol. 31, No. 1(A), pp. 72–82, 2015
Printed in Great Britain

0949-149X/91 $3.00+0.00

Comparison of Final Examination Formats in a Numerical Methods Course*

GARRICK ADEN-BUIE
Department of Industrial and Management Systems Engineering, College of Engineering, University of South Florida, 4202 E. Fowler Ave. ENB118, Tampa, FL, USA. E-mail: gadenbuie@mail.usf.edu

AUTAR KAW
Department of Mechanical Engineering, College of Engineering, University of South Florida, 4202 E. Fowler Ave. ENB118, Tampa, FL, USA. E-mail: kaw@usf.edu

ALI YALCIN
Department of Industrial and Management Systems Engineering, College of Engineering, University of South Florida, 4202 E. Fowler Ave. ENB118, Tampa, FL, USA. E-mail: ayalcin@usf.edu

With decreasing budgets for teaching assistants, large class sizes, and increased teaching loads, it is becoming ever more important to effectively utilize resources without sacrificing best practices of assessment. The objective of this study is to evaluate a hybrid multiple-choice final examination with optional partial credit (MC+PC) as a replacement for the same examination in constructed response (CR) or strict multiple-choice (MC) formats. In the hybrid MC+PC format, students were given multiple-choice options and were also allowed to submit constructed responses that would be graded for partial credit. The three examination formats were utilized once each in three offerings of a Numerical Methods course at the University of South Florida. Multiple linear regression and item analysis of student responses demonstrate that students approach the MC+PC format similarly to a CR exam, and the administrative requirements of the test were significantly reduced. This study finds the hybrid MC+PC format to be equally reliable and appropriate for a comprehensive final examination.

Keywords: final examination; examination formats; numerical methods

1. Introduction

As the discourse on educational strategy has shifted in recent decades from a focus on teaching to student-centered learning objectives, the role of student assessments has shifted from measurement of topic mastery to the "constructive alignment" of assessments with the learning process [1]. When examinations and other assessments undertaken during the progression of a course are constructively aligned with learning, they both measure student achievement and guide the learning process through structured formative feedback [2].

Comprehensive final examinations, in contrast, serve to measure overall achievement of the learning objectives of the course and, due in part to their timing in the U.S. semester-based course calendar, rarely serve to guide the learning process. In the combined experience of the authors as instructors of engineering curricula, less than 1% of students request to review the scored final examination. This indicates that a large majority of students do not perceive final examinations to be an opportunity for learning, but rather a straightforward measurement of their mastery of the skills acquired in the course.

While constructed response (CR) examinations expose the thought processes of individual students and thus facilitate constructive student-centered feedback, they are time- and resource-intensive to score [3]. Instructors who must balance teaching and research obligations or who strive to ensure the effective allocation of teaching support may reasonably question the efficiency of a constructed response final examination format. Ideally, a more efficient but equally effective grading method would save instructor resources without undermining the role of the assessment.

Multiple-choice (MC) examinations, in comparison to CR examinations, are often preferred by students [4] and are more easily and reliably administered by instructors. In a computation-intensive science, technology, engineering or math (STEM) course, this preference may be tempered by the general inability of a multiple-choice examination to differentiate between conceptual and procedural errors.

To overcome this limitation, this study presents a synthesis of MC and CR formats whereby students may opt to provide a written response to an MC item in addition to their selection of a multiple-choice option. If the student selects an incorrect item

* Accepted 25 September 2014.

option, the written response is scored and assigned partial credit. The above formats were evaluated by administering a comprehensive final examination to students in an undergraduate course in Numerical Methods in the three formats: constructed response, multiple-choice, and multiple-choice with partial credit.

The primary research question is to evaluate whether the multiple-choice with partial credit (MC+PC) examination format provides an equally reliable and appropriate evaluation of student achievement of learning objectives when compared with the CR and MC-only formats, in conjunction with reduced administration requirements.

2. Background

The question of choosing the most appropriate item format for student assessments is neither new nor definitely resolved and has been discussed since the appearance of MC tests in the early 1900s [5]. More specifically, a number of studies have evaluated the equivalence between, and advantages and drawbacks of, the CR and MC formats [3, 6, 7]. While it is acknowledged that a specific format may be more appropriate depending on the trait the examiner wishes to evaluate, Rodriguez summarized the consensus in the literature that, when carefully designed, both formats approach equivalency, particularly for qualitative and reading comprehension items. It is more essential, he advised, to define the measurement objectives of the test and design a test that "elicits the kind of behavior reflected in [that] definition" [5].

From an administrative perspective, CR examinations can be one to several orders of magnitude more costly to implement and score than MC examinations, especially as the size of the examinee population grows [3]. CR items are generally considered more reliable than MC items, as student guessing is minimized and more nuanced scoring is possible; however, maintaining validity and consistency requires strict maintenance and fair application of a grading rubric [1]. As a result, CR items require allocating students more time during the examination and increase the administrative demands in preparing for and scoring the examination and providing feedback to students. Scoring constructed response items requires graders with high-level domain knowledge and includes a certain degree of subjectivity that may inadvertently introduce variation or bias in students' scores.

From the student perspective, MC items are generally preferred, although at times for reasons counterproductive to learning goals [4]. MC items are perceived by students to be "easier," both to prepare for and during the examination. Students tend to believe that MC items are limited to testing basic knowledge and find comfort in the availability of options and the ability to guess if they are unsure of the correct answer [8]. Conversely, they tend to find CR items "fairer" in terms of demonstrating the depth of knowledge or skills being tested and also for the ability to achieve partial credit.

