Decision Sciences Journal of Innovative Education
Volume 3 Number 1
Spring 2005
Printed in the U.S.A.

Multiple-Choice Tests and Student Understanding: What Is the Connection?

Mark G. Simkin† and William L. Kuechler
College of Business Administration, University of Nevada, Reno, Nevada 89557,
e-mail: simkin@equinox.unr.edu

ABSTRACT

Instructors can use both "multiple-choice" (MC) and "constructed response" (CR) questions (such as short answer, essay, or problem-solving questions) to evaluate student understanding of course materials and principles. This article begins by discussing the advantages and concerns of using these alternate test formats and reviews the studies conducted to test the hypothesis (or perhaps better described as the hope) that MC tests, by themselves, perform an adequate job of evaluating student understanding of course materials. Despite research from educational psychology demonstrating the potential for MC tests to measure the same levels of student mastery as CR tests, recent studies in specific educational domains find imperfect relationships between these two performance measures. We suggest that a significant confound in prior experiments has been the treatment of MC questions as homogeneous entities when in fact MC questions may test widely varying levels of student understanding. The primary contribution of the article is a modified research model for CR/MC research based on knowledge-level analyses of MC test banks and CR question sets from basic computer language programming. The analyses are based on an operationalization of Bloom's Taxonomy of Learning Goals for the domain, which is used to develop a skills-focused taxonomy of MC questions. However, we propose that their analyses readily generalize to similar teaching domains of interest to decision sciences educators such as modeling and simulation programming.

Subject Areas: Constructed-Response Tests, Student Assessment, and Multiple-Choice Tests.

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, problem-solving exercises, and essay questions. Most of the alternates to MC and true-false questions are described in the literature as "constructed-response" (CR) 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 test formats, there are many reasons why most students and many instructors prefer MC tests over the other types of examination questions. However, the literature contains a considerable

†Corresponding author.

body of analysis and debate over this issue. Even after years of research on the trait-equivalence of MC versus CR examinations in different domains "... the evidence is inconclusive" (Martinez, 1999).

The effectiveness of MC questions alone in measuring student understanding of class materials is of particular interest to those decision sciences instructors teaching quantitative courses (such as Production courses) or procedural language programming classes (such as Visual Basic or simulation languages). For example, many decision science instructors have heavy teaching responsibilities, or must balance their preferences for penetrating but labor-intensive computational tests against their research and publication responsibilities. MC tests provide an easy alternative to such constructed-response tests, but nagging questions remain regarding the ability of such examinations to accurately and fairly measure student understanding of course concepts.

The next section of this article examines this debate in greater detail and highlights some of the issues involved in test formats in general and MC testing in particular. It discusses both the research in educational psychology that proposes MC tests can measure many learning outcomes as accurately as CR tests and the tenuous relationships that to date have been found and not found between MC and CR questions in specific educational domains.

The third section of the article begins with an overview of the many knowledge classification frameworks that have been developed. We then operationalize Bloom's Taxonomy of Learning Goals (Bloom et al., 1956) for the skills covered in beginning programming language instruction. Using this taxonomy, we perform a knowledge-level analysis of both MC test banks and multiple constructed response examination questions for basic computer language courses. We then develop a classification scheme for MC questions for our domain and compare the question classes to the knowledge levels exhibited in programming (CR) questions.

Our analysis suggests that carefully crafted MC examinations can overcome some of the problems that have been documented for this type of exam. The last section of this article summarizes this finding in the form of a research direction for expanding understanding of the relationship between MC and CR questions in basic computer language programming and similar domains and presents our conclusions.

REVIEW OF THE LITERATURE

Advantages and Concerns of Multiple-Choice Tests

Many instructors appear to use MC tests as a preferred assessment tool. In teaching economics, for example, investigators estimate MC test usage for student assessments of between 45% and 67% (Becker & Watts, 2001; Siegfried & Kennedy, 1995). Our sense is that the actual number is probably higher today because MC tests are so ubiquitous across the other disciplines (even in teaching compositional writing—see Bridgeman, 1991) and also because they are now used so frequently on entrance examinations such as SAT tests, online (Web-based) tests, mass-lecture courses, and certification tests such as the CPA exam.

Table 1: Advantages of multiple-choice tests over constructed response tests.

