WHY USE MULTIPLE-CHOICE QUESTIONS ON ACCOUNTING CERTIFICATION EXAMINATIONS?

Mark G. Simkin
College of Business Administration
University of Nevada
Reno, Nevada
USA

William E. Keuchler
College of Business Administration
University of Nevada
Reno, Nevada
USA

Arline Savage
Orfalea College of Business
California Polytechnic State University
San Luis Obispo, California
USA

Debra Stiver
College of Business Administration
University of Nevada
Reno, Nevada
USA

ABSTRACT

With the continued growth in popularity of professional certification examinations has come increasing interest in the composition and equity of the questions on such tests. Multiple-choice (MC) questions are easier to grade than constructed response (CR) questions, but are MC questions the best type of question to use on important certification examinations? This paper attempts to answer this question, using sample data from a number of university accounting classes. In an empirical study across a sample domain of 10 separate classes and nearly 450 students, we found only a weak

relationship between performance on the MC and CR portions of the tests. Consequently, the findings from this study suggest that certification examinations based solely on MC questions may not be testing applicants at the same level as CR tests might. Gender bias did not appear to be an issue in this study.

Key words: Multiple choice questions, constructed response questions, test formats, certification examinations, gender bias

Data availability: Data are available from the third author

INTRODUCTION

The accounting literature suggests that professional certification is important to, and desired by, employees, employers, and professional organizations. Potential personal benefits to individuals who earn certification include (1) pride in accomplishment, (2) official recognition of the recipient's knowledge, (3) enhanced career advancement opportunities, (4) impetus to improve work and communication skills, (5) increased job security, (6) ability to qualify for work in sensitive governmental or corporate areas, and (7) financial remuneration (Grigsby, 2000; Summerfield, 2008).

Employers also gain when they encourage their employees and job applicants to obtain professional certifications. Benefits include (1) an independent metric with which to screen job applicants, (2) a knowledgeable work force, (3) assurance of continuing education of employees, (4) greater currency in the field, (5) enhanced employee understanding of matters and concepts beyond their immediate job descriptions, (6) availability of objective criteria with which to award raises and promotions, and (7) an objective rationale for dismissing employees who repeatedly fail certification tests (Christensen, 1998).

Finally, benefits often accrue to the professional organizations that offer certifications, including (1) enhanced recognition and stature of the sanctioning organization, (2) ability to tailor certification tests to specific technical areas, (3) greater understanding of the strengths and weaknesses of test takers, (4) enhanced ability to provide feedback to relevant parties, and (5) revenues from examination fees (Grigsby, 2000; Summerfield, 2008).

Like many other professions, the field of accounting has a wide range of professional societies and concomitant certifications. Table 1 provides examples, listed alphabetically by certification name. One of the oldest and most respected–the certified public accountant (CPA) credential–has such stature that many university accounting programs "teach to" this exam and accounting programs are sometimes judged on the basis of the pass rates of their students on it (Jackson, 2006).

If certification examinations are important, then so are the types of questions used on them. Test takers, employers, and the test developers themselves have at least one objective in common: assurance that the questions on these assessments are fair and unbiased, and that the tests themselves accurately and equitably measure the applicants' knowledge in the subject area.

In most cases, the questions on certification examinations can be classified as either: (1) machine-gradable, multiple-choice (MC) questions that offer several possible alternative answers to each question, or (2) constructed response (CR) questions such as essays, simulations of real-

TABLE 1

Selected Certification Examinations Related to Accounting

Examination	Acronym	Sponsoring Organization	Cost	Duration (Hrs)	Composition
Certified Information Security Manager	CISM	Information Systems Audit and Control Association (ISACA)	$375 (members) $505 (non-members)	4	200 Multiple Choice (MC)
Certification in Control Self-Assessment®	CCSA	Institute of Internal Auditors (IIA)	$250 (members) $300 (non-members)	3.25	125 MC
Certified Financial Services Auditor®	CFSA	Institute of Internal Auditors (IIA)	$250 (members) $300 (non-members)	3.5	125 MC
Certified Fraud Examiner	CFA	Association of Certified Fraud Examiners	$990 and up	10	600 MC
Certified Government Auditing Professional®	CGAP	Institute of Internal Auditors (IIA)	$300	3.5	125 MC
Certified in the Governance of Enterprise IT	CGEIT	Information Systems Audit and Control Association (ISACA)	$325 (members) $455 (non-members)	4	120 MC
Certified Information Systems Auditor	CISA	Institute of Internal Auditors (IIA)	$375 (members) $505 (non-members)	4	200 MC
Certified Internal Auditor	CIA	Institute of Internal Auditors (IIA)	$575 (members	2.45 (each of 4 parts)	100 MC per part (400 total)
Certified Management Accountant	CMA	Institute of Management Accountants (IMA)	$700 (members)	4 (each of 2 parts)	100 MC and 2 30-minute essays per part
Certified Public Accountant	CPA	American Institute of Certified Public Accountants (AICPA)	varies by state	14 (all parts)	70% MC, 30% simulations

world problems, or computational problems. In the discussion that follows, the term "examination format" refers to the ratio of these two types of questions on a given certification examination.

Table 1 makes clear that the majority of accounting certification examinations rely heavily–or in most cases, solely–on multiple choice questions. The purpose of this article is to discuss the efficacy of using such questions to assess a test taker's understanding of a given body of knowledge. The next section of this paper discusses the theoretical merits and drawbacks of using such questions and reviews the empirical evidence. The third section reports the results of a new study that the authors conducted to empirically test the usefulness of MC questions as assessments of test taker knowledge. The fourth section provides further discussion of our results and also some caveats that limit our findings. The final section provides a summary and a set of conclusions for our work.

THE MC-CR CONTROVERSY

Like tests in academia, the objective of most certification examinations is to assess a candidate's knowledge in sufficient depth to assign a fair grade. Given the importance of many certifications to both applicants and employers, a great deal rides on the efficiency and equity of the process. This, in turn, has led both scholars and certification-exam developers to ask "what type of test format best performs this assessment task?"