A number of strategies for assessing partial knowledge using MC questions have been developed with the goal of minimizing guessing or determining the state of knowledge of the student [9]. Alternatives to standard single-item-correct, dichotomous scoring MC methods include differential item and option weighting or new item or response methods [10]. Option weighting methods assign partial credit weighting to item options according to correctness, based on the judgment of experts, such as the instructor, or by empirical evidence from previous administrations of the test. Dressel & Schmid [11] proposed the multiple-correct MC item format in which items may have more than one correct option and students are instructed to select all correct options. Coombs *et al.* [12] introduced elimination testing, a response method in which students explicitly eliminate incorrect item options and mark their selection for the correct option. Probability testing [13] and confidence marking [11] respectively ask students to assign a probability of correctness to each option or a confidence in the correctness of the student's selection. The number of variants to each of these strategies is significant, each with trade-offs in transparency, ease of communication, and time requirements for test writing and grading.

Recent studies have evaluated the use of MC examination formats in science and engineering courses. Scott et al. [14] provided a detailed analysis of the conversion of examination format in a large-scale introductory physics from CR to MC and conclude that MC examinations "[fulfill] their primary function of assessing student understanding and assigning the appropriate grade" while reducing student appeals and grading difficulties. Chan & Kennedy [15] reviewed stem-equivalent CR and MC items in two randomly assigned examination formats in a college-level economics course. Results of a comparison of MC and CR items showed mixed effects resulting from the inclusion of item options in which particular options may help students in "articulating the answer in unequivocal fashion" while other item options may cause students to "worry about erroneous factors that they otherwise would not have taken into consideration". Adair & Jaeger [16] introduced a computerized test format and scoring method in which students label each multiple-choice options as "correct", "wrong", or "not sure" thereby revealing partial knowledge of

the examined concepts by which students are graded accordingly. Finally, Stanger-Hall [17] evaluated the use of mixed CR and MC examinations throughout an introductory biology course and found that the inclusion of CR items improved critical thinking skills and studying strategies used by students.

3. Experimental design

Numerical Methods is an undergraduate, junior-level course that follows the mathematical course sequence of Calculus I, II, and III and Ordinary Differential Equations [18]. Three consecutive offerings of the course were included in this study, namely the Spring 2012, Spring 2013 and Summer 2013 semesters. Following local IRB procedures, students were invited to participate in the study via announcements made in the course.

For each participating student, the student's age, gender and performance in the prerequisite courses were recorded. Additionally, as students in the course are typically further into their academic careers, students were identified by transfer status: *first time in college* (FTIC), transfer students from a *community college* (CC) with a completed Associate of the Arts degree, or *other* (OT) which includes students transferring from another institution without a completed degree. All of the above data were collected from official institutional records.

Student achievement in the course was assessed through a combination of homework assignments [19], class activities and examinations, including the final comprehensive examination. The same topics were covered in each of the three semesters, drawn from eight chapters in a well-known Numerical Methods textbook [20]. Three examinations were administered over the course of each semester and together covered all of the material presented in the course. The in-course examinations consisted primarily of constructed response items with a few multiple-choice items.

The final examination contained three questions per chapter covered in the course; two of the three questions were based on the lower levels of Bloom's taxonomy [21]—*knowledge*, *comprehension*, and *application*—while the third was based on the higher levels—*analysis*, *synthesis*, and *evaluation*. The in-course examinations were similarly designed to measure learning at various levels. The three formats of the final examination were identical with respect to item stems and differed only in terms of item format and grading policy. Development of the examination relied fully on the 2nd author's 24 years of experience as an instructor of Numerical Methods, throughout which the content of the examination has stabilized and been proven valid. Three items were naturally multiple-choice and the format of these questions was not varied across the three semesters in this study. Each of the 24 questions was assigned a maximum score of 4 points, with the cumulative examination score being the sum of points received plus 4 additional points for a total maximum score of 100 points.

Each semester received one of the three final examination formats, administered as follows. The CR final examination was administered in the *Spring 2013* semester. With the exception of the three common multiple-choice questions, item stems were presented without options and students were asked to provide an answer and show all related work. Students were given 120 minutes to complete the examination. The final examinations were graded according to a rubric designed by the instructor and applied by a graduate teaching assistant, who worked with the instructor to ensure the rubric was followed closely. Correct final answers received full credit of 4 points; incorrect answers received as partial credit the 4 points reduced by 1 point for each procedural error (e.g., a sign or computational error) and 2 points for each conceptual error (e.g., correct application of less appropriate method).

The MC+PC final examination was administered in the *Spring 2012* semester. Item stems and four options were presented to students, who were instructed to select the correct option. Following the advice in Haladyna [22], item options were carefully constructed in such a way that distractors were non-obvious and required that the student understand the material or complete a calculation. Correct answers received full credit, while incorrect answers were then reviewed by a graduate teaching assistant following the same rubric and under the same guidance as the CR examination grader. Thus, if a student selected an incorrect MC option and elected to show their work, their answer was treated as if it was a CR item and the student received between 0 and 3 points (in integer increments). Students in this semester were again given 120 minutes to complete the examination.

