Using Multiple-choice Tests to Evaluate Students’ Understanding of Accounting

LYNN BIBLE*, MARK G. SIMKIN** and WILLIAM L. KUECHLER**

*Berry College, USA, **University of Nevada, USA

Received: April 2006
Revised: January 2007; July 2007
Accepted: July 2007

ABSTRACT Despite the wide diversity of formats available for constructing class examinations, there are many reasons why both university students and instructors prefer multiple-choice (MC) tests over other examination formats. However, it is not clear whether MC questions assess the same level of students’ understanding as constructed-response questions. This paper examines this debate within the context of teaching accounting classes, and reports the analysis of 238 test scores of students from the USA who were given both MC and short-answer questions on the same examination. The results indicate that the scores on the MC questions (and selected additional variables) explained about two-thirds of the variability in the scores on the short-answer questions, suggesting that MC questions can perform adequate assessment of subject mastery. This paper also provides some caveats in interpreting our results and suggests some extensions to the present work.

KEY WORDS: Assessment, multiple choice testing, constructed response assessment, standardized testing

I. Introduction

College instructors can use a wide variety of in-class test formats—for example, multiple-choice (MC) questions, true-false questions, fill-in-the-blank questions, short-answer questions, computational problems, or essays—for evaluating students’ understanding of key course topics. The principal alternative to a MC examination is a ‘constructed-response’ (CR) test. CR tests differ from MC tests because they require the student to formulate an answer independently, rather than selecting an answer from a set of prescribed alternatives.

There are several reasons why accounting instructors should be concerned with test formats. One is the desire to utilize examinations that accurately and fairly evaluate students’ knowledge. Another is the need to test students’ understanding of course material in sufficient detail to permit the assignment of fair grades. Yet a third is the interest in creating tests that assess students’ mastery of accounting knowledge in an
objective, efficient manner. A fourth reason is the administrative desire for feedback on the effectiveness of instructional methods and pedagogy.

A fifth concern is the importance that accounting instructors attach to the goal of preparing new graduates to enter the accounting profession. This preparation includes providing students with the knowledge necessary to pass professional certification examinations such as (in the USA) the Certified Public Accountant (CPA) and the Certified Management Accountant (CMA) tests. As of April 2004, the Uniform Certified Public Accountants (CPA) Examination is now administered entirely online. (Previously, each part of the CPA examination consisted of approximately 60% MC questions with the remaining 40% consisting of essay-type questions.) The new examination increases the percentage of MC questions on three out of its four parts to 80% (Snyder 2004). The remaining 20% are simulations. This change in test format challenges accounting faculty to look at their own tests—for example, to determine to what extent they use the same format as the CPA and CMA examinations, or to ask whether they disadvantage students by not using the same format.

A number of scholars have expressed an interest in, and conducted experiments to test, the relationship between ‘test format’ and ‘student performance’ in accounting education. Baldwin (1984), for example, recognized the many conveniences of MC questions in assessing students’ understanding of accounting principles, and developed guidelines for developing ‘high quality’ test questions. Further work by Peterson and Reider (2002) observed that a ‘computerized’ test format had a positive affect on the test takers’ perceptions of the test itself. Finally, several scholars have observed the usefulness of testing students repeatedly as an effective examination format (Liebler, 2003; Murphy and Stanga, 1994).

Multiple-choice test formats offer several advantages for both college instructors and students. For instructors, the advantages include (1) machine scoring, volume grading, and easily-computed statistical analyses of test results; (2) ‘shot gun’ coverage of course materials, leading to the ability to cover a wider range of course topics than might otherwise be possible; (3) compatibility with increasingly popular web-based courses; and (4) greater ‘referencing capabilities’ when disputes about the correct answers to specific questions arise, and therefore easier resolution of examination challenges.

For students, the advantages include (1) heightened confidence in the ability to guess the correct answer or to uncover the desired response by a process of elimination; (2) avoiding tests where good writing skills are required, or are allowed to compensate for the lack of factual recall; and (3) the perception that MC tests are more objective (Snyder, 2003; Kennedy and Walstad, 1997; Lumsden and Scott, 1995; Bridgeman, 1992; Gul, Teoh and Shannon, 1992; Zeidner, 1987; Collier and Mehrens, 1985).

