How Well Do Multiple Choice Tests Evaluate Student Understanding in Computer Programming Classes?

William L. Kuechler
Mark G. Simkin
College of Business Administration/026
University of Nevada
Reno, Nevada 89557
kuechler@unr.nevada.edu Simkin@unr.edu

ABSTRACT

Despite the wide diversity of formats with which to construct class examinations, there are many reasons why both university students and instructors prefer multiple-choice tests over other types of exam questions. The purpose of the present study was to examine this multiple-choice/constructed-response debate within the context of teaching computer programming classes. This paper reports the analysis of over 150 test scores of students who were given both multiple-choice and short-answer questions on the same midterm examination. We found that, while student performance on these different types of questions was statistically correlated, the scores on the coding questions explained less than half the variability in the scores on the multiple choice questions. Gender, graduate status, and university major were not significant. This paper also provides some caveats in interpreting our results, suggests some extensions to the present work, and perhaps most importantly in light of the uncovered weak statistical relationship, addresses the question of whether multiple-choice tests are "good enough."

Keywords: Multiple-Choice Versus Essay Tests, Computer Programming Education, Test Formats, Student Test Performance

1. INTRODUCTION

College instructors can use a wide variety of test formats for evaluating student understanding of key course topics, including multiple-choice (MC) questions, true-false, fill-in-the-blank, short answer, coding exercises, and essays. Other format choices include take home tests and oral examinations. Most of the alternatives to MC and true-false questions are described in the literature as constructed-response questions, meaning that they require students to create their own answers rather than select the correct one from a list of prewritten alternatives.

Despite the wide diversity of such test formats, there are many reasons why both students and instructors prefer multiple-choice tests over other types of examination questions. However, the literature contains a considerable body of analysis and debate over this issue. The purpose of the present study was to examine this question within the context of teaching computer programming classes. In such classes, short answer questions, and (especially) coding questions using a particular computer language are common. Among other objectives, our purpose was to answer the question "are MC tests able to effectively test student understanding of programming concepts using these alternate evaluators? "

The next section of this paper examines this debate in greater detail and highlights some of the issues involving MC-testing in particular. The third section of the paper describes an empirical investigation performed by the authors using four semesters worth of test data from an entry-level Visual Basic programming class. The fourth section of this paper discusses our results, compares them to earlier findings, and provides some caveats in interpreting our analyses. We close with a summary and conclusions.

2. ADVANTAGES AND CONCERNS OF MULTIPLE-CHOICE TESTS

There are many reasons why instructors like multiple-choice tests. Perhaps foremost among them is the fact that such tests can be machine-graded—an advantage of special importance in mass lecture classes where volume grading is required. But such tests can also help control cheating, because the instructor can create multiple versions of the same test with either different questions or different orderings of the same questions. Then, too, given the time

constraints typically imposed upon instructors to give examinations that can be completed in relatively short periods of time, such tests enable instructors to ask questions that cover a wider range of material, and to ask more such questions than, say, essay tests (Bridgeman and Lewis, 1994, Saunders and Walstad, 1990).

Machine-graded tests have several additional advantages over other types of tests. One benefit is the ability to capture test results in machine-readable formats at the same time the tests themselves are taken, thus eliminating time-consuming data transcription and facilitating the record keeping process. This advantage has been especially useful in web-based classes, which commonly use such web software as PageOut, WebCt, or Blackboard to test students online. Then, too, MC formats enable computer programs to create question-by-question summaries as well as perform sophisticated statistical analyses of test results. In computer programming classes, it is also possible to argue that it only makes sense to use computers to grade questions in courses that cover concepts *about* computers.

Another oft-mentioned advantage of MC tests is their perceived objectivity (Kreig and Uyars, 2001, Zeidner, 1987). This perception stems from the belief that each question on a MC test has exactly one right answer, which a student either does or does not identify during the examination period. Questions that are based upon specific textbook chapters and perhaps drawn from prewritten test banks enhance this perception of objectivity by providing easy references for both students and instructors in case of disagreement about correct answers. This referencing characteristic is no small advantage to those administrators called to investigate student challenges to examinations or the course grades based upon such tests, and is therefore also important to universities in our litigious society.

Both students and instructors appear to dislike essay tests or similar types of constructed-response examinations, although for different reasons. Many students do not like essays because they require higher levels of organizational skills to frame cogent answers, higher levels of recall about the subject matters, more integrative knowledge, and of course, good writing skills (Zeidner, 1987). Then, too (and unlike MC tests), essay examinations do not preclude a student's saying things that are just plain wrong, and for which an instructor may feel compelled to deduct points—a common event on constructed-response examinations. Finally, although student test-format preferences may not have been too important at one time, "customer satisfaction" appears to be an increasing concern to the university administrators of both public and private institutions of higher learning.

If the majority of both students and faculty prefer MC tests over other test formats,[1] it remains less clear whether multiple-choice questions examine the same student cognitive understanding of class materials, or do so as well as constructed-response tests. Most faculty objections to MC tests center on a perceived inability of MC questions to measure problem solving (analytic) ability or to determine whether or not the student can actually synthesize working, productive output from a multiplicity of memorized factoids. This is most significant for faculty who perceive the university as having a certification responsibility to the potential employers of its graduates.