The MC final examination was given in the *Summer 2013* semester. Item stems and options were identical to those presented to students on the MC+PC final examination. This treatment was differentiated by the use of the conventional MC format correct/incorrect grading style. Thus students received either 0 or 4 points with no opportunity for partial credit. Because students were not required to organize or structure their responses, students in this semester were given 90 minutes to complete the examination. In all semesters, only a few students required the entire allocated time.

Table 1. Student participation by semester

Treatment	N	Opted In	Opted Out	Incomplete	Participation (%)
Spring 2012	74	65	9	0	87.8
Spring 2013	83	75	8	0	90.4
Summer 2013	63	59	4	2	90.5

Table 2. Total number of students, gender and transfer status by semester

		Gender		Transfer Status		
Semester	Total	Male	Female	FTIC	CC	Other
Spring 2012	65	63	2	41	17	7
Spring 2013	75	65	10	41	29	5
Summer 2013	57	52	5	25	22	10
Total	197	180	17	107	68	22

4. Results

4.1 Student demographics and academic preparation

As seen in Table 1, participation was approximately 90% in each of the three semesters, with $N = 199$ total participants. Two students were excluded from the study due to incomplete prerequisite grade records. The number of students, age, transfer status and mean prerequisite GPA (PGPA) are shown in Tables 2 and 3.

Ideally, the composition of each class should be equal, both in terms of the origin of the student and their performance in the prerequisite courses. A Pearson's Chi-squared test was applied to determine if each of the classes contained similar students. Significance for this and all other tests presented in this study was set at a Type 1 error rate of 5%. The results of this analysis indicate that the composition of students in each class is not significantly different with respect to transfer status ($\chi^2 = 7.479$, $p = 0.113$). However, a two-sided Student's t-test comparing PGPA between semesters (Table 4) suggests that student performance in the prerequisite courses differs between Spr '12 vs Spr '13 and Spr '12 vs Sum '13, indicating that students' prior academic performance at the start of the course in the Spring 2012 semester differs from the other semesters.

4.2 Student performance on final and in-course examinations

Performance of the students in the course up to the final examination is measured by averaging the in-course examination grades. Homework grades were excluded as they are designed to encourage student participation, while examination grades are a stronger measure of mastery of the topics studied. Three examinations were given during the semester and collectively cover all of the topics in the syllabus. Thus, a student's performance on the in-course examinations can be directly compared to their performance on the final examination. Averaged in-course examination grades are moderately correlated with students' previous academic performance as measured by MPGPA, with an average correlation coefficient of 0.486. The mean, median and standard deviation of averaged in-course examination grades are presented in Table 5.

Student performance on the final examination is presented in Table 6 by raw score and according to final examination format. The CR and MC+PC examinations are scored by the equivalent and directly comparable partial credit method discussed in Section 3, with scores assigned in integer values from 0 to 4. Similarly, the dichotomously scored MC examination can be compared with the MC+PC format by removing the partial credit option (hereafter denoted MC-PC) and using the dichotomous 0 or 4-point scoring method.

Table 3. Mean age and PGPA of students by semester

	Age	PGPA	
Semester	Mean	Mean	SD
Spring 2012	22.52	3.22	0.52
Spring 2013	23.15	3.04	0.54
Summer 2013	23.39	2.98	0.53
Mean	23.02	3.08	0.53

Table 4. Student's t-test on PGPA between semesters

	Spr '12 vs Spr '13	Spr '12 vs Sum '13	Spr '13 vs Sum '13
t statistic	2.033	2.509	0.597
p value	0.044	0.013	0.552

Table 5. Averaged in-course examination grades by semester

Treatment	Mean	Median	SD
Spring 2012	77.19	78.75	10.17
Spring 2013	74.57	76.00	12.30
Summer 2013	75.03	75.67	12.49
Mean	75.60	76.81	11.65

Table 6. Final examination raw score by examination format

Scoring	Format	Semester	N	Mean	Median	SD	Min	Max
Partial Credit	CR	Spring 2013	75	58.1	58	13.7	27	86
	MC+PC	Spring 2012	65	69.6	71	13.3	22	94
Dichotomous	MC-PC	Spring 2012	65	59.9	60	14.8	20	92
	MC	Summer 2013	57	59.3	60	15.4	24	92

Mean final examination scores were significantly higher for the MC+PC students than for CR students under partial credit scoring, $t(136) = -5.03$, $p < 0.001$. A significant effect was not observed in the dichotomous scoring scenario between the MC-PC and MC formats, $t(117) = -0.23$, $p = 0.816$. The percentage of each point level awarded out of the total number of items graded in each examination format is presented in Fig. 1, where it can be seen that MC+PC were more likely to receive full credit.

Interestingly, students taking the MC+PC final examination were significantly more likely to choose not to select a multiple-choice option than the students taking the MC final examination. Among all of the responses of the MC students, only four responses were blank (where no option was selected). In contrast, in the MC+PC group, for each item an average of 16% of the students chose not to select a multiple-choice option despite the fact that absolutely no penalty was imposed for an incorrect selection. The implication is that, while multiple-choice tests are often criticized for encouraging guessing, when taking the MC+PC format examination students did not feel compelled to guess when they were not confident in their answer.

4.3 Validity of the final examination formats

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 in-course examinations. A Spearman's rank test indicates statistically significant strong correlation between performance on in-course 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 consistent evaluation of the student's mastery of the subjects presented in the course.

