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

Mr. Garrick A Aden-Buie, University of South Florida

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 predictive modeling for healthcare decision support and sustainable dynamic systems.

Prof. Autar Kaw, University of South Florida

Dr. Autar Kaw is a professor of mechanical engineering at the University of South Florida. He was named the 2012 U.S. Professor the Year (Doctoral Institutions) by the Carnegie Foundation for the Advancement of Teaching and the Council for Advancement and Support of Education. The U.S. Professor of the Year award is the highest honor in the nation for undergraduate teaching. He received his BE Honors degree in Mechanical Engineering from Birla Institute of Technology and Science (BITS), India in 1981, and his degrees of Ph.D. in 1987 and M.S. in 1984, both in Engineering Mechanics from Clemson University, SC. He joined University of South Florida in 1987.

Professor Kaw's main scholarly interests are in engineering education research, open courseware development, bascule bridge design, fracture mechanics, composite materials, and the state and future of higher education. His research has been funded by National Science Foundation, Air Force Office of Scientific Research, Florida Department of Transportation, and Wright Patterson Air Force Base.

Professor Kaw has written several textbooks on subjects such as composite materials, numerical methods, computer programming and engineering licensure examination.

Since 2002, under Professor Kaw's leadership, he and his colleagues from around the nation have developed, implemented, refined and assessed online resources for an open courseware in Numerical Methods (http://nm.MathForCollege.com). The courseware gets more than a million page views per year. This is in addition to 900,000 views of the YouTube lectures and 150,000 annual visitors to the "numerical methods guy" blog.

Professor Kaw's opinion editorials have appeared in the Tampa Bay Times and Tampa Tribune, and his work has been covered/cited in Chronicle of Higher Education, Inside Higher Education, Congressional Record, ASEE Prism, Campus Technology, Florida Trend Magazine, WUSF, Bay News 9, NSF Discoveries, Voice of America, Times of India, and Indian Express. He has written more than 80 refereed technical papers.

Professor Kaw is a Fellow of the American Society of Mechanical Engineers (ASME) and a member of the American Society of Engineering Education (ASEE). He has also been a Maintenance Engineer (1982) for Ford-Escorts Tractors, India, and a Summer Faculty Fellow (1992) and Visiting Scientist (1991) at Wright Patterson Air Force Base.

Dr. Ali Yalcin, University of South Florida

Dr. Ali Yalcin received his B.S., M.S., and Ph.D. degrees in Industrial and Systems Engineering from Rutgers University, New Brunswick New Jersey in 1995, 1997 and 2000. He is currently an Associate Professor at the University of South Florida, Industrial and Management Systems Engineering Department, and an Associate Faculty member of the Center for Urban Transportation Research. His research interests include systems modeling, analysis and control, data analysis and decision support in healthcare, information systems and engineering education research. His work has been funded by federal organizations including National Science Foundation and Army Office of Research and medical device manufacturing industry. He has taught courses in the areas of systems modeling and performance analysis, information systems design, production planning, facilities design, and systems simulation. He co-authored the 2006 Joint Publishers Book-of-the-Year textbook, Design of Industrial Information Systems, Elsevier.

Prof. Ram Pendyala, Arizona State University

Ram M. Pendyala is a Professor of Transportation Systems in the School of Sustainable Engineering and the Built Environment at Arizona State University. His expertise lies in the study of human activity-travel behavior, sustainable mobility strategies, public transportation systems, and the land use, travel, energy, and air quality impacts of a wide range of transportation policies and technologies. Dr. Pendyala has conducted more than $5 million in sponsored research and published nearly 100 peer-reviewed journal articles and book chapters. He serves on the editorial boards of a number of journals including Transportation, Transport Reviews, Journal of Choice Modeling, and Transportation Letters. He is the chair of the Travel Analysis Methods Section of the Transportation Research Board and the immediate past chair of its Committee on Traveler Behavior and Values. He is also the immediate past chair of the International Association for Travel Behaviour Research (IATBR). Dr. Pendyala has his PhD and Masters degrees in Civil Engineering with a specialization in transportation systems from the University of California at Davis. He obtained his undergraduate degree in Civil Engineering from the Indian Institute of Technology - Madras in Chennai, India.