Assessment Choices

MC examinations enjoy a number of important advantages. For example, they can be drawn randomly from computerized test banks and administered frequently to test takers; they can be graded easily, quickly, consistently, and accurately; they are usually viewed as objective; they can cover a wide range of subjects; and they can be returned to their test takers in relatively short periods of time (Zeidner, 1987; Snyder, 2003). These advantages probably account for the fact that, today, a host of professional certification examinations in such diverse disciplines as architecture, dentistry, engineering, law, medicine, optometry, pharmacology, computer programming, and veterinary sciences now use MC formats exclusively as assessment tools (Hamblen, 2006; Puliyenthuruthel, 2005). Historically, the one exception to this trend has been the AICPA exam, but 24 years ago, Mitchell Rothkopf, then director of the examination division, argued strongly for "no more essay questions on the uniform CPA exam" (Rothkopf, 1987).

On the other hand, the most common argument favoring CR questions is the widespread belief that they test a deeper understanding of the subject material (Hwang et al., 2008; Ingram and Howard, 1998; Bridgeman, 1992; Lukhele et al.,1994). A related perception is that CR test questions are better at evaluating a respondent's integrative skills–for example, because they demand a more–thorough grasp of the subject matter and its context–and therefore, better probe both the breadth and depth of the test taker's knowledge (Hancock, 1994; Rogers and Hartley, 1999; Bacon, 2003). The fact that CR questions also test the respondent's ability to organize thoughts into a cohesive answer adds to this preference (Shaftel and Shaftel, 2007). It is probably for these reasons that some certification examinations in such areas as engineering or information technology now have extensive, performance-based components. Similarly, essay questions have been re-introduced into the SAT examinations because its administrators felt that such questions better assess the accuracy of language and reasoning skills (Katz et al., 2000).

Professional accountants do not answer test questions for a living. Rather, they perform accounting tasks using the skills they have learned in school or acquired at work. Thus, a final

reason favoring CR tests is the greater likelihood of structural fidelity–i.e., the degree to which examination questions require the same problem-solving skills encountered in the work venues of a given field (Messick, 1993). This is particularly important in the accounting area, where employers are more interested in hiring competent accountants and auditors than good test takers.

Despite their advantages, CR questions have significant drawbacks, even for those who believe they are superior assessment tools. Perhaps the most important of them is that grading takes longer than for MC tests, tends to be more subjective, and often requires substantial prerequisite knowledge. This process is also more onerous for the evaluators themselves, who are more likely to be subject to both critical and litigious challenges. Finally, if the CR questions require writing samples or essays, some scholars (as well as many students) believe that CR questions naturally favor those individuals with superior writing skills, even if poorly-written answers have superior knowledge content (Zimmerman and Williams, 2003).

The MC-CR controversy includes one final component: the question of "gender bias" (Walstad and Robson, 1997; Hirschfeld et al.,1995). Certification test developers have a particularly strong interest in this matter because they have both a natural and a legal incentive to ensure that their examinations are "gender neutral"–i.e., that their tests do not favor males over females or vice versa. But are MC questions really gender neutral?

Given these concerns, the choice between using MC or CR questions on a given professional certification examination creates a natural dichotomy. Multiple-choice questions are more convenient to grade, but are considered by many researchers to be less effective at measuring deep conceptual understanding (Becker and Johnston, 1999), while CR questions are just the opposite. The issues of "test equity" and "test efficacy" therefore come down to the extent to which the two types of questions are related. If we can find a strong relationship between them, then certification test developers can use MC examinations almost exclusively on such examinations, knowing that whatever is measured by CR tests is also measured by the other, and saving thousands of hours of grading time in the process. Conversely, if only a weak relationship exists–or none at all–then examiners would appear to be remiss in relying exclusively on MC questions in certification examinations. What empirical evidence exists to answer this question?

Empirical Evidence

An extensive body of research has addressed the question of how well MC versus CR questions test understanding of content material. Much of the theoretical work comes from the areas of educational psychology and educational assessment (Martinez, 1999; Hancock, 1994; Nunnally and Bernstein, 1994; Simkin and Kuechler, 2005). While these works suggest that it is theoretically possible to construct MC items that measure many of the same cognitive abilities as CR items, the question remains of how well this theory holds up empirically.

Early studies of this hypothesis have led some scholars to conclude that MC tests and CR tests measure the same thing (Traub, 1993; Wainer and Thissen, 1993; Bennett et al., 1991; Bridgeman, 1991). After examining sample tests from seven different disciplines, for example, Wainer and Thissen (1993) concluded that this relationship was so strong that "whatever is … measured by the constructed response section is measured better by the multiple choice section…We have never found any test that is composed of an objectively and a subjectively scored section for which this is not true" (p. 116). Subsequent studies by Walstad and Becker (1994) and Kennedy and Walstad (1997) echoed these sentiments.

A number of recent studies dispute these findings. For example, using a two-stage least squares estimation procedure, Becker and Johnston (1999) found no relationship between student performance on MC and essay questions on economics examinations, and therefore concluded that "these testing forms measure different dimensions of knowledge" (p. 348). A similar study in physics instruction by Dufresne et al. (2002. p. 175) led the authors to conclude that student answers on MC questions "…more often than not [give] a false indicator of deep conceptual understanding." Finally, in the areas of accounting and information systems, works by Kuechler and Simkin (2003) and Bible et al. (2008) found only moderate relationships between the MC and CR portions of the study examinations.

A potential explanation for why earlier empirical work has not yielded more consistent findings is that the type of CR questions used in the analysis may vary by domain. For example, if the CR portions of most English literature examinations involve essay questions, the higher-level organizational and expository skills required to answer them are less likely to correlate with student performance on MC questions, ostensibly over the same material. But this may not be a problem in the accounting or engineering areas, where CR questions are often computational, usually require students to adhere to more rigid standards of logic, and correct answers have less room for variance. As a result, it may be reasonable to expect a closer relationship between student performance on the MC and CR portions of student examinations in these areas. Again, a closer relationship bodes well for professional certification examinations that rely solely on MC questions.