To ensure that the addition of the partial credit option does not affect the overall ranking of students in the course, a Spearman's rank correlation

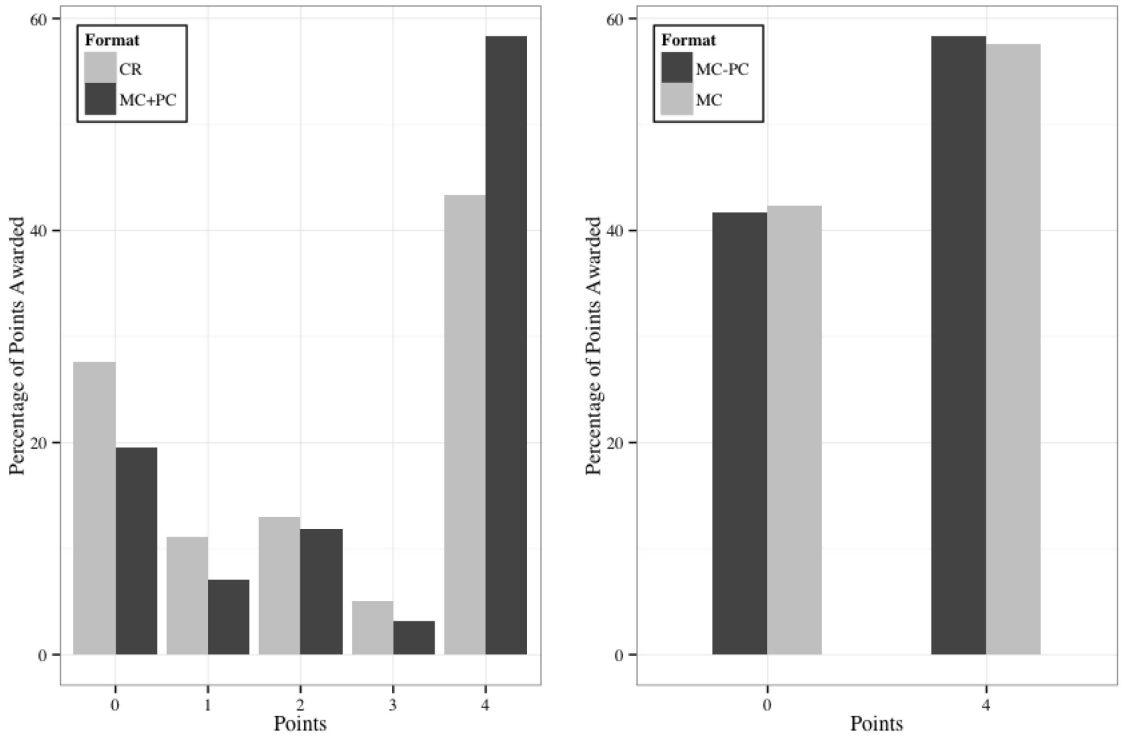

Fig. 1. Percentage of points awarded to students by format.

Table 7. Spearman's coefficient of correlation between the students' averaged in-course examination grades and final examination grade

Scoring	Semester	Grading Policy	Spearman's Coefficient	*p* value
Partial Credit	Spring 2013	CR	0.673	< 0.001
	Spring 2012	MC+PC	0.619	< 0.001
Dichotomous	Spring 2012	MC-PC	0.626	< 0.001
	Summer 2013	MC	0.676	< 0.001

Table 8. Cronbach's alpha for each final examination format

Scoring	Format	Semester	Cronbach's Alpha
Partial Credit	CR	Spring 2013	0.746
	MC+PC	Spring 2012	0.732
Dichotomous	MC-PC	Spring 2012	0.675
	MC	Summer 2013	0.682

test was also applied to the final examination grades of the Spring 2012 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 ($p < 0.001$).

4.4 Reliability of the final examination formats

The term *reliability* refers to the ability of a test to consistently assess or measure the same underlying ability or concept, insofar as in a fully reliable test the only source of measurement error is random error. Cronbach's coefficient alpha [23] is the most popular metric for evaluating reliability, and is considered a measurement of internal consistency, or the level of inter-item correlation within a test administered to a single group. The coefficient alpha estimation of reliability for each of the examination formats and scoring methods is shown in Table 8. For both the CR and MC+PC examination formats, alpha is near 0.74, while the dichotomously scored MC and MC-PC examination formats demonstrated reliability near 0.68.

While higher reliability is preferred, both scoring methods achieve adequate reliability for mastery-type, low-stakes tests used in conjunction with other grading and scoring methods, where a reliability coefficient of 0.60 or greater is considered acceptable [24, 25]. The reliability of a test can be increased by adding more items relevant to the test subject, and the new reliability of the test can be predicted by the Spearman-Brown prediction formula [26] according to the observed reliability of the test with the current number of items. Using this formula, the examinations under the dichotomous scoring method would require 8 additional items to be equivalent to the reliability of the partial credit final examinations.

4.5 Evaluation of the examination formats by multiple linear regression analysis

Multiple linear regression models were used to evaluate the three examination formats within the context of the two scoring methods, taking into account the student's age, gender, transfer status and academic performance prior to the final examination. Estimated coefficients and p values for the partial credit and dichotomous scoring methods are presented in Table 9. In both cases, student profile information and the examination format explained a significant portion of variance in final examination score: *partial credit*, $R^2_{adj} = 0.582$, $F(7, 132) = 28.682$, $p < 0.001$; *dichotomous*, $R^2_{adj} = 0.452$, $F(7, 114) = 15.249$, $p < 0.001$.