Advantage	Citations
Machine gradable, thereby increasing scoring accuracy	Holder and Mills (2001), Walstad (1999), Kniveton (1996), Walstad and Becker (1994)
Efficient way to collect and grade examinations from large numbers of test takers	Dufresne et al. (2002)
Helps certification examiners agree on questions to ask a large number of test takers	Snyder (2004), Holder and Mills (2001), Bridgeman and Rock (1993), Bridgeman (1991)
Facilitates referencing the correct answer in textbook or other source	Bridgeman and Lewis (1994)
Perceived objectivity in grading process	Zeidner (1987), Wainer and Thissen (1993), Becker and Johnson (1999)
Facilitates timely feedback for test takers in classes, and immediate feedback in web-based systems	Delgado and Prieto (2003), Epstein, Epstein, & Brosvic (2001), Epstein and Brosvic (2002), Kreig and Uyar (2001)
Enables instructors to ask a large number of questions on a wider range of subject materials	Becker and Johnson (1999), Walstad and Robson (1997), Lukhele et al. (1994), Saunders and Walstad (1994)
Helps students avoid losing points for poor spelling, grammar, or poor writing ability	Zeidner (1987)
Easier preparation by test takers	Carey (1997), Ramdsden (1988), Scouller (1998), Frederickson and Collins (1989).
Does not require deep understanding of tested material (student advantage)	Biggs (1973), Beard and Senior (1980), Entwistle and Entwistle (1992)
Reduces student anxiety	Snow (1993)
Multiple versions of the same MC examination helps thwart cheating	Kreig and Uyar (2001), Wesolowsky (2000)
Helps avoid inconsistent grading of essays	Kniveton (1996)
Availability of computerized MC test banks, answer keys, and test generators	
Test takers can increase the probability of guessing the right answer to a question by eliminating unlikely choices (student advantage)	Bush (2001), Hobson and Ghoshal (1996).
Electronic test items can easily be edited, pre-tested, stored, and reused	Haladyna and Downing (1989).

The literature on testing pedagogy suggests that both students and faculty often prefer MC tests, although they appear to do so for different reasons (Chan & Kennedy, 2002; Zeidner, 1987). Table 1 summarizes some of the advantages of using MC tests. From the standpoint of the instructor, these advantages include (1) the ease and accuracy with which such tests can be machine graded (or regraded) and returned to students on a timely basis, especially in large classes; (2) the ability to create multiple versions of the same examination, thereby better enabling

instructors to control cheating; (3) the ability to evaluate a test itself on a question-by-question basis; (4) the ease with which correct answers can be verified; and (5) the ability to cover a wide range of material.

From the standpoint of the student, these advantages include: (1) the perception that MC tests are objective and, therefore, avoid instructor bias; (2) the ability to guess correct answers (often without penalty); (3) the ability to earn partial credit if he or she is a slow test taker; and (4) the perceived ability to do better on MC tests than on essay or other forms of constructed response tests. Student preferences for a specific test format appear to have functional validity as well. It has been shown that a "small but statistically significant" subset of students may receive grades that are not indicative of their true mastery of material based on construction of the exam. This applies whether exams are composed purely of MC questions, essay questions, or some composite of the two (Kennedy & Walstad, 1997).

However, not all students or university faculty members prefer MC tests. In the Zeidner (1987) study, for example, about 25% of the student respondents indicated a preference for constructed-response test formats over MC formats. Scholars also note that essay-test formats enable students to "obfuscate under uncertainty," sometimes with advantageous results (Becker & Johnson, 1999).

Many college-level instructors express concern over the difficulty of constructing MC examination questions (when a test bank is not available) compared to constructing CR examination questions (Brown et al., 1997). This has raised the possibility that poorly written MC questions actually hide student knowledge rather than reveal it (Dufresne, Leonard, & Gerace, 2002; Becker & Johnson, 1999). A related matter concerns the wording of MC questions and particularly the possibility that students with poor verbal skills are disadvantaged by poorly or ambiguously worded MC tests (Paxton, 2000). In our experience, this latter consideration is particularly important to students for whom English is a second language.

Another criticism of MC tests is that instructors, in an effort to remove ambiguous questions, may construct examinations that are too easy and provide an inaccurate indicator of student understanding. This misleads faculty concerning students' grasp of course concepts or mastered course materials (Chan & Kennedy, 2002). A number of researchers have investigated the belief, widely held by educators in many fields (Martinez, 1999; Fenna, 2004), that the results of MC tests can be influenced by "testwiseness," the use by students of format-specific strategies during examinations. The most common technique is to eliminate one or more MC answers based on only a partial understanding of the knowledge being tested and thus generate misleadingly high test scores. Studies by Rogers and Hartley (1999) and Zimmerman and Williams (2003) both corroborate the influence of testwiseness on MC examinations.