As noted earlier, there is also the issue of gender bias. Here, too, the empirical evidence is inconsistent. Some studies have shown a possible advantage of males relative to females on MC tests (Lumsden et al., 1987; Bolger and Kellaghan, 1990; Lumsden and Scott, 1995). Bridgeman and Lewis (1994) estimated this advantage to be about one-third of one standard deviation. However, other studies have shown no significant difference between males and females when both are evaluated using a MC test rather than a CR test (Chan and Kennedy, 2002; Greene, 1997; Walstad and Becker, 1994).

It is possible that "gender bias" may depend upon a variety of factors, including the particular sample used to address this question as well as the discipline from which the sample is taken (Hamilton, 1999; Gallagher and De Lisi, 1994). The relevant question here is whether MC questions are gender-neutral on accounting certification examinations. To date, several studies have explored the relationship between "test format" and "gender bias" in the accounting area. One study by Lumsden and Scott (1995) found that males scored higher than females on MC questions when taking the economics section of the Chartered Association of Certified Accountants examination, while another study by Gul et al. (1992) found no gender differences in students taking MC examinations in an auditing class. Tsui et al. (1995) replicated the Gul et al. (1992) study, but did not report gender differences.

Finally, Bible et al. (2008) identified "gender" as a small, but statistically significant determinant of student performance on MC questions on intermediate accounting tests. However, these researchers also found that the statistical significance varied by class section, leading them to conclude that gender differentials were at best "small." They also noted that the language-processing skills often required on the essay portions of tests in other disciplines were not at play in the tests used in their study.

A NEW STUDY

In the authors' opinion, the empirical evidence on the relationship between performance on MC and CR examinations is notable for what it does not provide: a definitive answer to the question of how close these assessments are to one another in measuring a test taker's understanding of a given body of knowledge, and also whether or not MC questions have a gender bias. If a close relationship can be found, then certification examiners can find assurance that their MC tests are as useful a predictor as CR tests of the professional competencies of those who take them. Conversely, if a close relationship cannot be found, it is easy to ask "of what use are such tests?" Consequently, given the discussions above, the authors were interested in investigating two specific research questions: (1) can multiple-choice examinations capture a substantial amount of the variability of constructed response scores (i.e., performance on MC questions correlates significantly and directly with performance on CR questions), and (2) are there significant differences in test performance that can be attributed to gender differences in the test takers?

The relative strength of the relationship between student performance on MC and CR tests is important. What we would like is a means to predict student performance on CR tests using more tractable MC questions. It is this possibility that the present study sought to confirm. If the relationship is strong enough (a subjective matter), then test developers can have the best of both worlds–MC exams that are easy to grade and comfort in the knowledge that such tests evaluate as much mastery of professional materials as CR tests.

Because CR questions like word problems can create a richer "prompting environment" that is known to enhance recall under some conditions, these and other factors have led researchers to conclude that multiple-choice and constructed response questions tap into different learning constructs (Becker and Johnston, 1999). However, the intent of this research was to determine the degree to which whatever is measured by one type of measure captures a meaningful amount of the variation in whatever is measured by the other type of measure.

Methodology

To answer our research questions, the authors conducted a study over four semesters to investigate the relationship between the two types of performance measures. The sample data consisted of the test scores of 448 students who had enrolled in accounting classes at two different state universities. Each student was required to complete two parts of the same test: one part consisting of MC questions and one part consisting of CR questions. Although the enrollees in the classes obviously differed from semester to semester, the two types of questions on each test covered the same class materials and attempted to measure student understanding of the subject matters.

All the students in the study were taking the courses for credit. In this sample, 41 percent of the students were female and 59 percent were male. Because the influence of "gender" has been identified in prior research as potentially important, we also included this variable in our investigations to see if it affected student test performance. Following prior experimental designs, CR scores acted as the dependent variable and a multiple linear regression analysis was used to determine the degree to which MC questions and certain demographic variables (described in greater detail below) were useful predictors of performance on the CR portion of each examination.

In the student tests used in this study, each MC question referred to a separate aspect of the associated accounting material and had four possible answers, labeled A through D. Figure 1 illustrates a typical MC question. Students answered this section of the exam by blackening a square on a scantron scoring sheet for a particular question. The number of correct responses for the

A company purchased office supplies costing $3,000 and debited Office Supplies for the full amount. At the end of the accounting period, a physical count of office supplies revealed $1,600 still on hand. The appropriate adjusting journal entry to be made at the end of the period would be
A) Debit Office Supplies, $1,600; Credit Office Supplies Expense, $1,600.
B) Debit Office Supplies, $3,400; Credit Office Supplies Expense, $3,400.
C) Debit Office Supplies Expense, $3,400; Credit Office Supplies, $3,400.
D) Debit Office Supplies Expense, $1,400; Credit Office Supplies, $1,400.

Figure 1. A typical multiple-choice question

Prepare the journal entries to record the following transactions on Paula Worth Company's books using a perpetual inventory system. On February 6, Paula Worth Company sold $60,000 of merchandise to the Jones Company, terms 2/10, net/30. The cost of the merchandise sold was $30,000. On February 8, the Jones Company returned $10,000 of the merchandise purchased on February 6 because it was defective. The cost of the merchandise returned was $5,000. On February 16, Paula Worth Company received the balance due from the Jones Company.

Figure 2. A typical constructed-response (CR) question

different types of questions on each exam was scaled to a percentage to permit us to test student performance consistently across semesters.

In contrast to the MC questions in the sample tests, each CR question required a student to perform computations for some specific accounting event. Figures 2 and 3 illustrate typical questions. Each CR problem was worth several points–typically in the range of 5 to 25 points. All of these exam questions were graded using an answer template, limiting grading variations and restricting the subjectivity involved in the process. The instructor in each class used the same answer key to grade each CR question, awarding partial credit consistently and in the same proportion for partially-correct answers.