For both scoring formats, the students' in-course examination grade average is a statistically significant predictor of performance on the final examination, while performance in the prerequisite courses is statistically significant for the partial credit scoring method and nearly significant in the dichotomous scoring method. Under partial credit scoring, a significant effect was observed for the format of the examination, where the multiple-choice format increases final examination grades by approximately 9 points. Significant effects were not

Table 9. Results of multiple linear regression analysis comparing CR and MC+PC examination formats under partial credit scoring

	Partial Credit Scoring			Dichotomous Scoring		
Factor	**Estimate**	**SE**	**p**	**Estimate**	**SE**	**p**
(Intercept)	–0.82	8.60	0.926	–7.50	10.16	0.462
Format: MC+PC	8.90	1.66	< 0.001	2.50	2.14	0.244
Average In-Course Exam Grade	0.67	0.08	< 0.001	0.79	0.10	< 0.001
PGPA	4.82	1.77	0.007	4.43	2.26	0.053
Age	–0.22	0.22	0.328	–0.32	0.30	0.286
Gender: Female	2.78	2.92	0.343	3.68	4.48	0.413
Transfer: CC	–2.20	2.04	0.283	–1.23	2.61	0.638
Transfer: Other	–4.24	3.02	0.163	–2.46	3.29	0.456

observed for the remaining factors, thus a strong bias was not demonstrated for age, gender or transfer status. However, the negative coefficients of age and transfer status indicate that older and non-FTIC students tend to underperform when compared to FTIC and younger students.

4.6 Grading burden

In terms of grader effort, 58% of MC+PC items were correctly answered and required no additional review after the MC option selection was scored, leaving 41% to be graded by hand (1% of the items were unanswered). Students in the MC+PC group were motivated to approach the questions as if they were constructed response and 81% of the correctly answered questions included work. In comparison, while all of the items in the CR section had to be manually scored, 43% of the total items to be graded was answered correctly and thus required minimal scoring effort. The remaining 57% of the CR items had to be graded by hand and required more of the grader's time. Overall, a conservative estimate of the reduction in high-effort grading required for the MC+PC format was 28%.

4.7 Item analysis

4.7.1 Item difficulty index

For each examination format, the item difficulty index was calculated as the proportion of students who correctly answered an item among the total number of students. Naturally, this value is a positive number between 0 and 1, with universally easy or difficult questions indicated by an item difficulty of 1 or 0, respectively. It is generally recommended that, for norm-referenced examinations, the ideal average item difficulty index is halfway between chance and perfect scores—or 0.625 for four-option multiple-choice items [27]. Ideally, for mastery-type examinations, the average difficulty index should be higher such that a larger percentage of the examinees are correctly answering each item.

Mean item difficulty index and standard deviation for the four scoring strategies are presented in Table 10. While the constructed response and strict multiple-choice scoring strategies were on average equally difficult, the combination of the multiple-choice format with partial credit scoring decreased the difficulty of the examination as seen in the increased average difficulty index. Given that the examination under consideration is a cumulative final course evaluation, the average item difficulty index demonstrated by the MC+PC format is more suited to the objectives of the examination.

4.7.2 Item discrimination index

The ability of an item to discriminate between high and low achieving students was measured by the point biserial correlation of item score to total score on the examination [28]. The item discrimination index ranges from –1.0 to +1.0, where a positive discrimination index indicates that students who performed well on the test tended to correctly answer the item. Negative discrimination suggests that lower achieving students fared better on the item and is not desirable. For norm-referenced tests, a high discrimination index is desired to differentiate student performance within the class, while a lower discrimination index is expected for mastery-type tests [27].

Average item discrimination of each of the scoring formats is presented in Table 10. On average the partial credit scoring strategies are slightly more discriminating than the multiple-choice only formats.

4.7.3 Item characteristic curves

Item characteristic curves (ICC) provide further information on student performance and finer detail as to performance on an item across a range of student achievement levels [28]. Students were divided by quartiles based on their overall examination score within their course cohort. For each item and examination format, the ICC was generated by plotting average item score on the vertical axis by overall exam performance by quartile on the horizontal axis. Figure 2 shows the ICCs for the 24 final examination questions for each of the four scoring formats. A consistent and gradual slope is desired, indicating that more higher-achieving students correctly answer the item, as is seen in many of the questions. A flat ICC is less desirable and occurs in situations of higher item difficulty or lower item discrimination. Figure 2 suggests that the questions are on average well-balanced across all of the

Table 10. Item difficulty and discrimination indices for the four exam and scoring formats

		Item Difficulty		Item Discrimination	
Scoring	**Format**	**Mean**	**SD**	**Mean**	**SD**
Partial Credit	CR	0.564	0.221	0.299	0.174
	MC+PC	0.683	0.183	0.281	0.153
Dichotomous	MC-PC	0.583	0.214	0.227	0.141
	MC	0.576	0.189	0.234	0.122