Scholars also note that MC questions deny a student the opportunity to organize, synthesize, or argue coherently, to express knowledge in personal terms, or to demonstrate creativity—skills that many scholars feel are better revealed with essay exam formats (Tuckman, 1993; Lukhele et al., 1994; Bridgeman, 1992). Some observers express the concern that MC tests discourage critical thinking, fail to attract students to exciting areas in science and industry, and encourage students

to view course assessments as a "numbers game" (Carey, 1997). There is also evidence that students study course material differently depending on the type of test they anticipate and this alters the nature and quality of student learning. Studies are mixed in their detection of anticipation effects; however a majority of studies have found that response formats make a difference in anticipatory learning and that the expectation of CR tests favors concept learning while the anticipation of MC tests favors detail memorization (Martinez, 1999; Traub & MacRury, 1990).

Other research argues that MC tests suffer from gender, answer-choice, racial, or other types of bias that render inaccurate assessments of student understanding. With regard to gender differentials, for example, past research suggests that males may have a relative advantage on MC tests (Bell & Hay, 1987; Lumsden & Scott, 1987; Bolger & Kellaghan, 1990). 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 economics tests using fixed-response tests rather than constructed-response tests (Walstad & Becker, 1994; Greene, 1997; Chan & Kennedy, 2002). It is also not clear whether this potential gender advantage (or lack of it) applies to students taking computer classes.

Perhaps the most important drawback to CR examinations is the comparatively large amount of time required to grade them. For those university instructors who lack student graders, or who feel compelled to perform their own grading, CR tests impose a substantial time commitment, expose the grader to complaints of inconsistent evaluations or bias (Ashburn, 1938), and (in large classes) are all but impossible to return to students in a timely fashion. At institutions that primarily reward research productivity, CR tests also create an opportunity cost in the time and effort drawn away from more directly rewarded publication activities.

What Do Multiple-Choice Tests Really Measure?

Many educators who teach elementary programming language courses share the belief with the educational community at large that CR questions better measure a student's ability to solve real-world problems (Hancock, 1994; Rogers & Hartley, 1999; Bacon, 2003; Fenna, 2004). In programming language instruction and similar domains, this belief is reinforced by the lack of structural fidelity of MC examinations; structural fidelity refers to the congruence between performance called upon by the test and proficient performance in the referent domain (Fredericksen, 1984; Messick, 1993). The behavior of professional programmers much more closely approximates answering CR questions than answering MC questions. Still, universal preference for MC examinations would be assured if it could be demonstrated that such tests are as accurate as constructed-response tests in measuring student understanding of course subject matter. Given such demonstration, there would be little need to resort to more arduous examination formats and grading procedures, and faculty could be assured that their (MC) tests were both fair and incisive measures of student mastery.

We pose two fundamental questions in pursuit of a learning measure that is both tractable and complete:

Research Question 1: Can MC and CR tests measure the same levels of understanding in the area of basic computer language programming?

Research Question 2: To what degree do MC questions on this topic predict performance on CR questions on the same topic?

A fairly extensive body of research has addressed the issue of how well MC questions test student 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). This research, explored more fully in the section on knowledge levels, has demonstrated that it is theoretically possible to construct MC items that measure many of the same cognitive abilities as CR items. However, these general studies usually stress that empirical testing is necessary to validate the theoretical propositions in each learning domain. Within the stream of empirical, domain-specific MC-versus-CR research, one of the most difficult problems has been to determine a suitable measure for "student understanding." One stream of effort has used constructed-response questions for this test with underlying null hypothesis: "whatever is measured by the MC questions is equally (or at least adequately) measured by constructed response questions." Consequently, student performance on MC questions has been the independent variable, and student performance on constructed response questions has been the dependent variable. This pragmatic measurement position is termed *trait equivalence* and claims only that the same score on trait-equivalent MC and CR tests indicates equal levels of understanding of the subject matter. In contrast, the stronger claim of *psychological equivalence* indicates that the equivalent cognitive processes are being exercised in both types of tests and is rarely sought in domain-specific empirical research.