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

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]. In this context, examinations and other assessments undertaken during the progression of a course 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 are rarely used as a learning instrument. In the combined experience of the authors as instructors of engineering curricula, less than 5% 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 less problematic to grade and are often preferred by students and instructors. In a computation-intensive 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. If the student selects an incorrect item 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 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[4]. More specifically, a number of studies have evaluated the equivalence between, and advantages and drawbacks of, the CR and MC formats[3, 5, 6]. 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"[4].

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 introduce variation in students' scores between graders and open the instructor to complaints.

From the student perspective, MC items are generally preferred, although at times for reasons counterproductive to learning goals[7]. MC items are perceived by students to be "easier," both to prepare for and during the examination, as 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 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 methods 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] provide 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". Bauer *et al.*[16] explored several scoring algorithms for assigning partial credit in multiple-correct options MC items in the use of final examinations for third-year medical students and report that scoring systems that allocate partial credit to students provide better psychometric results than dichotomous scoring. 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. In the collective view of the above studies, MC tests can efficiently assess student achievement, in particular when used within a range of other learning and teaching strategies that encourage higher thinking skills and provide students and instructors alike with constructive feedback.

3 Experimental design

In the Department of Mechanical Engineering of University of South Florida, Numerical Methods is an undergraduate, junior-level course that follows the prerequisite mathematical course sequence of Calculus I, II, and III and Ordinary Differential Equations. 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) – started their college at University of South Florida, 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, 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[18]. 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[19]—*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 differ 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[20], 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 were 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 very few students required the fully allocated time.

4 Results

4.1 Student demographics

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 prerequisite GPA (PGPA) are shown in Tables 2 and 3.

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 1. Student participation 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

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

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 ($\kappa^2 = .479$, 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.

	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 3. Mean age and PGPA of students by semester.

	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 4. Student's t-test on PGPA between 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 throughout 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 student's previous academic performance as measured by PGPA, 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.

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 5. Averaged in-course examination grades by semester.

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.

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 Figure 1, where it can be seen that MC+PC were more likely to receive full credit.

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

Table 6. Final examination raw score by examination format.

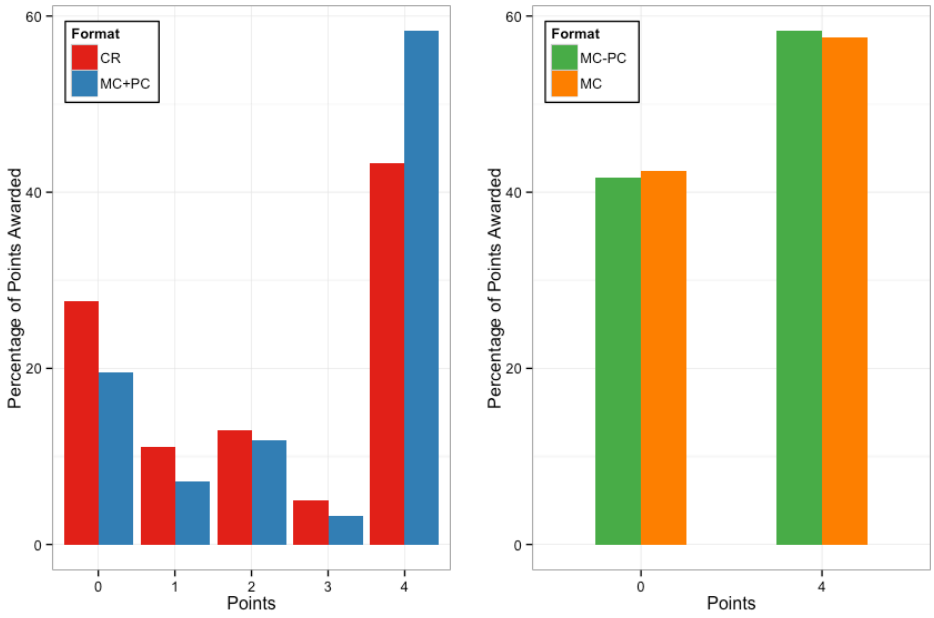

Figure 1: Percentage of points awarded to students by format

4.3 Validity and reliability 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-class examinations. A Spearman's rank test 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.

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

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 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 ($p<0.001$)

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[21] 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.

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

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

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 may be considered acceptable[22]. 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[23] 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.

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. Of the correctly answered items, 81% included work that would have otherwise contributed to partial credit. Of the 1560 total items requiring grading in the MC+PC section, only 17 items were left blank with no selected option, leaving 41% to be graded by hand. In comparison with the CR section, in which 1800 items required grading, 43% were answered correctly and thus required minimal scoring effort, while the remaining 57% had to be graded by hand.