Gender also appears to play a role in answering the question "how fair are multiple-choice tests compared to constructed-answer tests in evaluating student understanding of course materials." Past research suggests that males may have a relative advantage on MC tests (Bell and Hay, 1987; Lumsden and Scott, 1987; Bolger and Kellaghan, 1990; Mazzeo, Schmitt, and Bleistein, 1995). Bridgemand and Lewis (1994) estimated this male advantage at about one-third of a standard deviation, but these results have not been universally replicated. For example, several more-recent studies have found no significant gender differences among economic tests on fixed-response tests (Walstad and Becker, 1994; Greend, 1997; O'Neill, 2001, and Chan and Kennedy, 2002). It is also not clear whether this potential gender advantage (or lack of it) applies to students taking computer classes.

If most university instructors have a higher regard for constructed-response tests than they do for MC tests, the fact remains that graders with high levels of domain expertise must evaluate the questions on them—a more time-consuming task than grading MC questions, and (in the opinion of many scholars and students) a task requiring greater subjectivity (Zeidner, 1987). It is for precisely for this reason that, for the past few years, scholars have attempted to answer the question "how well do the results of student scores on MC tests correlate with scores on constructed-response tests." If the MC-constructed-response relationship is "high enough," then instructors may be able to get the best of both worlds with just MC questions—e.g., exams that are easy to grade and that evaluate a sufficient amount of student mastery of class materials to assign each one a fair grade. In view of budgetary constraints and the increasingly competitive market in which university teaching takes place, it is also important to ask "if student performance on MC tests does not perfectly correlate with that of constructed response tests, is the relationship good enough?"

3. AN EMPIRICAL STUDY

Although the relationship between student performance on MC questions and constructed response questions has been addressed extensively in such disciplines as economics, the issue has not been studied as well in technical fields such as computing. Accordingly, the authors were interested in exploring the correlation between multiple-choice and constructed response examination questions in the computer language training venue. Given the literature review above, we designed our study to test two specific hypotheses:

H1: Multiple choice coding scores capture a substantial amount of the variability of the constructed response scores (i.e., MC scores correlate significantly and strongly with coding scores), and

H2: Gender differences lead to statistically significant score differences on multiple choice tests.

3.1 Methodology

To test these hypotheses, the authors conducted a study over four semesters (two years) to investigate the relative effectiveness of two types of measures of computer language learning: (1) multiple choice question sets and (2) coding (constructed response) question sets. Because the influence of such demographic variables on learning as "gender" and "choice of major" have been identified in prior research as potentially important, these variables were also investigated to see if they significantly affected student performance on programming tests.

The study proceeded as follows. The sample included the 152 students who had enrolled in an introductory Visual Basic class that was taught in the college of business of a 15,000-student state university. All the students in the study were taking the course for credit. Of these students, approximately 50 percent were IS majors or IS minors for whom this course was required. The remaining students were from other majors in the college or university who were taking the course as an elective. In this sample, 38 percent of the students in the sample were female and 62 percent were male.

Each student in each class section was given the same type of midterm test (treatment) once during the semester. These examinations were administered as the normal midterm exams for an introductory Visual Basic course and took place during a one hour and fifteen minute class session. All exams were closed-book, although students were permitted to use notes, homework and class handouts. The first section of each test consisted of multiple choice questions and the second section required students to write short segments of Visual Basic code (the coding section).

In the sample tests, each MC question referred to a separate aspect of program development using Visual Basic and had four or five possible answers, labeled A through E. Figure 1 illustrates a typical MC question. Students answered this section of the exam by blackening a square for a particular question on a Scantron scoring sheet.

In contrast to the MC questions in the sample tests, each constructed-response ("coding") question required a student to create coding segments that accomplished a specific task. Figure 2 illustrates a typical question. The questions, as well as the written English instructions, referenced screen captures (illustrations) that were reproduced from the integrated development environment (IDE) of Visual Basic. Several coding questions may have referenced the same screen figure, but some questions did not directly refer to them. Each coding question was worth several points—typically in the range of 5 to 8 points.

A form shows the term "Main Form" in its banner at its top. This text can be set using the form's:
A. Top property
B. Banner property
C. Caption property
D. Text property
E. None of these

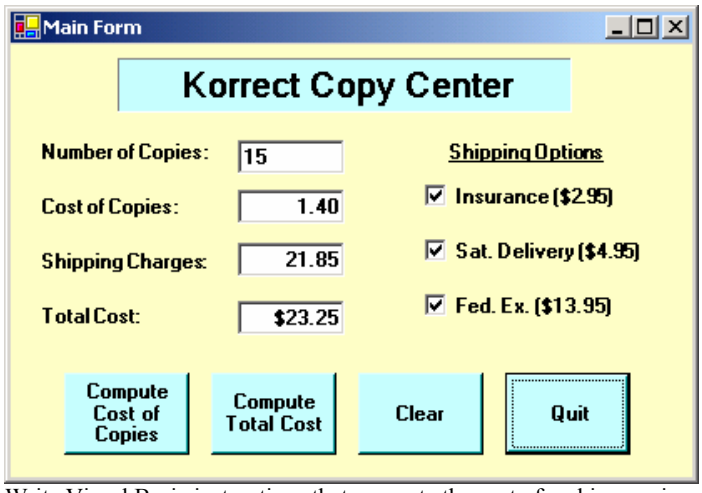
Figure 1. A typical multiple choice question