Early empirical tests of this hypothesis led some scholars to conclude that MC tests and constructed response tests do in fact measure the same thing (Taub, 1993; Wainer & Thissen, 1993; Bennett, Rock, & Want, 1991; Bridgeman, 1991; Sanders & Walstad, 1990). In a study by Lukhele, Thissen, and Wainter (1994), for example, the authors concluded that "the constructed response portion of the tests yielded little information over and above that provided by the multiple choice sections" on their achievement tests. Similarly, after examining sample tests from seven different disciplines, Wainer and Thissen (1993) thought this relationship was so strong that they concluded "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 Bridgeman and Rock (1993), Walstad and Becker (1994), and Kennedy and Walstad (1997) reached similar conclusions. Finally, a study by Bacon (2003) found MC testing to yield "equivalent reliability and validity" compared to short-answer tests.

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 student performance on MC and essay questions on economics examinations. They, therefore, concluded, "these testing forms measure

different dimensions of knowledge" (Becker & Johnston, 1999, p. 348). Similarly, a study of the use of MC questions in physics instruction by Dufresne et al. (2002) led the authors to conclude that student answers on MC questions "more often than not, [give] a false indicator of deep conceptual understanding."

Possible Confounds in Prior Research

There are many potential explanations for the lack of consistency in the empirical studies that have attempted to determine whether MC and CR questions measure learning equivalently. Small samples, non-random samples, variations in student levels, ages, or life experiences, and perhaps unwilling participants have undoubtedly contributed to the problem, as stated by the researchers themselves in the discussion and limitations sections of their papers. The lack of an adequate surrogate for "knowledge mastery" has also probably played a role. This has led some researchers to conclude that MC questions and constructed response questions are really not interchangeable, and that they probably examine different levels of cognition (Bridgeman & Rock, 1993; Walstad & Becker, 1994; Kennedy & Walstad, 1997; Kuechler & Simkin, 2004).

Curiously, however, many empirical studies of MC tests and what they measure make the implicit assumption that MC questions are homogenous entities that possess uniform power in measuring student understanding of subject material. Though we do not claim to have found all such studies, the articles by Bridgeman & Rock (1993), Walstad & Becker (1994), Kennedy & Walstad (1997), Dufresene et al. (2002), Kuechler & Simkin (2004), and Bacon (2003), all examine the MC versus CR question in specific educational domains and do not differentiate MC questions on any criteria—that is, *they treat them as homogenous entities*. In our opinion, corroborated by the theoretical work in educational psychology, this doesn't make sense. Just as a carpenter uses drill bits of various sizes to pierce wood to various depths, different MC questions test different levels of student understanding. This means that some MC questions—for example, questions regarding definitions—test only superficial knowledge or rote learning, while others—for example, questions implicitly requiring students to compute answers using unstated formulas—test deeper levels of understanding or mastery (Anderson & Krathwohl, 2001).

TOWARD A DEEPER UNDERSTANDING OF STUDENT ASSESSMENT

In this section of the article we present several analyses that show that carefully constructed MC questions on the topic of basic computer language programming can test a broad range of levels of understanding of that subject. In the following section, based on these analyses, we propose an empirical research direction to explore the relationship between performance on MC questions and constructed response questions *at different cognitive levels*. The basis for the analysis is Bloom's taxonomy of learning objectives, which we operationalize for our specific area of interest: introductory programming language instruction. We use this taxonomy to perform a detailed knowledge-level analysis of both constructed response and MC questions that have been used in our basic programming courses and commonly

used test banks on the subject. We then classify the MC questions according to the taxonomy and abstract rules for question construction at different knowledge levels. Finally, we compare the knowledge levels of MC and CR questions and propose a research direction.

Levels of Knowledge

Intuitively, teachers and students understand that examination questions, including MC questions, can span a wide range of difficulty. An early and extremely influential effort to explicate the vague concept of "difficulty" resulted from the working committee of educators under the leadership of Benjamin Bloom, who examined this question in 1956 (Bloom et al., 1956). The resulting rank ordering of "cognitive educational objectives" has become known as "Bloom's Taxonomy" (of educational goals). Bloom's original taxonomy has had a significant and ongoing impact on teaching theory and curriculum development since its original publication. Table 2 summarizes the levels of understanding set out in Bloom's report.