MC tests also have potential drawbacks. One concern is whether such tests evaluate the same level of understanding as CR tests. Another concern is possible ambiguity in the MC test questions themselves, which lead to doubts about what students are really answering (see Tsui, Lau and Font, 1995). Perhaps the most important objection to MC tests revolves around the perception that MC tests are unable to measure analytical skills adequately or to evaluate a student’s ability to combine individual facts into a meaningful and literate response. Additionally, a CR examination format provides an opportunity for students to summarize the most important concepts of the course. These integrative skills are particularly important to those accounting faculty interested in preparing students for future employment and professional examinations (Lukhele, Thissen and Wainer, 1994). After all, the problems faced by accounting practitioners more typically require ‘problem solving,’ ‘deduction,’ and ‘integration skills’ than the simplistic recall of definitions or lists.

Even if many university faculty regard CR examinations as a better tool than MC tests for evaluating students’ knowledge, the facts remain that grading CR examinations (1)
requires graders with high subject mastery; (2) is subjective to some extent no matter how much subject mastery is used in the evaluation process; and (3) requires graders to spend more time grading than would be the case with MC examinations. It is also true that most universities do not directly reward faculty for the quality of the examinations given in their classes or for the quality of the grading performed for such examinations, nor do any universities known to the authors penalize faculty for the lack of such qualities. Principally for these reasons, more universities and certification tests, including the CPA examination, are now using MC test formats; because of this trend, accounting faculty have a concomitant interest in answering a simple question: ‘How well do students’ scores on MC tests correlate with scores on CR tests?’ In the light of budgetary constraints, an increasingly competitive educational market, and the high likelihood that MC test results do not correlate perfectly with CR test results, it is also important to ask ‘is the relationship between the two close enough to assure an adequate assessment of students’ understanding of accounting material using just MC tests?’

The second section of this paper describes an empirical investigation performed by the authors using four semesters’ worth of test data from selected intermediate accounting classes. The third section discusses our results and compares them to earlier findings. The fourth section of the paper provides some caveats in interpreting our results and discusses some potential extensions of our work. The last section provides a summary and conclusions.

II. An Empirical Study

Many educators who teach intermediate accounting courses share the belief that CR questions measure a student’s ability to solve real-world problems better than MC questions (Bacon, 2003; Rogers and Hartley, 1999; Hancock, 1994). In educational domains, this belief is reinforced by the lack of ‘structural fidelity’ of MC examinations—i.e. the lack of congruence between the performance required by the test and the performance required in the referent domain (Messick, 1993; Frederiksen, 1984). In the accounting field, this means that the skills required of professional accountants more closely approximate the skills required to answer CR questions than the skills required to answer MC questions.

A fairly extensive body of research has addressed the issue of how well MC questions test students’ understanding, but with mixed results. The theoretical work that impels most MC versus CR research, including this study, comes from the areas of educational psychology and educational assessment (Martinez, 1999; Hancock, 1994, Nunnally and Bernstein, 1994). Early empirical tests have led some scholars to conclude that MC tests and CR tests do in fact measure the same thing (Saunders and Walstad, 1998; Traub, 1993; Wainer and Thissen, 1993; Bennett, Rock and Want, 1991, Bridgeman, 1991). However, a number of recent studies contradict these findings. Using a two-stage least squares estimation procedure, for example, Becker and Johnston (1999) found no relationship between students’ performance on MC and essay questions on economics examinations. Therefore, they concluded that ‘these testing forms measure different dimensions of knowledge’ (Becker and Johnston, 1999). A similar study in physics instruction by Dufresne, Leonard and Gerace (2002) led the authors to conclude that students’ answers to MC questions ‘more often than not, [give] a false indicator of deep conceptual understanding.’

It was not the intention of this study to duplicate these tests, but rather to determine if a means to predict students’ performance on such tests using more tractable MC questions exists. It is this possibility that the present study sought to confirm.

The relative strength of the relationship between students’ performance on MC and CR tests is important. If the relationship is strong enough (a subjective matter), then instructors
can get the best of both worlds—i.e. MC examinations that are easy to grade and assurance that such tests evaluate a sufficient amount of students’ mastery of course materials to assign fair grades. Although the answer to this question has been studied extensively in other disciplines, it has been less well investigated in the accounting field.