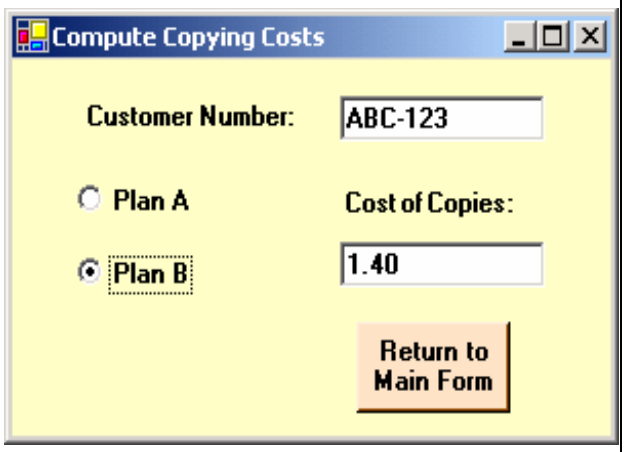
At the beginning of each examination period, the instructor in charge made clear that there were two parts to the test, that each part was worth so many points, that there was no penalty for wrong answers on the multiple choice portion of the examination, and that students should budget their time in order to complete the entire examination within the allotted time. In practice, most students answered the multiple choice questions first, but a few "worked backwards" and began with the coding questions. This alternate test-taking strategy was notable for those students who expressed preferences for constructed-response questions, as well as for some (but not all) of the foreign students for whom English was a second language and who sometimes had difficulty dissecting the wording nuances of selected MC questions.

Most students were able to complete the entire examination in the time allotted. Due to the higher point weighting, per question, it is possible that subjects focused more attention on the coding questions than on individual multiple choice questions. Also, because screen captures 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 multiple choice and constructed response questions tap into different learning constructs (Becker and Johnson, 1999). However, the intent of this research was to determine the degree to which ***whatever is measured by one type of question captures a meaningful amount of the variation in whatever is measured by the other type of question.***

3.2 Dependent Variable

Following prior experimental designs, multiple-choice scores acted as the dependent variable and a multiple linear regression model was used to determine the degree to which coding scores and selected demographic variables (described in greater detail below) were useful predictors of performance on the multiple-choice portion of the exam. Three of the four examinations had 50 multiple-choice





Write Visual Basic instructions that compute the cost of making copies at a duplicating shop. The number of copies to make is stored in a Textbox named "txtNumber" and the cost is based on the following schedule. For Plan A, the cost of 15 copies is $1.20 (=.08 x 15 copies).

Number of copies:	10 or less	11-100	over 100
Charge each:	10 cents	8 cents	5 cents

Figure 2. Typical coding question

questions, each worth a single point out of 100 points. In the final semester of the study, the midterm examination had 40 MC questions, each worth a single point out of 100 points, and a coding section worth 60 points. Accordingly, the data for all semesters were scaled to percentages for consistency

3.3 Independent Variables

From the standpoint of this investigation, the most important independent variable was the student's score on the coding (constructed-response) portion of the examination. As illustrated in Figure 2 above, these questions required students to create coding segments that would enable a computer to perform a stated task. Students were permitted to use their notes, class handouts, and prior homework to help them create the instructions for these tasks, but were not allowed to actually test their answers on a computer. (In the opinion of the authors, this format better examines what students know, rather than what they can experimentally learn with the help of a computer, during the examination period.)

To ensure consistency for the coding answers, the same instructor manually graded all the coding questions on all the examinations using prewritten grading sheets. These sheets indicated the correct answer(s) to each coding question as well as a list of penalties to assess for common errors. In a few instances, a student's answer for a given question used programming instructions that were not covered during class time. In each such instance, the coding answer was keyed into a small computer program and tested for accuracy. (In the vast majority of cases, the code failed, but in a few instances, the instructor learned something!)

The same functional learning material—e.g., coding for such events as button clicks—was covered in all semesters and therefore tested in the coding questions in the midterm examinations. However, because the examinations had different, and differently-worded, coding questions in each successive semester in which the study was conducted, dummy (0-1) variables were added to the regression equation for each semester's examination. The authors were especially concerned about differences between the most recent semester and prior semesters because the most significant material change, from VB version 6 to VB.NET, took place between those semesters.

Perhaps the second-most-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" (Harrison and Rainer, 2001; Heinssen, 1987; Gutek and Bikson, 1985). However, it is also known that gender differences may be the result of such mediating variables as "computer related anxiety" or "differential skill sets." Because all of the participants in this study self-selected the course or at least the field of study, gender differentials may be less significant for the study group than the population at large.

Also, many of the studies finding gender effects between modes of assessment (multiple-choice vs. constructed-

response) used essay questions in which verbal arguments were set forth or elaborate verbal descriptions constructed in qualitative disciplines such as history, English or economics (Walsted and Robeson, 1997; Tobias, 1992).

In contrast, the coding section of the instrument used in this study was much less sensitive to natural-language abilities than these studies. Indeed, as discussed in the following section of this paper, the coding questions assess functional knowledge far more than natural-language fluency. Thus, the well documented verbal and argument construction advantage of female students may not make a differential contribution to the coding portion of the test score. Potentially more significant for this study are gender differences in cognitive tendencies, which could have an effect on measurement issues directly (Perrig and Kintsch, 1988).