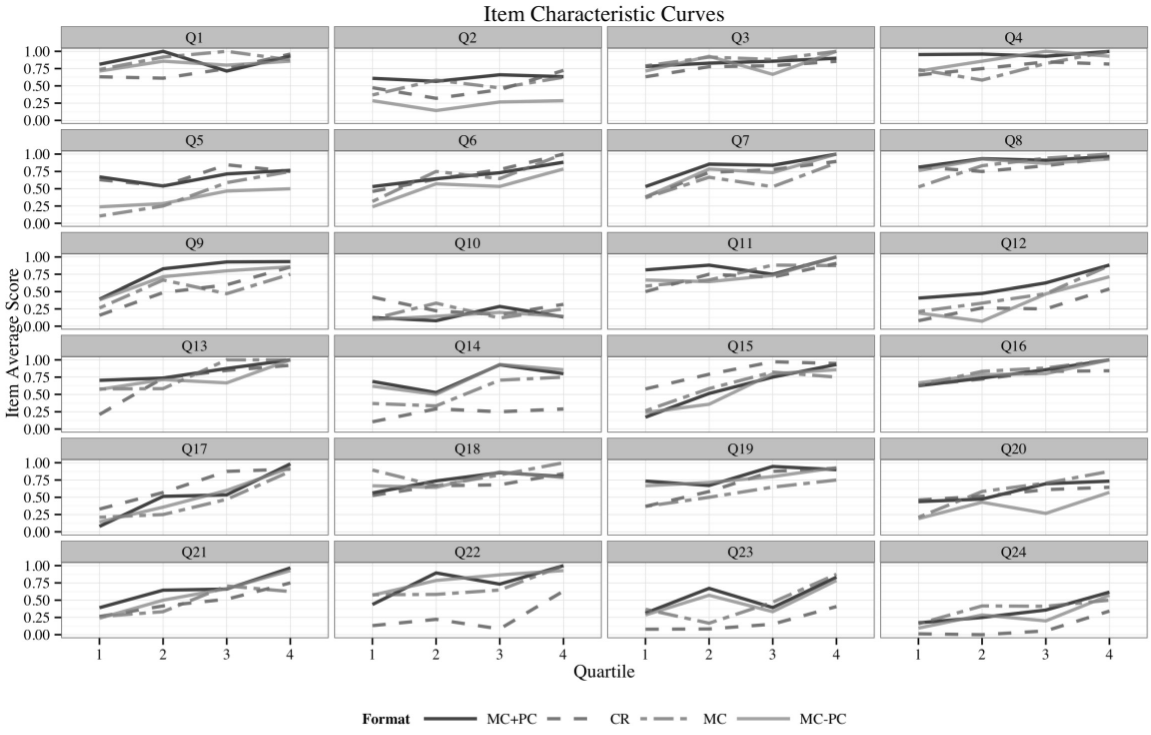

Fig. 2. Item characteristic curves: average item correctness by class-wise quartiles of the full final examination score for each item and format.

scoring formats as the curves are similar for all formats with a few exceptions.

4.7.4 Example questions

To illustrate the difference in student responses to questions in the three formats, two questions from the final examination have been reproduced in Box 1. The only difference between the items across the formats is the addition of the multiple-choice options, which were the same for the MC and MC+PC formats. The item stems were constructed carefully so that no additional information was presented in the multiple-choice options that were not already available in the item stem, other than the limited options.

In Question 14, students were asked to suggest the best location for the placement of two velocity sensors in a pipe of radius 2 meters such that flow rate would be most accurately observed from the sensors. In this question, the item difficulty index across the four scoring strategies studied was 0.230 (CR), 0.708 (MC+PC), 0.708 (MC-PC), and 0.526 (MC). The item discrimination index for this question did not demonstrate such large differences; all four item discrimination indices were near the average of the four (M = 0.150, SD = 0.042). In this case, it was determined that the options forced students to consider the physical limits imposed in the problem in terms of the placement of the probes. Students without the options present were more likely to suggest infeasible probe placement (e.g. outside of the pipe) or to provide a location for a single probe rather than two (despite the clear indication in the item stem that two answers were expected). The low difficulty index in the CR format and the absence of any difference between MC+PC and MC-PC difficulty indices show the influence of providing choices to the students. Largely, students either knew or did not know the correct answer.

In Question 15, the item options include answers derived from common mistakes. The item difficulty index was higher for the CR student group than for the students presented with multiple-choice options—0.817 (CR), 0.569 (MC+PC), 0.523 (MC-PC), and 0.561 (MC). Average item discrimination for the question is 0.427 (SD = 0.059). This question has a higher item difficulty index for the CR format because while students tended to begin to approach the question correctly and thereby obtain reasonable partial credit, they often faltered at the end when coming up with the final formula for the question. In the case of MC+PC and MC, the item difficulty index is lower in part because students may have been overly confident in their answer upon finding a multiple-choice option that matched their calculations. The item discrimination index is more than acceptable on this question.

Question 14

You are asked to estimate the water flow rate in a pipe of radius 2 meters at a remote are location with a harsh environment. You already know that velocity varies along the radial location, but you do not know how it varies. The flow rate, Q, is given by $Q = \int_0^2 2\pi r V dr$. To save money, you are allowed to put only two velocity probes (these probes measure velocity and send the data to the central office in New York, NY via satellite) in the pipe. Radial location, r, is measure from the center of the pipe, that is $r = 0$ meters is the center of the pipe and $r = 2$ meters is the pipe radius. The radial locations in meters you would suggest for the two velocity probes for the most accurate calculation of the flow rate would most nearly be at $r =$ _____ and $r =$ _____ .