Since 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 propose that MC and CR 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 test captures a meaningful amount of the variation in whatever is measured by the other type of test. Given the discussions above, the authors propose the first hypothesis:

\[ H_1: \text{MC test scores capture a substantial amount of the variability of constructed response scores, (i.e. performance on MC questions varies significantly and strongly with performance on CR questions).} \]

According to some research, ‘gender’ may also play a role in determining a student’s ability to perform well on MC tests compared to CR tests. Some studies have shown a possible advantage of males relative to females on MC tests (Bolger and Kelleghan, 1990). Bridgeman and Lewis (1994) estimated this advantage to be about one-third of one standard deviation. Research on this subject, however, has not universally found the same result. 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).

There is little in the accounting education literature regarding gender differentials caused by testing formats. 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) performed a replication of the Gul et al. (1992) study and did not report gender differences, but it is not clear this variable was analyzed. Due to the conflict between the Lumsden and Scott (1995) study and the Gul et al. (1992) study, the authors believe that another look at the gender issue is warranted. The second hypothesis is:

\[ H_2: \text{Gender will not result in statistically significant differences between the scores on MC tests and the scores on CR response tests.} \]

**Methodology**

To test these hypotheses, the authors conducted a study over four semesters to investigate the relationship between the two types of performance measures. The sample consisted of 238 students who had enrolled in the intermediate accounting classes of a 15,000-student state university in the USA. 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 measured students’ understanding of the same material.

All the students in the study were taking the course for credit. In this sample, 41% of the students were female and 59% were male. Following prior experimental designs, CR scores acted as the dependent variable and an ordinary least squares 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.

Three of the examinations that provided data for this study were administered as the normal final examinations for an intermediate accounting course in the college of business. The fourth examination was a mid-term examination. The same instructor administered and graded all examinations, copies of which are available upon request from the authors. The final examinations took place during a two-hour class session. The mid-term examination took place during a one-hour-and-fifteen-minute class session. All examinations were closed book.

As stated above, the two measures of students' understanding in this study were the scores on separate sections of the same examination. One of the three final examinations had 50 MC questions, each worth a single point out of a total of 100 examination points. Two of the three final examinations had 30 MC questions, each worth two points out of 100 points, and a problem-solving/essay questions section worth 40 points. The fourth examination used in the study was a mid-term examination that had 10 MC questions, each worth three points out of 100 points, and a problem-solving/essay section worth 70 points. The number of correct responses for the different types of questions on each examination was scaled to a percentage to permit testing students' performance consistently across semesters.

In the sample tests, each MC question referred to a separate aspect of intermediate accounting and had four possible answers, labeled A to D. A typical MC question is illustrated in Figure 1. Students answered this section of the examination by blackening a square on a Scantron scoring sheet for a particular question.

In contrast to the MC questions in the sample tests, each CR question required a student to record the accounting entries for a specific economic event. Figure 2 illustrates a typical question. Each (CR) problem was worth several points—typically in the range of 5 to 15 points. An obvious concern in grading CR questions is the potential for inconsistency in the evaluation process. In this study, however, there was little subjectivity involved in the grading because all evaluations of students' answers were performed by the same instructor using a pre-constructed grading sheet, which showed the correct calculations and answers possible for each question.

The MC portion appeared first in each examination, which was followed by the problem-solving questions. The cover page of the examination contained test instructions; information about how much each question was worth; and the maximum amount of time available for the examination. In practice, most students began by answering the MC questions, but the 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 foreign students for whom English is a second language (Kuechler and Simkin, 2004). If the weight of total MC questions to

<table>
<thead>
<tr>
<th>A. $60,000 gain</th>
<th>B. $24,000 gain</th>
<th>C. $20,000 loss</th>
<th>D. $16,000 gain</th>
</tr>
</thead>
</table>

Figure 1. An illustration of a multiple choice question
total examination points is high enough, some subjects may focus more attention on the MC questions than the problem-solving/essay (CR) questions, a strategy that might also affect their test performance.

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 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, the authors preferred to perform the 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. A theoretical adjustment of the data points would lower the MC percentage scores, which would have the effect of lowering the variance and the total sum of squares. This would increase the explained variance in the regression model (because \( R^2 = \frac{SSR}{total SS} \)). However, because any adjustment would be arbitrary at our current state of understanding, a more conservative approach was chosen.

**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 using prewritten grading sheets. These sheets indicated 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, it is reiterated that the same functional learning material was covered in all semesters and was tested by both the CR and MC questions in the sample examinations. However, it is also true that the examinations in each semester had different, and

---

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

---

**Example Questions**

(a) Prepare the entry to record the interest expense at 1 April, 2002. Assume that interest payable was credited when the bonds were issued (round to nearest dollar).