The MC questions appeared first in each examination, followed by the CR (problem-solving) questions. The cover page of the exam contained test instructions, information about how much each question was worth, and the maximum time available for the exam. In practice, most students began by answering the MC questions, but test takers could also "work backwards" if they wished and begin with the CR questions. This alternate test taking strategy is common among students who prefer CR questions as well as for some international students for whom English is a second language (Kuechler and Simkin, 2003). Due to the higher point weighting, it is also possible that subjects focused more attention on the multiple-choice questions than the problem solving/essay (CR) questions.

Some prior studies of examination effectiveness have adjusted the scores of MC tests in an attempt to correct for the effects of guessing. One of the most recent investigations of the influence of guessing on MC tests by Zimmerman and Williams (2003) concurs with most prior research on that topic in stating that "guessing contributes to error variance and diminishes the reliability of

Bank Reconciliation

Part 1 - Multiple Choice Questions

1. Which journal(s) are normally associated with the General Checking GL Account (and were used as such in your project)?
    a. General Journal
    b. Cash Receipts Journal
    c. Cash Disbursements Journal
    d. All of the above

2. Last month's bank reconciliation is used for this month's bank reconciliation to identify:
    a. Deposits in transit that have finally cleared the bank
    b. Outstanding checks that have finally cleared the bank
    c. Both A & B
    d. None of the above. Last month's bank reconciliation is NOT used except as a template.

3. Which of the following reconciling items on the bank reconciliation require an adjustment to the balance of the cash general ledger account?
    a. A monthly service charge fee   d. errors by the bank
    b. Outstanding checks   e. All of the above require adjustment
    c. Deposits-in-transit

Part 2 - Constructive Response Question

See attached bookkeeping and accounting information for Your Company (General Ledger, General Journal, Cash Receipts Journal, Cash Payments Journal, Bank Statement, last month's reconciling items for the bank reconciliation).

Complete the accounting task below. Include complete and proper headings:

Start with the Cash Operating account and do a bank reconciliation for the General Checking bank account. Record and post any necessary adjustments in the books.

Figure 3. A multiple-choice and related constructed-response (CR) question

tests." These authors also note that the variance depends on multiple factors, including the ability of the examinee, in ways that are not fully understood. For this reason, rather than adjust the raw scores, we preferred to perform our regressions on unadjusted scores and inform the reader of the effect of the likely increase in variance in our data sets and its effect on regression analyses. Again, our objective was not to explain the variability of MC or CR tests, but rather to investigate whether performance on MC tests and CR tests are related.

Data

To address these questions, the authors gathered data from a wide range of accounting courses and students. Table 2 describes the 10 accounting classes used in our study and the number

TABLE 2

Sources of Sample Data

Class	Class Title	Number of Students
Class 1	Introductory Financial Accounting	46
Class 2	Introductory Financial Accounting	46
Class 3	Introductory Financial Accounting	45
Class 4	Introductory Financial Accounting	45
Class 5	Introductory Financial Accounting	46
Class 6	Accounting Information Systems	54
Class 7	Accounting Information Systems	39
Class 8	Accounting Information Systems	37
Class 9	Intermediate Financial Accounting	45
Class 10	Intermediate Financial Accounting	34

of students enrolled in each of them. To ensure independent sample observations, we excluded secondary test scores from the few students who were simultaneously or consecutively enrolled in two of these classes.

Dependent and Independent Variables

The dependent variable for the study was the (percentage) score on the CR portion of each class examination. From the standpoint of this investigation, the most important independent variable was the student's score on the MC portion of the examination. As illustrated in Figure 2, the CR questions required students to create journal entries or perform similar tasks. To ensure consistency in grading CR answers, the same instructor manually graded all the CR questions on all examinations for a given class, indicating the correct answer(s) to each CR question as well as a list of numerical penalties to assess for common errors.

Another important independent variable used in this study was "gender." As noted above, a number of prior studies have detected a relationship between "gender" and "computer-related outcomes" (Hamilton, 1999; Gutek and Bikson, 1985). Accordingly, "gender" was included in the linear regression model as a dummy variable, using "1" for females and "0" for males. Finally, the examinations in each class had different, and differently-worded, questions, and the tests were taken by different students. Accordingly, we added dummy (0-1) variables to the regression equation for each class to account for these differences.

In summary, the regression equation tested in this study used student performance on the CR portion of a final or midterm examination as the dependent variable, and student performance on the MC portion of the exam, gender, and semester dummy variables as independent variables. Table 3 lists these independent variables and provides brief descriptions of them.

Results

Table 4 displays the results from the regression analysis described above. Coefficient values, t-statistics, and probabilities are only shown for nine of the ten classes because we used 0-1 dummy

TABLE 3

Independent Variables for the Regression Model

Independent Variable	Description
MC	The (percentage) score on the multiple-choice portion of an examination
Gender	A dummy variable: 0 = male, 1 = female
Class (1, 2, etc.)	A dummy variable (0-1) to account for the different examination questions given to each class

TABLE 4

Linear Regression Results

	Coefficient (β)	t-statistic	Probability
Intercept	47.11	13.89	<.001
MC %	0.56	13.66	<.001
Gender	1.30	1.21	.229
Class 1	-7.41	-3.16	.002
Class 2	-8.27	-3.09	.002
Class 3	-9.15	-3.37	.001
Class 4	-6.60	-2.58	.010
Class 5	-0.73	-0.19	.848
Class 6	10.03	-3.75	<.001
Class 7	-7.52	-2.92	.004
Class 8	-4.86	-1.88	.061
Class 9	1.95	0.76	.446

$N = 448$, $F = 18.5$, $df = 11$, $P < .001$, $R^2 = .301$

variables for them and the tenth class acted as a "base" for the others. Table 5 shows the correlation coefficients for all the regression variables.