The other independent variables used in this study represented selected demographic variables. They include: (1) a dummy variable for whether or not a student was a graduate student (Grad)—a possible surrogate for maturity, (2) a dummy variable for IS major (IS)—the programming course was required for IS majors but was an elective for non-IS majors, and (3) dummy variables for three of the four semesters of the study as described above.

3.4 Outliers

Before estimating the coefficients for the regression equation, the authors created the scatter diagram shown in Figure 3. In this plot of the MC variable against the coding variable, six data points become immediately apparent as outliers. The figure suggests that these data points have a similar slope to the best-fit line for the remainder of the data, yet are noticeably below the majority of the sample observations. That is, treated as a separate group, they would have a very similar regression coefficient as the main portion of the data but a different (lower) Y-axis intercept. We examined these points in detail, and found that they belonged to the lowest-performing students in our CIS program. We thus created a special dummy variable, Low Performance (LP), for them, assigning it a value of "1" for each of the six outlying observations and "0" for the remaining data.

In Figure 3, note that all the LP data points have higher MC scores than coding scores, which we attributed to the ability of the test takers to guess the correct answers to the multiple-choice questions. Such guessing is virtually impossible on the coding questions involved in this study, given the nature of the tasks at hand. This is important distinction between this study's experimental venue and those of the more traditional tests containing essay questions, where the ability to verbally "obfuscate under uncertainty" is both possible and potentially advantageous (Becker and Johnson, 1999).

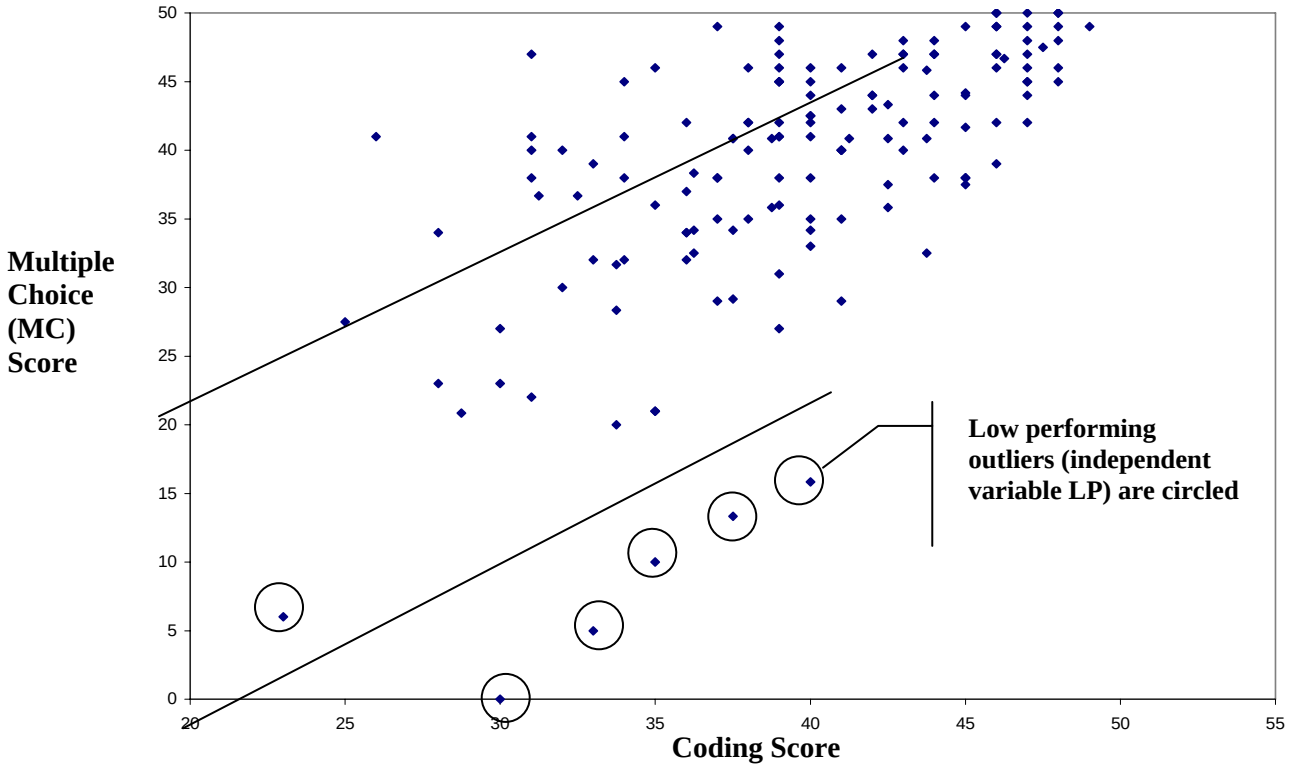


Figure 3. Scatter diagram for student examinations, showing outliers

In summary, the regression equation tested in this study used student performance on the multiple choice portion of a midterm examination as the dependent variable, and student performance on the constructed response portion as the major independent variable. In addition, the study included a number of dummy variables for such factors as "gender," "graduate status," and "IS major" as described above. Figure 4 lists all the independent variables and their descriptions.