(b) Prepare the entry to record the conversion on 1 October, 2002. Use the book value method. Assume that the entry to record amortization of the bond and interest payment has been made.

---

Doyle Co. issued $5,000,000 of 12%, 5-year convertible bonds on 1 December, 2001 for $5,026,000 plus accrued interest. The bonds were dated 1 April, 2001 with interest payable on 1 April and 1 October. Bond premium is amortized each interest period on a straight-line basis. Doyle Co. has a fiscal year-end of 30 September. On 1 October, 2002, $2,500,000 of these bonds was converted into 35,000 shares of $15 par common stock. Accrued interest was paid in cash at the time of conversion.
differently worded, questions, and that the tests were taken by different students. Accordingly, dummy (0–1) variables were added to the regression equation for three of the four semesters to account for these differences. Including these factors therefore served as an important statistical control.

In summary, the regression equation tested in this study used students’ performance on the CR portion of a final or mid-term examination as the dependent variable, and students’ performance on the MC portion of the examination, gender, and semester dummy variables as independent variables. The complete regression equation is, therefore:

\[
\text{CR} = \text{Intercept} + \beta_1 \times \text{MC} + \beta_2 \times \text{Gender} + \beta_3 \times S_1 + \beta_4 \times S_2 + \beta_5 \times S_3
\]

where: \( \text{MC} \) = the percentage score on the multiple-choice portion of an examination; \( \text{Gender} \) = a dummy variable: 0 = male, 1 = female; \( S_k \) = a dummy variable (0–1) to account for the different examination questions given each semester (\( k = 1,2,3 \)).

### III. Results

To test hypothesis 1 we first calculated correlation coefficients. There was a moderately strong positive correlation of CR test scores and MC test scores (\( r = 0.407 \)). While the correlation coefficient indicates that there is a somewhat strong relationship between CR and MC test scores, it does not answer the question of whether MC test scores capture a substantial amount of the variability of constructed response scores. To examine this, an ordinary least squares regression analysis was performed using SPSS Version 13 (see Table 1). The adjusted total R-square value of ‘0.66’ indicates that the independent variables of our model explained approximately two-thirds of the variability of the dependent (CR) variable. This finding compares favorably to similar models tested by Kuechler and Simkin (2004), and noticeably exceeds a comparable value of ‘0.45’ found in that study of information systems students.

The t-statistics for all of the equation regressors were statistically significant (\( P < 0.05 \)) and most were significant at \( P < 0.01 \). The beta coefficient of ‘0.63’ for the MC questions is noteworthy. Its positive sign shows that students’ performance on MC questions is positively correlated with students’ performance on CR questions, and its magnitude of ‘0.63’ means that a student was likely to earn, on average, two-thirds of a percentage point on each CR question for every one percentage point earned on the MC portion of the same test. Thus, the statistical evidence supports our first hypothesis that MC test scores explain a substantial amount of the variability of CR test scores, or, in simple terms, MC questions appear to be potentially acceptable surrogates for CR questions.\(^1\)

#### Table 1. Summary of linear regression results (all semesters combined). \( F = 92.1, \ df = 5, \ P < 0.001, \ \text{adjusted R}^2 = 0.66 \)

<table>
<thead>
<tr>
<th>Intercept or independent variable</th>
<th>Coefficient (( \beta_k ))</th>
<th>t-statistic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>31.28</td>
<td>8.09</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>MC (% correct)</td>
<td>0.63</td>
<td>11.39</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Gender (dummy variable)</td>
<td>4.27</td>
<td>2.20</td>
<td>&lt;0.029</td>
</tr>
<tr>
<td>Semester 1 (dummy variable)</td>
<td>-8.84</td>
<td>3.60</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Semester 2 (dummy variable)</td>
<td>38.93</td>
<td>15.50</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Semester 3 (dummy variable)</td>
<td>28.54</td>
<td>10.16</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
The regression model also enabled us to test hypothesis 2. We found a small, but statistically significant, value for the ‘gender’ variable (see Table 1). Since the underlying dummy variable was coded as ‘0’ for males and ‘1’ for females, the regression coefficient of ‘4.27’ indicates that females have slightly more than a four per cent advantage over males on CR questions. As noted above, this finding is consistent with earlier studies which found similar gender differentials favoring males on MC tests (Hirschfeld, Moore and Brown, 1995; Bridgeman and Lewis, 1994; Bolger and Kelleghan, 1990; Gutek and Bikson, 1985) or females on CR tests (Harris and Kerby, 1997; Lumsden and Scott, 1995). However, this result must be regarded cautiously. Several other studies have found no significant correlation between MC and gender (Chan and Kennedy, 2002; Bacon, 2003; Gul et al., 1992).