The t-statistics for almost all of the equation regressors were statistically significant (for $p < 0.05$) and most were significant at $p < 0.01$. The beta coefficient of "0.56" for the MC questions is noteworthy. Its positive sign shows that student performance on MC questions is positively correlated with student performance on CR questions, and its magnitude of "0.56" means that a student was likely to earn, on average, a little more than a half percentage point on each CR question for every one percentage point earned on the MC portion of the same test. Because the relative proportion of MC and CR questions varied from class to class and test to test, this value of 0.56 is an aggregate scaling factor for the relationship between the two parts, rather than a direct measure of the equivalence of these two parts.

TABLE 5

Correlation Matrix for All Variables

	CR	MC	Gender	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9
MC	0.499**											
Gender	0.053	-0.009										
Class 1	0.035	0.156**	0.014									
Class 2	-0.053	0.007	-0.047	-0.157**								
Class 3	-0.064	0.017	-0.033	-0.153**	-0.093*							
Class 4	0.065	0.148**	-0.031	-0.188**	-0.114*	-0.111*						
Class 5	0.033	0.063	0.143**	-0.084	-0.051	-0.050	-0.061					
Class 6	0.011	0.191**	0.001	-0.170**	-0.103*	-0.100*	-0.124**	-0.055				
Class 7	-0.100*	-0.100*	0.016	-0.170**	-0.103*	-0.100*	-0.124**	-0.055	-0.112*			
Class 8	0.012	-0.022	0.016	-0.170**	-0.103*	-0.100*	-0.124**	-0.055	-0.112*	-0.112*		
Class 9	0.054	-0.236**	0.001	-0.170**	-0.103*	-0.100*	-0.124**	-0.055	-0.112*	-0.112*	-0.112*	
Class 10	-0.004	-0.215	-0.026	-0.146	-0.088	-0.086	-0.106	-0.048	-0.096	-0.906	-0.096	-0.096

Note: for significant correlation coefficients, * indicates .05 and ** indicates .01

The results for the gender variable were inconclusive. Because the underlying dummy variable was coded as "0" for males and "1" for females, the regression coefficient of "1.3," if significant, would imply that females have a strong advantage over males on CR questions. However, the P value of .229 indicates that gender is not statistically significant. Because the conclusion of a strong positive bias for females on MC tests is inconsistent with earlier studies (Bible et al., 2008; Kuechler and Simkin, 2003; Gutek and Bikson, 1985), all of which found small gender differentials favoring males on MC tests, we consider the non-significant positive result for females an artifact of this particular data set. The lack of gender significance in fact reflects the meta-level inconsistency found for gender effects in prior studies.

Finally, we found that seven out of nine of the coefficients for our class (dummy) variables were statistically significant. As noted earlier, this finding merely confirms our assertions that a myriad of factors differed from test to test, and therefore, between classes and semesters–for example, how hard each test was relative to the others.

Perhaps the most important statistic in our linear regression was the adjusted R^2 value of ".301." This finding is consistent with similar models tested by Kuechler and Simkin (2003), who found a comparable value of "0.45" in a similar study of information systems students. The ultimate assessment of whether such a result is "good" or "bad" is subjective, but (for the reasons explained above) we expected a higher value.

Individual Class Regressions

To isolate the possible interactive effects of the differentiating factors caused by a different set of tests and students, the authors also performed regressions for each class separately, using only an intercept, MC, and gender as explanatory variables. Table 6 reports our findings.

As with the results in Table 4, most of the regression coefficients for these latter models were statistically significant. However, the actual estimation values differed from semester to semester, reflecting the differences in test questions, variations in teaching effectiveness, and different student compositions. Of particular interest were the coefficient estimates for the MC variable, which ranged in value from ".20" (for class 10) to "0.83" (for class 2). Again, the positive signs of these estimates reinforce the intuitive notion that student performance on MC and CR questions are related to one another, and again, none of these estimates were statistically differentiable from a desirable coefficient value of "1.0"–the ideal value indicating that performances on MC and CR tests are perfectly synchronized.

We note that the regressions for five of the ten classes taken individually have adjusted R^2 values less than the summary regression containing all data points. We also note that "gender" was significant for only one class (Class 2). Both of these observations lend credibility to the results of the all-class regression (Table 4), which also reports a small R^2 value and a non-statistically-significant gender coefficient (p =.229).

Two Additional Statistical Tests

To further examine the statistical properties of our sample data, we performed both a matched-pair test for the mean difference in scores and a Wilcoxon signed rank test for each of the 10 different classes in our sample. Again, we tested our class data separately in order to reduce the confounding effects potentially inherent in combining the data from different class subjects, varying amounts of course support, alternate instructors, and similar artifacts. Because we could not

TABLE 6
Estimates of Regression Coefficients for Individual Classes

	Class 1	Class 2	Class 3	Class 4	Class 5
Intercept	33.17*	15.39*	27.34	70.48*	51.25*
MC (%)	0.77*	0.83*	0.69*	0.21*	0.53*
Gender	-1.08	6.65*	2.30	2.30	all female
Adjusted R^2	0.56	0.45	0.17	0.31	0.26
df, F, p	2, 59.77, <.001	2, 16.66, <.001	2, 4.78, .015	2, 3.10, .054	2, 4.56, .060

	Class 6	Class 7	Class 8	Class 9	Class 10
Intercept	17.08	33.52*	43.62*	65.64*	72.26*
MC (%)	0.82*	0.62*	0.57*	0.31*	0.20*
Gender	-4.07	4.06	-2.96	2.20	-0.48
Adjusted R^2	0.31	0.29	0.48	0.20	0.12
df, F, p	2, 11.22, <.001	2, 9.89, <.001	2, 21.27, <.001	2, 6.39, .004	2, 3.20, .054

* Statistically significant at p=0.05 level

TABLE 7
Summary of Results for Matched-Pair Tests and Wilcoxon Signed Rank Test

	Matched-Pair t-tests			Wilcoxon Sgined Rank Tests		
	N	t value	p value	N	t value	p value
Class 1	46	1.69	0.098	45	1.55	0.120
Class 2	46	4.03	0.000	44	3.98	0.000
Class 3	45	3.16	0.003	45	3.33	0.001
Class 4	45	5.72	0.000	45	4.72	0.000
Class 5	46	-0.27	0.789	43	0.52	0.603
Class 6	54	2.18	0.034	53	1.97	0.049
Class 7	39	2.57	0.014	39	2.45	0.015
Class 8	37	1.28	0.209	37	1.59	0.113
Class 9	45	10.03	0.000	45	5.80	0.000
Class 10	34	8.38	0.000	34	4.96	0.000

definitively determine whether these differences were normally distributed, we also performed a non-parametric Wilcoxon test. Table 7 summarizes the results of our analyses.