Dependent Variable	Description
Coding	The score on the coding portion of the exam.
Gender	A dummy variable: 0 = male, 1 = female
Grad	A dummy variable: 1 = graduate student, 0 = undergraduate student
CISMaj	A dummy variable for required course (1) or elective (0)
LP	A dummy variable indicating significantly lower than average performance.
Semester (F_01, etc.)	A dummy variable to account for the different examinations given each semester.

Figure 4. Independent variables for the regression model.

4. RESULTS

To analyze the initial data sample, the authors computed the Pearson correlation coefficients shown in Figure 5. This table also includes the mean, standard deviation and range for each of the three interval variables, Coding, MC and Total. P-values are shown in parentheses only for those correlations that are significant at or better than a $p < .05$ level.

In our results, the only variables that correlated significantly with MC were Coding and F_02—the dummy variable for the semester in which the class switched from VB6 to VB.Net. However, because gender, graduate student ranking, and required-course have been shown to be potentially significant factors in earlier studies, the regression model included these variables as well.

The results of the regression analysis described above are shown in Figure 6. The t-statistics for the Intercept, Coding, LP, and F_02 coefficients were all highly significant ($p < .001$). Recall that the fall semester of 2002 (F_02) was the semester in which the switch was made from VB 6 to VB.NET. None of the other variables contributed significantly to the model. The residuals for all variables for the model were normally distributed, had constant variance, and were independent, strongly supporting the applicability of the linear regression model.

The positive sign of the intercept in the results for this equation is important. Its value of "20.91" means that, even if a student earned no points for the coding portion of the examination, he or she would have earned almost 21 points on the MC portion of the examination. This finding echoes earlier studies suggesting that MC formats enable test takers to better guess correct answers compared to constructed response tests.

The estimated value of ".50" for the coefficient of the coding variable is also noteworthy. Its positive sign suggests that students who do well on the constructed-response portion of the test will also do well on the MC portion of the test. Given a student's ability to guess on the MC questions, however, it is surprising that this value is not greater than one. A larger value would mean that students who did well on the more-challenging coding portion of the test would do even better on the MC portion of the test. The absence of such a finding reinforces the claims found in the literature that MC and constructed-response questions probably test different cognitive processes. There is also the possibility that those students who do well on constructed-response questions sometimes read too much into MC questions, thereby confusing themselves and (as they sometimes report in the aftermath of such examinations) "talk themselves out of the correct answers."

4.1 Discussion

The fact that the coding measure (along with LP and several additional demographic variables) was able to explain only 45% of the variability of the multiple-choice measure is disappointing, and is low in comparison with the R-square statistics reported in similar studies. Again, one possible explanation for the low value found here is that it supports the belief that the two measures tap into different constructs, and that neither is an adequate substitute for the other. This is contrary to our first hypothesis, and also conflicts with several findings in the field of economics education, where up to 69% of the variation in constructed response was accounted for by multiple-choice measures (Walsted and Becker, 1994, p. 196; Bridgeman and Rock, 1993, p. 326).

The differences in findings in this study and those in the economics literature can be partially explained by considering the difference in the nature of the constructed-response questions for economics (and also English and history) and those for this study. For example, Figure 7 is a typical economics essay question that was found on the midterm exam for a junior level economics class at the authors' school. Notice that the skill set required to successfully answer this question includes (1) a recall of facts, (2) English language fluency and (3) rhetoric. In contrast, the constructed response (coding) questions for our study are typical of those used in computer language examinations, where competence in the programming language itself is much more valuable than fluency in

Range	29-98	23-49	0-50								
Mean	78.77	39.73	39.04								
Std. Dev.	13.68	5.59	9.46								
	Total	MC	Coding	LP	S_03	F_01	Gender	Grad	CISmaj	F_03	S_02
Total	1.000	(p<.001) 0.841	(p<.001) 0.947	(p=.04) -0.180	(p=.034) -0.183	-0.013	-0.047	0.064	0.062	-0.098	0.114
MC		1.000	(p<.001) 0.622	-0.038	-0.102	0.105	-0.066	0.048	0.089	(p=.025) -0.194	0.148
Coding			1.000	(p=.006) -0.237	(p=.018) -0.204	-0.081	-0.030	0.064	0.037	-0.026	0.077
LP				1.000	0.146	-0.042	-0.068	-0.017	0.085	-0.060	-0.052
S_03					1.000	-0.241	0.094	0.062	-0.026	(p<.001) -0.409	(p<.001) -0.354
F_01						1.000	-0.060	0.007	0.050	(p=.004) -0.247	(p=.005) -0.241
Gender							1.000	0.008	-0.150	0.054	-0.081
Grad								1.000	(p=.021) -0.200	-0.051	-0.027
CISmaj									1.000	-0.058	0.076
F_03										1.000	(p<.001) -0.409
S_02											1.000

Figure 5. Matrix of Pearson Correlations

	Coefficient (β)	t-statistic	p-value	Adjusted R^2
Intercept	20.91	9.76	<.001	0.449
Coding	0.50	9.95	<.001	
Gender	0.07	.09	0.93	
LP	8.97	3.78	<.001	
Grad	0.38	0.20	0.85	
CIS Major	0.21	0.29	0.77	
F_02	-2.30	-2.42	0.02	
S_03	-0.38	-0.37	0.71	
F_01	-1.61	-1.65	0.10	

Figure 6. Regression Results

English. Further, no construction of an extended argument in English is required, and more importantly, no points were awarded for fluent or cogent arguments.