Finally, we found that all the coefficients for our semester (dummy) variables in Table 1 were statistically significant. These variables represent semester-to-semester adjustments to the intercept for the regression equation, and their differing values and statistical significance suggests that there was substantial variation in baseline students’ performance on the CR portion of the examinations between the various semesters.

One possible explanation for this is that the examinations for each semester contained different questions. However, the questions were taken from the same test bank, were very nearly the same length (in number of questions), covered the same material, and they were administered at the same point in the course for each semester. Although the courses were taught by the same instructor, a second potential differential is the possibility that the instructor did not teach the same material in exactly the same way. Even small differences in presentation, coverage of course materials, or depth of explanations may account for some of the observed variability. Finally, a third explanation is the fact that different students took the examinations each semester. Such differences in student composition can be considerable from semester to semester, with some classes containing bright, motivated students while other classes contain more passive, less-motivated ones.

**Semester-by-semester Regressions**

The wide variation in the values of the coefficients of the semester dummy variables suggests that a considerable amount of the differences in students’ test performance were attributable to changes in the course from semester to semester. To isolate the possible interactive effects of the differentiating factors caused by a different set of tests, instructor presentations, and students, the authors also performed semester-by-semester regressions using only an intercept, MC, and Gender as explanatory variables. Table 2

<table>
<thead>
<tr>
<th></th>
<th>Semester 1</th>
<th>Semester 2</th>
<th>Semester 3</th>
<th>Semester 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-7.96a</td>
<td>-17.05a</td>
<td>-3.88</td>
<td>48.34a</td>
</tr>
<tr>
<td>MC (% correct)</td>
<td>1.16a</td>
<td>0.76a</td>
<td>0.79</td>
<td>0.33a</td>
</tr>
<tr>
<td>Gender</td>
<td>-2.09</td>
<td>8.43a</td>
<td>1.19</td>
<td>6.84a</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.64</td>
<td>0.44</td>
<td>0.38</td>
<td>0.20</td>
</tr>
<tr>
<td>df, F, P</td>
<td>2/55, 51.5, &lt;0.001</td>
<td>2/61, 26, &lt;0.001</td>
<td>2/37, 13.1, &lt;0.001</td>
<td>2/73, 10.5, &lt;0.001</td>
</tr>
</tbody>
</table>

aStatistically significant at P = 0.05 level.
reports the results of our analyses. The number of students in each semester is equal to the larger of the degrees of freedom shown in the bottom row of Table 2.

As with the results in Table 1, 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, the different number of test questions, possible variations in teaching effectiveness, and the different student compositions in the four classes. Again, the positive sign of the estimate for the MC variable reinforces the intuitive notion that students’ performance on MC and CR questions are related to one another, and again, none of these estimates was statistically differentiable from a desirable coefficient value of ‘1.0.’

Of particular interest were the coefficient estimates for the MC variable, all of which where statistically significant but which ranged in value from ‘1.16’ (for semester 1) to ‘0.33’ (for semester 4). The different magnitudes of these values are mostly attributable to alternate scales—i.e. the different number of MC questions in each examination. For example, because there were 50 MC questions in the first semester’s examination, the coefficient of ‘1.16’ means that a difference of one examination question results in a predicted change of ‘2.32’ percent on the CR portion of that test. In contrast, because there were only 10 MC questions on the test in semester 4, a one-question change would result in a change of 3.33% on the CR portion of that test—a value that differs from ‘2.32’ but not by a factor of nearly four as suggested by the statistical results in Table 1. The decreasing number of MC questions relative to the total number of test questions in successive semesters (1–4) probably also explains the decreasing explanatory power (i.e. the R-square values) of the models themselves.

Finally, we note that the adjusted R-squared for the full regression including all semesters (in Table 1) is higher than for any of the semesters individually (Table 2). The higher R-squared value found there is to be expected due to the larger variance in the full (all semester) data set. That is, as noted above, there are substantial differences between the mean scores in different semesters.