The full taxonomy as developed in Bloom (1956) specifies sublevels for all six major levels. Table 2 shows sublevels only for the primary-level named "Comprehension." The levels from "Knowledge" through "Evaluation" are ordered along a dimension that has typically been termed "understanding."

Bloom's taxonomy was developed prior to the very substantial body of research into learning and cognition that has occurred over the last 45 years. However, the insights and intuitions of the experienced instructors on the Bloom committee have proven extremely durable and accurate. Recently a commission of psychologists and educators re-evaluated the original taxonomy in light of "the full scope of research into thinking, cognition and problem solving" that had been developed since 1956 (Anderson & Krathwohl, 2001). This commission produced a revised version of Bloom's taxonomy that validated the original by mapping six well-researched cognitive processes to a set of knowledge levels derived directly from the original taxonomy. The revised taxonomy is shown in Table 3.

In addition to the revised Bloom taxonomy of Table 3, Anderson and Krathwohl (2001) also describe 11 other knowledge-level taxonomies proposed by different researchers for different purposes over the past 30 years. Anderson and Krathwohl also suggest that a simplified "cognitive process" taxonomy, suited to the empirical research we propose, is inherent in the six column headings in Table 3. Most of the levels in these taxonomies map to closely corresponding levels in other taxonomies; many would have served well in providing a partial order for levels of knowledge in this research. However, we chose Bloom's taxonomy for several reasons. First, it is the most widely known and, therefore, the most accessible to our audiences—decision science educators. Second, it is used in more prior (Hancock, 1994; Lipscomb, 1985) and current studies (e.g., Kastberg, 2003) than any other taxonomy. This permits our work to be more easily compared to prior work. Third, Bloom's taxonomy is regarded as a stricter hierarchy than many other taxonomies (Krathwohl, 2002) with less overlap between levels. Finally, a hierarchical taxonomy has significant benefits when proposing a domain-specific operationalization for creating examinations because each question requiring

Table 2: Bloom's original cognitive taxonomy.

Level	Description	Evidence of Ability
1. Knowledge	Rote memory; recognition without (necessarily having) the ability to apply learned knowledge	Answer strongly cued T/F or multiple-choice questions
2. Comprehension	Information has been assimilated into students frame of reference	Student can *understand* problems similar to those given in class
Translation	Gives meaning to information	Can put into own words
Interpretation	Changing from one form of representation to another	Can classify material according to experience
Extrapolation	Use information in new context	Ability to predict consequences
3. Application	Abstracts from learned material to solve new (analogous) situations	*Uses* learned techniques and knowledge in the production of solutions to novel (but structurally similar) situations
4. Analysis	Decompose learned material into components and understand the relationships between them	Recognize unstated assumptions; identify motives; separate conclusions from supporting evidence
5. Synthesis	Combine the elements of learned knowledge (abstracted in the application level and explicated into separate units in the analysis level) into new integrated wholes	Knowledge creation; fill gaps in existing knowledge or procedures to solve unstructured problems
6. Evaluation	Makes judgments about the value or worth of learned information	Produces judgments of worth concerning directions of knowledge acquisition

specific evidence of achievement is more precisely traced to a specific level of understanding.

Operationalizing Bloom's Taxonomy

It is important to understand that the general definitions of the educational goal levels of the Bloom taxonomy from the original report must be interpreted for different domains. Bloom himself felt that "Ideally each major field should have its own taxonomy of objectives in its own language—more detailed, closer to the special language and thinking of its experts, reflecting its own appropriate sub-divisions and levels of education, with possible new categories, combinations of categories and omitting categories as appropriate" (Anderson & Krathwohl, 2001, p. xxviii). We have found the goal hierarchy itself (the six levels of Table 2) to be entirely adequate. However, we have had to interpret *Evidence of Ability* attributes for each

Table 3: Bloom's taxonomy revised.

Type of Knowledge	Type of Cognitive Process					
	Remember	Understand	Apply	Analyze	Evaluate	Create
Factual	Recognize facts					
Conceptual		Interpret and infer concepts				
Procedural	Recognize procedures			Executing and implementing procedures		
Metacognitive						Planning strategies

level in the specific domain of introductory programming language teaching from descriptions of the use of these levels in nontechnical areas such as history, economics, and English (Mayer, 2002; Larsen, 1997; Reeves, 1990; Barnett & Ceci, 2002).