Matched-pair tests examine the difference between each pair of values in a sample–in this case, a student's score on his or her MC test and the corresponding score on his or her CR test. The null hypothesis here is that the scores come from the same population (i.e., the mean difference in

scores is zero). The analyses in Table 7 suggest that only the data from classes 1, 5 and 8 show no significant differences. In all other classes, the t-test statistic was significant at an alpha level of .05.

Parametric tests require several assumptions regarding the normality of the distribution for the test statistic of interest, and are also sensitive to outliers whose magnitudes can overwhelm the conforming behavior of the remaining data. Accordingly, we also performed a series of Wilcoxon signed rank tests to overcome this potential problem in our ten sets of class data. Like a matched-pair test, a Wilcoxon test also computes the difference in each student's CR and MC scores, but then assigns a rank order to each absolute difference. The test statistic is the mean of these ranks (after reassigning the positive or negative sign to each rank from the original difference). Differences of zero are removed from further consideration, explaining why the number of observations, n, differs for the same class in some rows of Table 7.

Under the null hypothesis that the MC and CR scores come from the same population, the average of these terms should cluster around zero. For each class tested, the results for the Wilcoxon signed rank test were consistent with the results of the paired t-test. Again, only classes 1, 5 and 8 were statistically insignificant at an alpha level of .05.

Our finding of statistical significance for the majority of our data somewhat contrasts with some of our regression results in terms of students performing consistently between the two test formats. The null hypothesis for both statistical tests reported in Table 7 is that the scores from the MC and CR portions of our student exams came from the same population. The statistical outcomes for seven out of ten tests performed here were that they did not–i.e., that student performance on the separate assessment measures was meaningfully different. In fact, in our sample classes, students consistently performed better on the CR portions of their exams, with mean differences ranging from 3.9 percentage points to 19.7 points. We speculate that a student's ability to earn partial credit on the CR portions of these tests may explain some of this.

DISCUSSION AND CAVEATS

Our statistical analysis suggests that a positive relationship exists between the MC and CR scores. This is to say that, while there may be differences in the magnitude of scores, the scores at least vary in a similar direction. The good news in this finding is the implication that using MC tests to measure understanding of a given subject area at least points in the right direction–i.e., those individuals who do well on MC tests are also likely to do well on CR tests, and vice versa. This is heartening news to certification test developers who seek exactly this relationship.

The results of both our matched-pair tests and our non-parametric tests are somewhat less encouraging, however–indicating that the MC and CR portions of each exam were not equivalent tests. The MC portions of all these student tests, even after adjusting for "gender" and other potential differences in our sample classes, explained about 30 percent of the variation in the scores of the CR portions of those same tests. Certification test developers are likely to be disappointed that this statistic was not higher. It indicates that the independent variables of our model–particularly student performance on the MC portions of the sample examinations–explained a little less than one third of the variability of the dependent (CR) variable. To us, it also suggests that the MC and CR portions of these tests are testing different types of understanding, or perhaps testing the same understanding at different depths.

The reasons for using MC tests instead of CR tests are well documented, and the authors acknowledge the convenience of employing them on certification examinations. We would also be comfortable with such formats if we can show that there is a clear and consistent positive

relationship between the two test formats. Our analysis, using data that were drawn from ten different accounting classes, suggests that this relationship, while positive, appears to be relatively weak. We are not alone in this finding–a number of scholars performing parallel tests on other student subjects and at other universities have reached the same conclusion in other disciplines (Becker and Johnston, 1999; Dufresne et al., 2002; Kuechler and Simkin, 2003).

This leads us back to the initial question posed in the title of this paper: "Why Use Multiple Choice Questions on Accounting Certification Examinations?" If the purpose of such examinations is to assess the professional qualifications of their test takers, we cannot defend their use and therefore have no answer to this question.

Caveats

In our analyses, we recognize that a host of factors may have led us astray. For example, any study is likely to have some sampling error. Here, we discuss some additional possibilities. One problem with any study using student participants is the possibility that its researchers have studied the wrong population. We recognize this possibility, but are not sure it applies here. After all, most certification exams require their test takers to complete a four-year degree in accounting or a related field at an accredited college–i.e., the test takers in the population from which our sample was drawn. Nonetheless, sampling bias is still possible.

Another possible explanation for our lack of stronger results is the likelihood that the different test formats are not equivalent–i.e., the possibility that the MC questions and CR questions test different levels of knowledge and that such questions therefore require different skill sets of their test takers. We readily admit this possibility, and in fact embrace it. We do not want to prove that these test formats are different. We already believe that they are different. Rather, what we sought to find was a relationship between student performance on these two, very different types of tests. If we had found a strong connection, we could then take comfort in the fact that certification tests, if carefully constructed, can indeed rely solely on MC test formats to perform the assessment tasks desired of them. Empirically, however, we were disappointed that we did not find a closer relationship.

We also note that we are all individuals, and that most people are more comfortable taking one type of test than the other. Some students complain, for example, that a forthcoming test will only be multiple choice–a test format they feel they "don't do well with"–while others express the opposite preference. We acknowledge the possibility that our sample contained a preponderance of individuals with one disposition rather than the other–e.g., we happened to pick students who, say, liked MC tests. In such circumstances, the connection between performance on MC and CR tests would be tenuous, leading to the results we did, in fact, observe here. Given our large sample and ten classes, however, this possibility seems remote.