**Discussion**

How do the results shown here compare to those of earlier studies? In terms of the explanatory power of the main model as reported in Table 1, better than some and worse than others. Research that supports the notion that MC examinations are effective surrogates for CR examinations report correlation coefficients ranging from 0.47 for music to 0.84 for chemistry (Wainer and Thissen, 1993), 0.64 for macro-economics and 0.69 for micro-economics (Walstad and Becker, 1994) and 0.71 for marketing (Bacon, 2003). Our correlation coefficient of 0.41 places us on the low end of the scale. However, in terms of the explanatory power of the main model as reported in Table 1, our R-square of 0.66 compares favorably with other studies. Becker, Highsmith, Kennedy and Walstad (1991) reported R-squares ranging from 0.49 in English to 0.86 for Chemistry. Kuechler and Simkin reported an R-square of 0.45 in an introductory computer-programming course. Both of these studies concluded that MC and CR examinations measure essentially the same knowledge levels.

The statistical significance of the gender variable is noteworthy here. Although gender differences may be the result of such mediating variables as ‘differential skill sets’ which has been suggested in research conducted with high school students (see Hamilton, 1999; Gallagher and De Lisi, 1994), research that examines college students in courses within their major have found no gender differences (see Bacon, 2003; Gul et al., 1992). Therefore, because all of the participants in our study self-selected the course, or at least the field of study, gender differentials should be smaller than for the population at
large. This is precisely the effect shown in the supplementary regressions of Table 2—the gender variable was significant in only two of the four semesters.

It is also noteworthy that many of the prior studies studying gender effects took place in such qualitative disciplines as history, English, or economics, and that the gender effects found between the alternate modes of assessment (MC versus CR) used essay questions in which verbal arguments or elaborate verbal descriptions were required (see, for example, Chan and Kennedy, 2002; Walstad and Robson, 1997; Lumsden and Scott, 1995). In contrast, the authors feel that the CR section of the tests used here were less sensitive to natural-language abilities than the CR questions in these alternate studies. Thus, the well-documented verbal and argument-construction advantages of female students should not make as large a differential contribution to test score.

Finally, the authors feel that the statistical variations in our results from semester to semester are particularly noteworthy. Most prior studies have limited their investigations to individual examinations, which of course are taken by a single set of students. Our longitudinal study found substantial differences from semester to semester, suggesting that a multitude of other factors can affect the strength of the connection between students’ performance on MC questions and CR questions. While troublesome, this finding also suggests a rich avenue for further study.

IV. Caveats and Extensions

The results present here should be interpreted with care. One caveat is that the sample was limited to the students taking intermediate accounting classes at a single institution in the USA. Since these students had chosen accounting as a major, a selection bias may result in more uniform populations than would be found in entry-level history or economics courses, which are often required of all students at a university. The variation in the results in our semester-by-semester analyses also suggests that repeated samplings at other institutions would be desirable.

A further consideration when comparing the results of different studies of the type discussed here is the nature of the CR questions themselves. Consider, for example, an essay question such as the following:

Using the concept of social utility, discuss the net value to society of the child labor laws enacted in America at the beginning of the 20th century. (20 points)

This question supports a wide range of responses, from the conventional to the novel. For typical accounting CR questions where a test-question’s answer can be more easily judged as either ‘functional’ or ‘non functional,’ such factors as novelty, English language fluency, and knowledge of external factors usually play smaller roles, thus making comparisons with tests that either directly or indirectly reward such peripheral matters difficult.

A third caveat involves the reliability of the grading in the CR portion of our sample tests. While ideally the tests would have been scored by several graders and their reliability assessed with metrics such as Cohen’s Kappa, the authors stress that the CR questions used in this study left little room for subjectivity. The question tasks, such as preparing journal entries, are highly constrained in form, and absolutely objective in numeric content. Further, the grader worked from prepared grading sheet in all cases, further minimizing grading variance. The MC portion of the examination, of course, was machine graded.

A final caveat stems from the implicit assumption the authors have observed in many earlier studies that MC questions are themselves homogeneous in content, scope, and
measurement accuracy, or are consistent in any other way. For many test banks, this assumption is patently false. The type of MC questions used in investigative studies is important simply because these questions examine different cognitive levels of understanding (Simkin and Kuechler, 2004).