Different interpretations of both knowledge-level definitions and Evidence of Ability attributes in different domains are both necessary and problematic. They are problematic because they complicate meta-level analyses of MC versus CR studies. Even within the same domain, different evidence of ability interpretations may lead to different conclusions by different researchers. Interpretations are necessary because different educational domains vary widely in the cognitive skills emphasized for mastery of knowledge in the domain. Specific to this study, computer language problem solving is an area that requires both verbal skills to parse and understand the textually stated problem or requirements document and logic and mathematical skills to implement the solution and prove its correctness. Prior studies in the individual domains of English language comprehension or mathematics do not adequately address the cognitive richness of programming. We discuss our domain-specific interpretations, summarized in the third column of Table 2, in more detail below.

Given our evidence of ability interpretations, we suggest that only the first three levels of the taxonomy apply to undergraduate, introductory programming classes. The first level, *Knowledge* is demonstrated by simple recall of facts and is certainly a component of learning a computer language. The second degree of understanding, *Comprehension*, is demonstrated by low-level methodological mastery—that is, the ability to follow a set of problem-solving steps on test material that is similar to what students have seen in class or in textbooks. This level is further divided into *Translation*, the ability to restate the core concepts from the class, *Interpretation*, the ability to determine when concepts are applicable, and *Extrapolation*, the ability to predict consequences from the application of concepts, even if the consequences have not been fully explored in class or in the textbook. This level too is applicable to computer language learning at any level.

The *Application* level is demonstrated by skills we feel are near the upper bound of the understanding required for undergraduate programming classes—that

is, the ability to transfer the knowledge to new, but structurally similar, domains. We have interpreted understanding at this level to demonstrate what psychologists term *analogical reasoning*, a high-level cognitive process, whereby the *structure* of a problem rather than its specific details suggests the application of a solution. A student relating what he or she knows about planetary motion to the Bohr model of the atom is an often-cited example of analogical reasoning. To test Application-level skills, the question should require the student to transpose or apply the knowledge taught in the classroom to a different environment or problem situation. The situation should be "strictly analogous" (i.e., require a one-to-one mapping between problem and context elements), but no development of new knowledge should be required.

The fourth level, *Analysis*, is demonstrated by the ability to, for example, recognize and discuss why one type of sort was used rather than another or why one type of data structure was chosen over another in a given application. We feel this ability is beyond the skill set required for most undergraduate first courses in computer languages in an IS curriculum, based on our own experience, a review of online syllabi for introductory IS programming courses and a review of the ACM model information systems curriculum (ACM, 2002). The two levels above *Analysis*—*Synthesis* and *Evaluation*—require a student to create new knowledge in the domain of study and to evaluate the utility of learned knowledge, respectively, and are clearly at the Masters' or PhD degree level of understanding. Martinez (1999) and Hancock (1994) both concur that MC items can reliably measure the same knowledge levels as CR items only for the first four of Bloom's levels.

A Knowledge-Level Analysis of CR Questions for Introductory Programming Classes

In this section of the article we present our analysis of constructed response questions in introductory programming classes using the first three levels of the original Bloom taxonomy. As discussed earlier, many educators in IS and Computer Science feel that constructed response questions better measure a student's ability to solve real-world problems than do MC or True/False questions. The rationale for this feeling rests in the belief that reasonably complex constructed response questions (for programming classes, at least) more closely mimic the types of problems routinely solved in actual programming environments (i.e., structural fidelity).

Consider the CR question shown in Figure 1, taken from one of the authors' Internet programming classes. The question encompasses literally all the skills required to parse a common type of web-solution-amenable problem and implement the solution. "Real-world" problems are usually more involved. They also differ from test questions in length—e.g., the number of items on the screen and number of server-side computations, but not in concept. That is, repeated application of the understanding demonstrated in the test question is adequate to solve a more detailed problem, completely typical of a class of real-world web-based problem solutions deployed in industry. Using terminology traditional to computer programming instruction, the most meaningful skills involved in the solution of this problem are:

1. *Code reading:* the ability to mentally construct the screen generated by the code.

Figure 1: A constructed response (coding) question.