A similar comment applies to the potential for an ordering bias in our study, caused by the fact that the student examinations used in this study consistently placed the MC sections of the tests before the CR portions. At least two items mitigate this possibility. One is the fact that students were free to answer the questions in any order, and all students were given enough time to finish all parts of their examination. The other, more important one, is the fact that no students were asked to leave the exam room before they had finished the entire test.

We also recognize that certification examiners invest substantial resources in creating and validating their tests, and that the MC questions they contain are not necessarily equivalent to the ones we used in our analyses. We certainly believe that carefully-worded MC questions are better

than ambiguous or poorly-worded ones. On the other hand, we have no reason to believe that the MC questions used here were different enough to invalidate our study results. Similarly, for test takers, we acknowledge that the motivation for studying and passing certification examinations may be greater than for passing college examinations, and that other differences in testing environments are also likely to be at work here.

Finally, we note that the theoretical advantages of CR examination questions in assessing deep understanding of subject matter can be diminished significantly by improper item composition or hurried or inadequate grading. However, proper item composition is a factor with MC examinations as well, and increases in impact at the higher levels of knowledge that might be targeted by such questions. Moreover, we would hope that for those certification examinations composed by public certification groups and intended to be administered to multiple audiences over an extended period of time, both the composition of examination items and the grading of examinations would be at the highest levels.

SUMMARY AND CONCLUSIONS

There are many reasons why employers, employees, and the accounting profession itself all value professional certification, an esteem reflected by the large number of certification examinations available to individuals. Our review of such certification examinations found that almost all of them rely solely on multiple-choice questions for testing purposes–a format that has many grading advantages but little structural fidelity. For a wide variety of reasons, MC tests are themselves controversial and not universally accepted as the best assessment tools for measuring professional accounting competencies–constructed response tests such as computational questions are usually preferred as theoretically superior metrics.

The question we sought to answer was whether there was any relationship between performance on these two types of tests. In our empirical study of this relationship across a sample domain of 10 separate classes and nearly 450 students, we found a positive relationship between performance on the MC and CR portions of their tests, but not a particularly strong one. At best, what this means is that the two types of questions measure different, but perhaps useful, things. At worst, it means that MC questions only examine rudimentary levels of knowledge–or perhaps nothing other than the MC test taking ability of the applicants.

Our inability to find a stronger relationship between CR and MC test formats does not mean that MC tests do not assess some form of knowledge. They almost certainly do. The issue is whether whatever knowledge is tested by less tractable, but perhaps more probing and structurally viable CR tests, can also be tested efficiently by MC tests.

What does all this mean to certification examinations? The findings from this study suggest that certification examinations based solely on MC questions may not be testing applicants at the same level as CR tests might. Of greater concern is the possibility that the final scores on such certification tests may actually be giving false signals about what applicants do or do not know. It is possible such tests may be denying some applicants the opportunity to demonstrate what they do know. For these reasons, we encourage certification examiners to follow the lead of the CPA exam developers and include at least some CR questions on their tests.

REFERENCES

Bacon, D. R. 2003. Assessing Learning Outcomes: A Comparison of Multiple-Choice and Short Answer Questions in a Marketing Context. *Journal of Marketing Education* (Vol. 25) 31-36.

Becker, W. E., and C. Johnston. 1999. The Relationship Between Multiple Choice and Essay Response Questions in Assessing Economics Understanding. *Economic Record* (Vol. 75) 348-357.

Bennett, R. E., D. A. Rock, and M. Want. 1991. Equivalence of Free-Response and Multiple-Choice Items. *Journal of Educational Measurement* (Vol. 28) 77-92.

Bible, L., M. G. Simkin, and W. L. Kuechler. 2008. Using Multiple-Choice Tests to Evaluate Students' Understanding of Accounting. *Accounting Education* (Vol. 17) S55-S68.

Bolger, N., and T. Kelleghan. 1990. Method of Measurement and Gender Differences in Scholastic Achievement. *Journal of Educational Measurement* (Vol. 27) 165-74.

Bridgeman, B. 1991. Essays and Multiple-Choice Tests as Predictors of College Freshman GPA. *Research in Higher Education* (Vol. 32) 319-332.

_____. 1992. A Comparison of Quantitative Questions in Open-Ended and Multiple-Choice Formats. *Journal of Educational Measurement* (Vol. 29) 253-271.

_____., and C. Lewis. 1994. The Relationship of Essay and Multiple-Choice Scores with Grades in College Courses. *Journal of Educational Measurement* (Vol. 31) 37-50.

Chan, N., and P. E. Kennedy. 2002. Are Multiple-Choice Exams Easier for Economics Students? A Comparison of Multiple Choice and Equivalent Constructed Response Examination Questions. *Southern Economic Journal* (Vol. 68) 957-971.

Christensen, L. 1998. Annual Professional Certification Testing - An Idea Whose Time Has Come? *Armed Forces Comptroller* (Vol. 43, No. 4) 20.

Dufresne, R. J., W. J. Leonard, and W. J. Gerace. 2002. Making Sense of Students' Answers to Multiple-Choice Questions. *The Physics Teacher* (Vol. 40) 174-180.

Gallagher, A. M., and R. De Lisi. 1994. Gender Differences in Scholastic Aptitude Test - Mathematics Problem Solving Among High-Ability Students. *Journal of Educational Psychology* (Vol. 86) 204-211.

Greene, B. 1997. Verbal Abilities, Gender, and the Introductory Economics Course: A New Look at an Old Assumption. *Journal of Economic Education* (December 22) 13-30.

Grigsby, J. 2000. The Real Benefits of Professional Certification. *Receivables Report for America's Health Care Financial Managers* (Vol. 15, No. 1) 1, 12.

Gul, F. A., H. Y. Teoh, and R. Shannon. 1992. Cognitive Style as a Factor in Accounting Students' Performance on Multiple Choice Examinations. *Accounting Education: An International Journal* (Vol. 1) 311-319.

Gutek, B. A., and T. K. Bikson. 1985. Differential Experiences of Men and Women in Computerized Offices. *Sex Roles* (Vol. 13) 123-136.