A further consideration is that fact that the answers to essay questions such as the example in Figure 7 support a wide range of responses, from the conventional to the novel. Creativity (within limits) is also sometimes valued and rewarded in responses to such questions. For program coding questions, especially at the introductory level, there is far less room for such creativity, which is usually not valued: an answer is either functional or non-functional. The implication is that English language fluency is less a factor in answering constructed-response questions for programming classes. This observation is also supported by inspection of the data set, which shows that foreign students typically score better in the coding section of exams than on the multiple choice sections (where a firmer grasp of English rhetoric probably helps).

Our second hypothesis, that gender would be a significant factor as demonstrated in multiple studies in non-IS fields, was also not supported by our study results. The difference in constructed response question type between IS and non-IS fields of study may account for the inability to replicate gender significance. In reconciling this finding with earlier studies, we note that, even in those studies where gender was significant, it accounted for relatively little score variance.

> Using the concept of Social Utility and Pareto's distinction between the maximum utility of and the maximum utility for a community, discuss the net value to society of the child labor laws enacted in America at the beginning of the 20th century. (20 points)

Figure 7. A sample constructed response (essay) question for an economics examination

4.2 Caveats and Extensions

Our results must be interpreted with care. One difficulty is that our sample was limited to the students taking a specific, entry-level, procedural programming language class in Visual Basic. While all the students taking this class were included in the study, our sample observations are hardly random, and in fact were self selective for non-IS majors. Further, because many students take this class from other disciplines—e.g., computer science, accounting, or even graphic arts—our sample was probably less homogeneous than if we had collected similar data from upper-level CIS-only classes. This self-selection caveat may also explain the lack of statistical significance of the "gender" variable.

As noted above, a second reservation concerns the degree to which the coding questions used in this study parallel the essay questions often used in the tests of other, more-qualitative disciplines. In particular, the programming constructs required in the tests of this study required very specific domain knowledge and such requirements as "trends," "integration of separate themes," or similar tasks had very little place in them. Similarly, while it is easy to classify the coding portions of our midterm examinations as those requiring "constructed responses," it is quite another to claim that such coding questions require the same cognitive skills and understanding as, say, the "compare-and-contrast" questions often found on economics or history tests. Thus, comparisons between our coding questions and the constructed response questions used in many prior investigations are problematic.

A third caveat stems from the implicit assumption the authors have observed in many earlier studies that multiple-choice questions are themselves homogeneous in content, scope, and measurement accuracy, or are consistent in any other way. Indeed, it seems almost heroic to argue that a MC question that tests knowledge about a definition, for example, is equivalent to a MC question that tests understanding of a fundamental principle. In the present study, in fact, the midterm examinations used a mix of questions that ranged widely in quality and composition.

Bloom's widely-accepted taxonomy of learning (Bloom, 1956) proposes that different levels of understanding can be achieved across subject areas. Differently constructed questions of either multiple choice or constructed response types have the potential to test understanding at any of these levels. However, none of the prior research of which we are aware of has treated multiple-choice questions as anything other than a unidimensional construct. Thus, it is not known whether student performance on selected types of MC questions will correlate more closely with performance on constructed-response questions, but a priori, this seems possible. This issue represents a particularly promising avenue for further inquiry.

A number of additional efforts could extend the work begun here. One possibility would be to include alternate variables with which to explain the variations in student scores on MC tests. Examples might include native-language ability, level of mathematical skills, choice of undergraduate major, or course grades in prerequisite classes. We hasten to add that greater correlations between these variables and student performance on MC tests might increase the explanatory power of our linear regression models, although they would not strengthen the argument that MC and constructed response examinations necessarily test the same level of student understanding of underlying course materials.

A second avenue for extending our work is to replicate our study in more advanced programming instruction, and/or to other courses within the IS area such as database design or systems analysis and design classes. Indeed, the more advanced the class in a given subject area, the more likely it is that higher levels of synthesis, integration, and cognitive competency are required. But the fundamental question—can MC questions accurately, or at least adequately, test these abilities in such advanced-level classes—remains.

5. SUMMARY AND CONCONLUSIONS

There are many reasons why both students and instructors express preferences for multiple choice tests over constructed-response tests. Among them are (1) the advantages of machine scoring, volume grading, and easily-computed statistical analyses of test results, (2) heightened student confidence in the ability to guess the correct answer, (3) student aversion to tests where good writing skills compensate for the lack of factual recall, (4) "shot gun" coverage of course materials, leading to the ability to cover a wider range of course topics that might otherwise be possible, (5) the perception that MC tests are more objective, (6) compatibility with increasingly popular web-based courses, (7) higher "referencing capabilities" when tests disputes arise, and therefore easier resolution of test-related grading problems, and (8) the more labor-intensive tasks inherent in grading constructed-response examinations.

Most instructors believe that constructed-response tests examine a higher level of cognitive reasoning than do multiple choice tests. Then too, a subset of earlier research suggests that MC questions involve a small gender bias in favor of male students. The present study attempted to investigate how closely MC questions and constructed response questions were related as evaluators for the 152

students who had enrolled in an entry-level computer programming class. MC test results was used as the dependent variable in a regression analysis whose independent variables included scores on a separate coding ("constructed-response") portion of the test, gender, graduate status, major (versus non-major) status, low-performance student (a dummy variable), and semester the class was taken (three dummy variables). In the results of the regression analysis, only the intercept, coding questions, a dummy variable for "low-performance," and a dummy variable for a particular semester were statistically significant. But all these variables together were able to explain less than half of the total variation in performance on the MC portion of the tests.