Question 1: Problem solving using HTML and ASP

Assume a form as shown below:

```
<HTML>
<HEAD>
<TITLE>Test Prep Form</TITLE>
</HEAD>
<BODY>
<H1>CIS 460 <BR>
Exam 2 Preparation Question Form</H1>
<H3>Please Enter Your Widget Order and click SUBMIT</H3>
<FORM ACTION="testProcess.asp" METHOD=POST>
<P>Please enter your Name</P>
<INPUT TYPE="TEXT" NAME="name"></P>
<P>Please enter the number of widgets you require</P>
<INPUT TYPE="TEXT" NAME="numWidgets"></P>
<P><INPUT TYPE="RESET" VALUE="Clear Form">
<INPUT TYPE="SUBMIT" VALUE="SUBMIT"></P>
</FORM>
</BODY>
</HTML>
```

Assume a user responded to the above form by entering her name as "Wanda" and an order for 100 widgets. Write an ASP program that will produce the following output from that input. Be sure to name your program appropriately.

Widget Order Confirmation

Hi Wanda!

Hope you have $5000 handy 'cause that's what 100 widgets cost!

2. *System behavior comprehension:* understanding that when the "SUBMIT" button of the form is clicked, the information in the fields will be encoded and sent to the server program "testProcess.asp"—the program the question requires students to write, and other details of client-server behavior.
3. *Problem comprehension:* parse the English language paragraph(s) and understand that the ASP program must multiply the cost per item by the number of items and display the computed quantity and the number of items in a stylized format. The item cost must also be calculated.
4. *Code writing:* generate a small ASP (a MS BASIC dialect) program to solve the problem.

The knowledge levels exhibited in the correct solution of the problem are orthogonal to the skills listed above. Code reading and writing both call for Knowledge-level recall of facts. Making sense of the HTML code also calls for Comprehension skills through the *Extrapolation* sublevel, that is, the ability to predict the consequences of executing the code. Both "problem comprehension" and "writing code" call for an Application-level understanding, assuming this specific problem was not discussed in class.

We note that the Application level has a broad span that is not immediately obvious from its definition. The degree to which Application-level skills are

demonstrated depends on the degree of similarity of the test problem to those solved in class. Questions that simply "change the numbers" from problems shown in class call for very limited Application-level skills. Questions for which the problem situation is less obviously similar to class or homework problems call for greater Application-level understanding. For example, given that string handling and database access had both been separately covered, but never discussed together in class, substantial Application-level understanding would be demonstrated by a successful answer to a question requiring a student to:

1. parse a string entered in a form *as a necessary precursor to*
2. determine a key for database retrieval.

Note also that problem wording is very important in determining the level of understanding actually exhibited in the solution of the problem. In the problem shown in Figure 1, for example, the sentence "Be sure to name your program appropriately." constitutes a prompt—in effect, a hint—to seek (and value) information, which the student might not otherwise have found significant.

A Knowledge-Level Analysis of MC and T/F Questions for Introductory Programming Classes

As shown in the analysis in the prior section, a typical constructed response programming question requires understanding at the *Knowledge, Comprehension,* and *Application* levels of Bloom's taxonomy. To determine the knowledge levels required for MC and T/F questions, we analyzed several textbook test banks and our own tests from classes in Internet Programming and Visual Basic. Just as for the constructed response questions, the skill types required to solve the MC questions were determined and matched to descriptions in the Evidence of Ability column of Table 2 to determine the cognitive goal level required for a successful answer to the question. We then abstracted additional attributes from multiple examples of questions at the same knowledge level to aid in identifying and constructing questions at that level. Rules for constructing MC questions at different knowledge levels and exemplar questions are shown in Figures 2 and 3. Figure 2 classifies *knowledge-level* MC questions and gives examples of each type, and Figure 3 does the same for *comprehension-level* MC questions.

Pure" knowledge questions should be clear and require little interpretation because they test student recall at the "stimulus-response" level. Note that there is less prompting (less stimulus; fewer cues) for the T/F questions (Type K3, Figure 2) in this category than for most MC questions. Prompting has been widely demonstrated to increase recall, sometimes dramatically (Tombaugh & Hubley, 2001; Rubin, 1985; Soloso & Biersdorff, 1975).

Comprehension questions require Knowledge-level recognition of facts *and* the use of that knowledge in cognitive tasks of varying type and complexity. Question C1 in Figure 3 tests the depth of understanding of a non-trivial concept by asking for the best interpretation of the term from multiple, superficially similar answers. Question C2 tests the ability to recognize a number of different programming constructs *in-situ*. Question C3 tests low-level procedural understanding—the ability to trace the flow of a simple program and determine the consequences of

Figure 2: Knowledge-level MC question descriptions and exemplars application level.