Hamblen, M. 2006. Users Struggle on Road to Cisco Certification. *Computerworld* (Vol. 40, No. 28) 20.

Hamilton, L. S. 1999. Detecting Gender-Based Differential Item Functioning on a Constructed-Response Science Test. *Applied Measurement in Education* (Vol. 12) 211-236.

Hancock, G. R. 1994. Cognitive Complexity and the Comparability of Multiple-Choice and Constructed-Response Test Formats. *Journal of Experimental Education* (Vol. 62) 143-157.

Hirschfeld, M., R. L. Moore, and E. Brown. 1995. Exploring the Gender Gap on the GRE Subject Test in Economics. *Journal of Economic Education* (Vol. 26) 3-15.

Hwang, N. R., G. Lui, and M. Yew Jen Wu Tong. 2008. Cooperative Learning in a Passive Learning Environment: A Replication and Extension. *Issues in Accounting Education* (Vol. 23) 67-75.

Ingram, R. W., and T. P. Howard. 1998. The Association Between Course Objectives and Grading Methods in Introductory Accounting Courses. *Issues in Accounting Education* (Vol. 13, No. 4) 815-832.

Jackson, R. E. 2006. Post Graduate Educational Requirements and Entry into the CPA Profession. *Journal of Labor Research* (Vol. 27, No. 1) 101-114.

Katz, I. R., R. E. Bennett, and A. E. Berger. 2000. Effects of Response Format on Difficulty of SAT Mathematics Items: It's Not the Strategy. *Journal of Educational Measurement* (Vol. 37) 39-57.

Kennedy, P. E., and W. B. Walstad. 1997. Combining Multiple-Choice and Constructed-Response Test Scores: An Economist's View. *Applied Measurement in Education* (Vol. 10) 359-75.

Kuechler, W. L., and M. G. Simkin. 2003. How Well Do Multiple-Choice Tests Evaluate Student Understanding in Computer Programming Classes? *Journal of Information Systems Education* (Vol. 14, No. 4) 389-400.

Lukhele, R., D. Thissen, and H. Wainer. 1994. On the Relative Value of Multiple-Choice, Constructed Response, and Examinee-Selected Items on Two Achievement Tests. *Journal of Educational Measurement* (Vol. 31, No. 3) 234-250.

Lumsden, K. G., A. Scott, and A. Becker. 1987. The Economics Student Reexamined: Male-Female Differences in Comprehension. *Journal of Economic Education* (Vol. 18, No.4)365 - 375.

_____, and _____. 1995. Economic Performance on Multiple Choice and Essay Examinations: A Large-Scale Study of Accounting Students. *Accounting Education: An International Journal* (Vol. 4) 153-167.

Martinez, M. E. 1999. Cognition and the Question of Test-Item Format. *Educational Psychologist* (Vol. 34) 207-218.

Messick, S. 1993. Trait Equivalence as Construct Validity of Score Interpretation Across Multiple Methods of Measurement. In Bennett, R. E., and W. C. Ward. *Construct Versus Choice in Cognitive Measurement*. (Hillsdale, NJ: Lawrence Erlbaum) 61-73.

Nunnally, J., and I. Bernstein. 1994. *Psychometric Theory*. (New York, McGraw Hill).

Puliyenthuruthel, J. 2005. How Google Searches-for Talent. *Business Week Issue* (April 11) 52.

Rogers, W. T., and D. Harley. 1999. An Empirical Comparison of Three- and Four-Choice Items and Tests: Susceptibility to Testwiseness and Internal Consistency Reliability. *Educational and Psychological Measurement* (Vol. 59) 234-247.

Rothkopf, M. 1987. No More Essay Questions on the Uniform CPA Examination. *Accounting Horizons* (Vol. 1, No. 4) 79-85.

Shaftel, J., and T. L. Shaftel. 2007. Educational Assessment and the AACSB. *Issues in Accounting Education* (Vol. 22, No. 2) 215-232.

Simkin, M. G., and W. L. Kuechler. 2005. Multiple-Choice Tests and Student Understanding: What is the Connection? *Decision Sciences Journal of Innovative Education* (Vol. 3) 73-97.

Snyder, A. 2003. The New CPA Exam-Meeting Today's Challenges. *Journal of Accountancy* (Vol. 196, No. 6) 11-12.

Summerfield, B. 2008. GIAC. The Hands-On Security Certification. *Certification Magazine* (Vol. 10, No. 6) 24-25.

Traub, R. E. 1993. On the Equivalence of the Traits Assessed by Multiple-Choice and Constructed-Response Tests. In *Construction Versus Choice in Cognitive Measurement,* edited by Bennett, R. E. and W.C. Ward. (Hillsdale, NJ: Lawrence Erlbaum) 29-44.

Tsui, J. S. L., T. S. C. Lau, and S. C. C. Font. 1995. Analysis of Accounting Students' Performance on Multiple-Choice Examination Questions: A Cognitive Style Perspective. *Accounting Education: An International Journal* (Vol. 4) 351-358.

Wainer, H., and D. Thissen. 1993. Combining Multiple-Choice and Constructed-Response Test Scores: Toward a Marxist Theory of Test Construction. *Applied Measurement in Education* (Vol. 6) 103-118.

Walstad, W. B., and W. E. Becker. 1994. Achievement Differences on Multiple-Choice and Essay Tests in Economics. *American Economic Review* (Vol. 84, No. 2) 193-197.

_____, and D. Robson. 1997. Differential Item Functioning and Male-Female Differences on Multiple-Choice Tests in Economics. *Journal of Economic Education* (Vol. 28) 155-171.

Zeidner, M. 1987. Essay Versus Multiple-Choice Type Classroom Exams: The Student's Perspective. *Journal of Educational Research* (Vol. 80) 352-358.

Zimmerman, D., and R. Williams. 2003. A New Look at the Influence of Guessing on the Reliability of Multiple-Choice Tests. *Applied Psychological Measurement* (Vol. 27) 357-371.