4.4 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, $F(7, 114) = 15.249$, ($p<0.001$).

	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

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

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 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.

5 Conclusions

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 all students. Similarly, students' previous academic performance, both in the course and in prerequisite courses, was a significant predictor of performance on the final examination, although the effect of prerequisite course performance was slightly less than significant in the partial credit scoring method.

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 results presented indicate that the combination of MC items with the partial credit option provides an ideal middle ground between a CR- and a MC-only examination. Grading demands decreased significantly for the MC+PC examination when compared to the CR examination format, however the MC+PC demonstrated reliability equivalent to the CR-only format. Thus, the results suggest that the MC+PC examination format may provide a desirable balance between the high levels 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.

6 Acknowledgements

This material is based upon work supported partially by the National Science Foundation under Grant Nos 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.

7 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*, vol. 3, no. 2, pp. 145–158, Jul. 2002.

[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*, vol. 6, no. 2, pp. 103–118, Apr. 1993.

[4] M. C. Rodriguez, "Choosing an Item Format," in *Large-scale assessment programs for all students: validity, technical adequacy, and implementation*, G. Tindal and T. M. Haladyna, Eds. Mahwah, N.J.: Lawrence Erlbaum Associates, 2002, pp. 213–231.

[5] R. E. Bennett, D. A. Rock, and M. Wang, "Equivalence of Free-Response and Multiple-Choice Items," *Journal of Educational Measurement*, vol. 28, no. 1, pp. 77–92, 1991.

[6] G. R. Hancock, "Cognitive Complexity and the Comparability of Multiple-Choice and Constructed-Response Test Formats," *The Journal of Experimental Education*, vol. 62, no. 2, pp. 143–157, Jan. 1994.

[7] K. Struyven, F. Dochy, and S. Janssens, "Students' Perceptions About Evaluation and Assessment in Higher Education: A Review," *Assessment & Evaluation in Higher Education*, vol. 30, no. 4, pp. 325–341, Aug. 2005.

[8] M. Zeidner, "Essay Versus Multiple-Choice Type Classroom Exams: The Student's Perspective," *The Journal of Educational Research*, vol. 80, no. 6, pp. 352–358, 1987.

[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*, vol. 21, no. 1, pp. 65–88, Mar. 1997.

[10] F. Lord, M. Novick, and A. Birnbaum, *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley Pub. Co, 1968.

[11] P. L. Dressel and J. Schmid, "Some Modifications of the Multiple-Choice Item," *Educational and Psychological Measurement*, vol. 13, no. 4, pp. 574–595, Dec. 1953.

[12] C. H. Coombs, J. E. Milholland, and F. B. Womer, "The Assessment of Partial Knowledge," *Educational and Psychological Measurement*, vol. 16, no. 1, pp. 13–37, Mar. 1956.

[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*, vol. 2, no. 2, p.020102, Jul. 2006.

[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*, vol. 68, no. 4, pp. 957–971, 2002.

[16] D. Bauer, M. Holzer, V. Kopp, and M. R. Fischer, "Pick-N Multiple Choice-Exams: A Comparison of Scoring Algorithms," *Advances in Health Sciences Education: Theory and Practice*, vol. 16, no. 2, pp. 211–21, May 2011.

[17] K. F. Stanger-Hall, "Multiple-Choice Exams: An Obstacle For Higher-Level Thinking In Introductory Science Classes," *CBE Life Sciences Education*, vol. 11, no. 3, pp. 294–306, Jan. 2012.

[18] A. Kaw, E. Kalu, and D. Nguyen, *Numerical methods with applications*. Raleigh, NC: Lulu.com, 2010.

[19] B. S. Bloom, *Taxonomy of educational objectives. Handbook I: Cognitive domain*. New York: David McKay Company, Inc., 1956.

[20] 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*, vol. 15, no. 3, pp. 309–333, 2002.

[21] L. J. Cronbach, "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika*, vol. 16, no. 3, pp. 297–334, Sep. 1951.

[22] L. M. Rudner and W. D. Schafer. (2001, April) "Reliability ERIC Digest," *ERIC clearinghouse on assessment and evaluation* [Online]. Available: http://www.ericdigests.org/2002-2/reliability.htm

[23] C. Spearman, "Correlation Calculated from Faulty Data," *British Journal of Psychology*, vol. 3, no. 3, pp. 271–295, Oct. 1910.