Lastly, we turn to the question posed at the beginning of this paper: "If the relationship between student performance in MC questions and constructed response questions is imperfect, is it still close enough to enable faculty to rely on *only* MC tests when evaluating student understanding?" The answer to this question rests on the subjective matter of determining an operational value for "close enough." Within the confines of the present study, the fact that "coding" (along with "gender" and a substantial set of dummy values) were able to explain less than half the variability in "multiple choice" performance suggests an answer of "no." But restricted budgets, larger class sizes, increasing teaching loads, the lack of rewards for creating and manually evaluating constructed-response tests, and/or the lack of penalties on faculty for using only MC tests all potentially mitigate the sole use of MC tests for student evaluation tasks. These considerations suggest that the more-reasonable answer may be "yes."

6. REFERENCES AND BIBLIOGRAPHY

Bar-Hillel, Maya and Yigal Attali (2002) "Seek Whence: Answer Sequences and Their Consequences in Key-Balanced Multiple-Choice Tests" *The American Statistician* vol. 56, no. 4 (November), pp. 299-303.

Becker, William E., and Carol Johnston (1999) "The Relationship Between Multiple Choice and Essay Response Questions in Assessing Economics Understanding" *Economic Record* vol. 75, no. 231 (December), pp. 348-57.

Becker, William E. and William B. Walstad (1999) "Data Loss From Pretest to Posttest as a Sample Selection Problem" *Review of Economics and Statistics* (February), pp. 184-188.

Bloom, B. S (1956) *A Taxonomy of Educational Objectives* New York, David McKay Company.

Bridgeman, Brent and Charles Lewis (1994) "The Relationship of Essay and Multiple-Choice Scores with Grades in College Courses" *Journal of Educational Measurement* vol. 31, no. 1 (Spring), pp. 37-50.

Bridgeman, Brent and Rick Morgan (1996) "Success in College for Students with Discrepancies Between Performance on Multiple-Choice and Essay Tests" *Journal of Educational Psychology* vol. 88, no. 2, pp. 333-340.

Bridgeman, Brent (1992) "A Comparison of Quantitative Questions in Open-Ended and Multiple-Choice Formats" *Journal of Educational Measurement* vol. 29, pp. 253-71.

Chan, Nixon and Peter E. Kenedy (2002) "Are Multiple-Choice Exams Easier for Economics Students? A Comparison of Multiple Choice and Equivalent Constructed Response Exam Questions" *Southern Economic Journal l* vol. 68, no. 4 (April), pp. 957-971.

Douglas, Stratford and Joseph Sulock (1995) "Estimating Educational Production Functions with Correction for Drops" *Journal of Economic Education*, Spring, pp. 101-12.

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

Grimes, Paul W. and Paul S. Nelson (1998) "The Social Issues Pedagogy vs. The Traditional Principles of Economics: An Empirical Examination." *The American Economist*, Spring, pp. 56-64.

Heckman, James M (1979) "Sample Selection Bias as a Specification Error." *Econometrica*, January, pp. 153-61.

Katz, Irvin R., Randy E. Bennett, and Aliza E. Berger (2000) "Effects of Response Format on Difficulty of SAT Mathematics Items: It's Not the Strategy." *Journal of Educational Measurement* vol. 37, pp. 39-57.

Kennedy, Peter E., and William B. Walstad (1997) "Combining Multiple-Choice and Constructed-Response Test Scores: An Economist's View" *Applied Measurement in Education* 10:3, pp. 59-75.

Kniveton, Bromley H. (1996) "A Correlational Analysis of Multiple-Choice and Essay Assessment Measures" vol. 56 (November), pp. 73-84.

Kreig, Randall G. and Uyar, B. (2001) "Student Performance in Business and Economics Statistics: Does Exam Structure Matter?" *Journal of Economics and Finance* vol. 25, no. 2 (Summer), pp. 229 – 240.

Lumsden, Keith G. and Alex Scott (1987) "The Economics Student Reexamined: Male-Female Difference in Comprehen-sion." *Journal of Economic Education*, Fall, pp. 65-75.

Maxwell, Nan L. and Jane S. Lopus (1995) "A Cost Effectiveness Analysis of Large and Small Classes in the University." *Educational Evaluation and Policy Analysis*, Summer, pp. 167-78.

Messick, Samuel (1993) "Trait equivalence as construct validity of score interpretation across multiple methods of measurement" in Construction Versus Choice in Cognitive Measurement, edited by Randy E. Bennett and William C. Ward. Hillsdale, NJ: Lawrence Erlbaum, pp. 61-73.

O'Neil, Patrick B. (2001) "Essay Versus Multiple Choice Exams: An Experiment in the Principles of

Macroeconomics Course" *American Economist* vol. 45, no. 1 (Spring), pp. 62-70.

Perrig, W. and W. Kintsch (1985) "Propositional and Situational Representations of Text." *Journal of Memory and Language* 24, pp. 503-518.

Ray, Margaret A. and Paul W. Grimes (1992) "Performance and Attitudinal Effects of a Computer Laboratory in the Principles of Economics Course." *Social Science Computer Review*, Spring, pp. 42-58.