**Study Extensions**

Further study could extend the work begun here in several different ways. Many universities offer the introductory accounting classes in large lecture halls that can accommodate up to 250 students and most business schools require all business majors to take such courses. Therefore, a wide variety of student interests would be present, along with less motivation to do well in the class. Repeating the study in such a setting would help determine the significance of ‘self-selection’ in the current study.

Many educational researchers note the difficulty of simple extrapolation of research results in one area of study to other areas due to the differences in abilities and cognitive skills required in different contexts (Modell, 2005; Anderson and Krathwohl, 2001; Walstad and Becker, 1994; Traub, 1993). Unlike financial accounting and auditing which focus on generally accepted accounting principles and/or auditing standards, for example, the tax area concentrates on code law, which may require a different set of skills or knowledge. Thus replication of this study in other accounting areas such as the tax area would not only have the potential to provide some additional insight into the MC/CR question debate but would also strengthen the application of results across the accounting field (Modell, 2005).

Finally, it is noted that the accounting examinations used in the present study employed a mix of questions that ranged widely in quality and scope. However, it is unknown whether such mixes measure students’ understanding better than, say, MC questions that focus entirely on (say) problem-solving concepts. A number of widely accepted taxonomies of learning (cf. Bloom, 1956) propose that different levels of understanding can be achieved across subject areas, and these can be tapped by carefully selected questions. However, no prior research of which the authors are aware has treated MC questions as anything other than a one-dimensional construct. Thus, empirically testing whether or not students’ performance on selected types of MC questions will correlate more closely with performance on CR questions represents an interesting avenue for future work.

**V. Summary and Conclusions**

There are various reasons why students and instructors may prefer MC tests to CR tests. However, some instructors believe that CR tests examine a higher level of cognitive understanding than do MC tests. This study investigated how closely MC questions and CR questions were related as evaluators for the 238 students who had enrolled in a junior-level, intermediate accounting class. CR test results were used as the dependent variable in a multiple linear regression analysis the independent variables in which included scores on a separate MC portion of the test, a dummy variable for ‘gender,’ and separate dummy variables for the semester in which the test was taken. The coefficients of the independent variables were all statistically different from zero, suggesting in particular that MC questions may be adequate surrogates for CR questions on students’ examinations. Taken as a whole, the independent variables explained almost two-thirds of the variability of the CR variable.

The answer to our initial question of whether the relationship between students’ performance in MC questions and CR questions is close enough to enable faculty to
rely on MC examinations to evaluate students’ understanding has both an objective and subjective component. Objectively, psychometric theory proposes that the reliability and validity of MC tests can be measured (Nunnally and Bernstein, 1994). Ultimately, however, the answer to the question of whether properly constructed MC tests can be substituted for CR tests rests upon a subjective interpretation of the term ‘close enough.’

Within the confines of our study, the relatively high $R^2$ values and statistically significant findings for the MC beta coefficient in our linear regressions suggest that students’ performance on MC questions is positively related to students’ performance on CR tests. This is encouraging news for institutions that, due to budget constraints, are resorting to larger class sizes and computerized examinations, and is important to professional examination administrators whose goal is to determine if an applicant has the necessary skills to warrant certification into the profession.

Notes
1. The $t$-statistic for the MC coefficient is for a ‘different-from-zero’ test. Also of interest to the researchers was a ‘different-from-one’ test—i.e. a test of the ‘ideal’ relationship between students’ performance on MC and CR questions. In particular, we would wish to reject our first hypothesis (H1) if the value of this coefficient differed statistically from ‘1.’ Recomputing the $t$-statistic for this second test resulted in a new $t$-value of ‘0.59,’ with associated $P$-value of 0.22.
2. The beta value of ‘4.27’ in Table 1 does not mean that females will on average score 4.27 points better on an examination, as this depends upon the relative proportion of CR questions to MC questions in a specific test’s composition.
3. To further investigate the effects of gender, we also regressed gender against MC directly and found that gender was not significant ($P = 0.65$), as desired for an independent variable in our initial model. We also regressed gender against MC and all combinations of ‘semester’ in search of potential interaction effects. Again, we found no significant results ($P$-values $= 0.295$ for semester 1; 0.576 for semester 2; 0.499 for semester 3; and 0.669 for semester 3). We are inclined to treat these later regressions as more meaningful than the ‘all variables’ regression. However, we also note that, even when significant, regression results must be interpreted in context.

References