(a) 0,2
(b) 1,2
(c) 0,1
(d) 0.42, 1.58

Question 15

The force vs. displacement data for a linear spring is given below. F is the force in Newtons and x is the displacement in meters. Assume displacement data is known more accurately.

Displacement, x (m)	10	15	20
Force, F (N)	100	200	500

If the data is regressed to $F = kx$, the value of k by minimizing the sum of the square of the residuals is most nearly _____ N/m.

(a) 16.11 N/m
(b) 17.78 N/m
(c) 19.31 N/m
(d) 40.00 N/m

Box 1: Example questions from the final examination with multiple-choice options included.

5. Discussion and limitations

This article presents and compares the performance of 197 students on the final examination of an undergraduate course on Numerical Methods, using three examination formats—constructed response, multiple-choice and a hybrid multiple-choice with partial credit—and under two scoring methods—partial credit and dichotomous scoring. Performance on the final examination was found to be highly correlated with performance on the in-course examinations for each of the three student groups.

Similarly, students' academic performance in prerequisite courses and during the numerical methods course were significant predictors of performance on the final examination for both the partial credit and dichotomous scoring strategies, although the effect of prerequisite course performance was slightly less than significant under dichotomous scoring. While students performed better when presented with multiple choices with the opportunity for partial credit than when asked to independently construct their response, no significant effects were observed with respect to the student profile.

Additionally, the hybrid MC+PC format was found to provide a similar level of reliability when compared with the other formats under their respective scoring methods. The observed average item difficulty and discrimination indices further indicate that the MC+PC format is an appropriate examination format for a comprehensive final evaluation. Furthermore, item analysis helps improve the design of the exam by providing important information on item performance when format changes are under consideration.

The results of this study are limited by the inclusion of a single numerical methods course rather than a broad selection of STEM courses. Similarly, the context of the study—an upper-level undergraduate course at an American university—limits the broader applicability of the results to other university systems with different course structures and schedules.

However, the results presented demonstrate that, within these limitations, the combination of MC items with optional partial credit provides an ideal

middle ground between a CR- or MC-only examination. Grading demands decreased significantly for the MC+PC examination when compared to the CR examination format, while the MC+PC demonstrated reliability equivalent to the CR-only format. The pressure to answer the multiple-choice portion of the item was reduced by the partial credit option, thus reducing guessing. Whereas the MC-only students left only 4 items blank, on each question an average of 16% of the MC+PC students did not select a multiple-choice option, relying instead on their written response for credit. Thus, the results suggest that the MC+PC examination format may provide a desirable balance between the high level of detail provided in student responses to a CR format examination and the reduced test burden for both instructors and students when using the MC format.

These findings are in line with similar studies involving the use of MC items in STEM courses. Scott, Stelzer, and Gladding [14] presented a successful conversion of all exams in an introductory physics course to multiple-choice format, but cautioned that instructors need to ensure that they still see and grade student work. The present study focuses on only the final examination, however the MC+PC format provides a mechanism for this type of feedback. Chan and Kennedy [15] observed similar patterns to those discussed in Section 4.7.4 in comparing MC items to CR equivalents in an economics course. For some items MC options helped students formulate the correct answer, while for other items the options seemed to mislead students. In the collective view of the literature (see Section 2) and the present study, MC tests can efficiently assess student achievement, in particular when used within a range of other learning and teaching strategies that encourage higher-order thinking skills and provide students and instructors alike with constructive feedback.

6. Conclusions

The introduced multiple-choice with constructed response partial credit format is a novel and simple balance between multiple-choice or constructed response only formats. Examination reliability and average item difficulty and discrimination indices indicate that this format is appropriate for a final examination. Furthermore, in comparison to traditional multiple-choice questions, the incentive to guess is reduced under the hybrid multiple-choice format. Similarly, overall grading requirements are also reduced when compared to the constructed response format.

Acknowledgements—This material is based upon work supported partially by the National Science Foundation under Grant Numbers 0717624 and 1322586, and the Research for Undergraduates Program in the University of South Florida (USF) College of Engineering. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. With permission from ASEE, portions of a 2014 ASEE conference proceedings paper written by the authors are used in this article.