Saunders, Philip (1991) Test of Understanding in College Economics, Third Edition, *Examiner's Manual*, New York: Joint Council on Economic Education.

Scouller, Karen (1998) "The Influence of Assessment Method on Students' Learning Approaches: Multiple Choice Question Examinations versus Assignment Essay" Higher Education vol. 35, pp. 453-472.

Siegfried, John J. and Rendig Fels (1979) "Research on Teaching College Economics: A Survey" *Journal of Economic Literature* (September), pp. 923-69.

Siegfried, John J. and William B. Walstad (1998) "Research on Teaching College Economics," in *Teaching Undergraduate Economics: A Handbook for Instructors*, edited by W.B. Walstad and P.Saunders. Irwin/McGraw Hill: Boston, pp. 141-66.

Siegfried, John J., and Peter E. Kennedy (1995) "Does Pedagogy Vary with Class Size in Introductory Economics?" *American Economic Review*, Papers and Proceedings 85:3, pp. 47-51.

Snow, Richard E. (1993) "Construct Validity and Constructed-Response Tests" in *Construction Versus Choice in Cognitive Measurement*, edited by Randy E. Bennett and William C. Ward., Hillsdale, NJ, Lawrence Erlbaum, pp. 45-60.

Sykes, Robert C., and Wendy M. Yen (2000) "The Scaling of Mixed-Item Format Tests with One-Parameter and Two-Parameter Partial Credit Model" *Journal of Educational Measurement* 37, pp. 221-44.

Tate, Richard (2000) "Performance of A Proposed Method for the Linking of Mixed Format Tests with Constructed Response And Multiple Choice Items" *Journal of Educational Measurement* 37, pp. 329-46.

Traub, Ross 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 Randy E. Bennett and William C. Ward. Hillsdale, NJ: Lawrence Erlbaum, pp. 29-44.

Wainer, Howard, and David Thissen (1993) "Combining Multiple-Choice and Constructed-Response Test Scores: Toward a Marxist Theory of Test Construction" *Applied Measurement in Education* 6:10, pp. 3-18.

Walstad, William B. (1998) "Multiple Choice Tests for the Economics Course" in *Teaching Undergraduate Economics: a Handbook for Instructors*, edited by William B. Walstad and Phillip Saunders, New York: McGraw-Hill, pp. 287-304.

Walstad, William B. and William E. Becker (1994) "Achievement Differences on Multiple Choice and Essay Tests in Economics" *American Economic Review*, vol. 84, no. 2 (May), pp. 193-6.

Walstad, William B., and Denise Robson (1997) "Differential Item Functioning and Male-Female Differences on n Multiple-choice Tests in Economics" *Journal of Economic Education* 28:1, pp. 55-71.

Welsh, Arthur L., and Phillip Saunders (1998) "Essay Questions and Tests" in *Teaching Undergraduate Economics: A Handbook for Instructors* edited by William B. Walstad and Phillip Saunders, New York: McGraw-Hill, pp. 305-318.

Zeidner, Moshe (1987) "Essay versus Multiple-Choice Type Classroom Exams: The Student's Perspective" *Journal of Educational Research* vol. 80, no. 6 (July/August), pp. 352-358.

7. END NOTE

[1] Not all students prefer MC tests over constructed-response tests. In the Zeidner (1987) study, for example, about 23% of the students preferred essay tests. Student rationales for this included (1) a preference for communicating written answers in more personal formats, (2) the possibility of earning partial credit for incomplete answers, (3) greater confidence in their writing abilities than recall skills, (4) an opportunity to defend an unpopular position in a forceful way, (5) an increased probability of making errors in MC formatted tests, and (6) the possibility that subjective grading works in their favor regardless of what they say.

AUTHOR BIOGRAPHIES

Mark G. Simkin is a professor of Information Systems at the University of Nevada. He earned his BA degree in mathematics from Brandeis University and his MBA and Ph.D. degrees from the University of California, Berkeley. He is the author or coauthor of 15 textbooks, including 3 on Visual Basic and 8 in Accounting Information Systems. He is also the author or coauthor of over 100 research articles, some of which have been published in *Decision Sciences, JASA, the Communications of the ACM, the International Journal of Information Management, the Journal of Accountancy, the Journal of Computer Information Systems, Interfaces, and the Journal of Systems Management.*

William L. Kuechler is an associate professor of Information Systems at the University of Nevada. He holds a BS in Electrical Engineering from Drexel University and a Ph.D. in Computer Information Systems from Georgia State University. His current research areas benefit from his twenty-year career in business software systems development and includes studies of inter-organizational workflow and coordination, web-based supply chain integration and the organizational effects of inter-organizational systems. He has published *in IEEE Transactions on Knowledge and Data Engineering, Decision Support Systems, Journal of Electronic Commerce Research, IEEE Transactions on Professional Communications, Information Systems Management, Annals of Information Technology Cases,* the proceedings of *WITS, HICSS* and other international conferences and journals. Dr. Kuechler is a member of IEEE and the ACM.

Information Systems & Computing
Academic Professionals

STATEMENT OF PEER REVIEW INTEGRITY

All papers published in the Journal of Information Systems Education have undergone rigorous peer review. This includes an initial editor screening and double-blind refereeing by three or more expert referees.

ISSN 1055-3096