References

1. J. Biggs and C. Tang, *Teaching for Quality Learning at University*, McGraw-Hill International, 2011.
2. C. Rust, The impact of assessment on student learning: How can the research literature practically help to inform the development of departmental assessment strategies and learner-centred assessment practices?, *Active Learning in Higher Education*, **3**(2), 2002, pp. 145–158.
3. H. Wainer and D. Thissen, Combining multiple-choice and constructed-response test scores: Toward a marxist theory of test construction, *Applied Measurement in Education*, **6**(2), 1993, pp. 103–118.
4. K. Struyven, F. Dochy and S. Janssens, Students' perceptions about evaluation and assessment in higher education: a review, *Assessment & Evaluation in Higher Education*, **30**(4), 2005, pp. 325–341.
5. M. C. Rodriguez, Choosing an item format, in G. Tindal and T. M. Haladyna *Large-Scale Assessment Programs for All Students: Validity, Technical Adequacy, and Implementation*, Lawrence Erlbaum Associates, 2002, pp. 213–231.
6. R. E. Bennett, D. A. Rock and M. Wang, Equivalence of free-response and multiple-choice items, *Journal of Educational Measurement*, **28**(1), 1991, pp. 77–92.
7. G. R. Hancock, Cognitive complexity and the comparability of multiple-choice and constructed-response test formats, *The Journal of Experimental Education*, **62**(2), 1994, pp. 143–157.
8. M. Zeidner, Essay versus multiple-choice type classroom exams: The student's perspective, *The Journal of Educational Research*, **80**(6), 1987, pp. 352–358.
9. A. Ben-Simon, D. V. Budescu and B. Nevo, A comparative study of measures of partial knowledge in multiple-choice tests, *Applied Psychological Measurement*, **21**(1), 1997, pp. 65–88.
10. F. Lord, M. Novick and A. Birnbaum, *Statistical Theories of Mental Test Scores.* Addison-Wesley Pub. Co., 1968.
11. P. L. Dressel and J. Schmid, Some modifications of the multiple-choice item, *Educational and Psychological Measurement*, **13**(4), 1953, pp. 574–595.
12. C. H. Coombs, J. E. Milholland and F. B. Womer, The assessment of partial knowledge, *Educational and Psychological Measurement*, **16**(1), 1956, pp. 13–37.
13. T. S. Wallsten, D. V. Budescu, R. Zwick and S. M. Kemp, Preferences and reasons for communicating probabilistic information in verbal or numerical terms, *Bulletin of the Psychonomic Society*, 1993.
14. M. Scott, T. Stelzer and G. Gladding, Evaluating multiple-choice exams in large introductory physics courses, *Physical Review Special Topics—Physics Education Research*, **2**(2), 2006, p. 020102.
15. N. Chan and P. E. Kennedy, Are multiple-choice exams easier for economics students? A comparison of multiple-choice and "equivalent" exam questions, *Southern Economic Journal*, **68**(4), 2002, pp. 957–971.
16. D. Adair and M. Jaeger, A scoring method based on simple probability theory that considers partial knowledge and omission of answers in multiple-choice testing, *International Journal of Engineering Education*, **29**(4), 2013, pp. 974–985.
17. K. F. Stanger-Hall, Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes, *CBE Life Sciences Education*, **11**(3), 2012, pp. 294–306.
18. A. Kaw, A. Yalcin, J. Eison, C. Owens, G. Besterfield, G. Lee-Thomas, D. Nguyen, M. Hess and R. Pendyala, A holistic view on history, development, assessment, and future of an open courseware in numerical methods, *ASEE Computers in Education Journal*, **3**(4), 2012, pp. 57–71.

19. G. Thomas, A. K. Kaw and A. Yalcin, Using online endless quizzes as graded homework, Proceedings of 2011 ASEE Conference & Exposition, 2011.
20. A. K. Kaw, E. Kalu and N. Duc, *Numerical Methods with Applications*. Lulu.com, 2010.
21. B. S. Bloom, *Taxonomy of Educational Objectives. Handbook I: Cognitive Domain*. David McKay Company, Inc., 1956.
22. T. M. Haladyna, S. M. Downing, and M. C. Rodriguez, A review of multiple-choice item-writing guidelines for classroom assessment, *Applied Measurement in Education*, **15**(3), 2002, pp. 309–333.
23. L. J. Cronbach, Coefficient alpha and the internal structure of tests, *Psychometrika*, **16**(3), 1951, pp. 297–334.
24. L. M. Rudner and W. D. Schafer, Reliability eRIC digest, *ERIC clearinghouse on assessment and evaluation*. 2001.
25. K. Allen, T. Reed-Rhoads, R. A. Terry, T. J. Murphy and A. D. Stone, Coefficient alpha: an engineer's interpretation of test reliability, *Journal of Engineering Education*, **97**(1), 2008, pp. 87–94.
26. C. Spearman, Correlation calculated from faulty data, *British Journal of Psychology*, **3**(3), 1910, pp. 271–295.
27. T. L. Flateby, A guide for writing and improving achievement tests, University of South Florida, Office of Evaluation; Testing, 813, 1996.
28. R. M. Furr and V. R. Bacharach, *Psychometrics: An Introduction*. SAGE Publications, 2013, p. 472.

Garrick Aden-Buie is a doctoral student in the Department of Industrial and Management Systems Engineering at the University of South Florida. He received a B.S. in Applied Mathematics and a B.A. in Spanish from Lehigh University. His research interests include data mining and predictive modeling for healthcare decision support systems.

Autar Kaw is a Professor of Mechanical Engineering and Jerome Krivanek Distinguished Teacher at the University of South Florida. With major funding from National Science Foundation, he is the principal contributor in developing award-winning online resources for an undergraduate course in Numerical Methods. His current research interests include engineering education research, rehabilitation engineering, body-armor design and bascule bridge design. He is a Fellow of ASME and the recipient of the 2012 Council for Advancement and Support of Education (CASE) and the Carnegie Foundation for the Advancement of Teaching (CFAT) Professor of the Year Award.

Ali Yalcin is an Associate Professor of Industrial and Management Systems Engineering Department at the University of South Florida, and an Associate Faculty member of the Center for Urban Transportation Research. His research interests include modeling, analysis and control of discrete event systems, production planning and control, industrial information systems, data analysis and knowledge discovery, and engineering education research. He has taught courses in the areas of systems modeling and analysis, information systems design, production planning, facilities design, and systems simulation. He also co-authored the 2006 Joint Publishers Book-of-the-Year textbook, Design of Industrial Information Systems, Elsevier.