(Type K1) Completion questions where the answer is specified:
Describing within a program why and how you coded a section of a program the way you did is called: (A) documentation, (B) welshing, (C) coding, (D) Boolean logic
(Type K2) Direct questions where the full answers are chosen from a list:
In the statement **FOR J = X TO Y STEP Z**, what is the default value for Z when "**STEP Z**" is omitted? (A) the same as X, (B) the same as Y, (C) 0, (D) 1
(Type K3) "Direct" unambiguous T/F questions:
To concatenate strings in Visual Basic use the symbol "$". **T**rue or **F**alse

its operation. However, no Comprehension-level question (as we have interpreted Bloom's taxonomy) requires *construction* of a solution to a problem.

The authors were surprised to find that it is extremely difficult to construct a MC question at the Application level because the nature of MC question constructs precludes a student's ability to create a solution. However, construction of a programming solution to a question posed in a different knowledge representation, typically text, is the functional definition by which we determine a question gauges understanding at the Application level! The limitation is due to the fact that with MC items cognition must eventually lead to convergence on a single response selected from a small option set; this precludes what the educational assessment literature terms *divergent production* (Boodoo, 1993), an integral part of sophisticated computer language programming. Even the most sophisticated MC questions that use code sections, such as Comprehension example C3a in Figure 3, can be answered by *tracing the flow* of the preconstructed programs, a cognitive process quite distinct from solution construction.

The closest MC approximation to an Application-level question in the programming domain that we could determine was the "find and fix the error" type of question similar to the example shown in Figure 4. One method of solution to the problem of Figure 4 is to mentally program a solution to the English language description of the function of the code, however, even in this quite sophisticated example the most straightforward solution strategy is to implement each suggested program change (answers A–D) in turn and trace the program flow to final output. Tracing program flow, however, is a Comprehension-level skill utilizing different cognitive processes than ad hoc construction of a solution. To meet the evidence of ability for the Application level a MC item would require (1) a statement of a problem, which required the student to transpose knowledge stated in class to a different domain, and which (2) had a single correct answer (to be chosen from the MC list) that could only have been arrived at by the desired transposition of knowledge. A very challenging item construction task!

The knowledge-level analyses above seem to indicate that no combination of MC questions can replicate exactly the skill set required to solve constructed

Figure 3: Comprehension-level MC question descriptions and exemplars.

(Type C1) Multiple choice requiring the student to "pick the best *paraphrase*":

An overloaded operator in an object oriented language: (A) will fail if used beyond its limits (B) does more than one thing, (C) displays a selection of choices to the operator (D) performs analogous operations on different data types

(Type C2) Classification questions:

The condition in the statement
IF ((SQRT(Z) > 12 && (COST*QUAN >5000))THEN RETURN
(A) uses inequalities, (B) is an example of Boolean logic, (C) demonstrates VB's function library (D) all of the above

(Type C3) Simple consequence questions:

How many lines of output are produced by the following program segment?

```
For i = 1 to 3
    For j  = 1 to 3
        For k  = i to j
            lstBox.Items.Add("Programming is fun!")
        Next
    Next
Next
```

(A) 8 (B) 9 (C) 10 (D) 11 (E) none of these

(Type C3a) More complex consequence questions which call for understanding the programmed implementation of more advanced concepts:

Which of these program fragments computes the sum of the series 1/2 + 2/3 + 3/4 + . . . + 99/100?

```
(A) for n = 1 to 99
        s += n / (i+n)
    next
(B) for q = 100 to 1
        s += (q + 1) / q
    next
(C) for d = 2 to 99
        s += 1 / d + d /  (d+1)
    next
(D) for x = 1 to 100
        s += 1 / (x + 1)
    next
```

response questions. However, what we seek is not to duplicate constructed response tests, but rather a means to predict performance in such tests using more tractable MC questions. Our analysis does show that, as expected, there are very substantial differences between MC questions, as classified with Bloom's taxonomy. This could well account for the lack of correlation found between performance on MC questions and constructed response questions (Kuechler & Simkin, 2003). In our earlier study, as in most MC versus CR research to date, we treated MC items as a homogenous entity; however, research in educational psychology predicts greater correlation between MC and CR questions at the higher cognitive levels (Martinez, 1999).

Figure 4: A sophisticated MC question approximating application-level understanding.

When the odd numbers are added successively, any finite sum will be a perfect square. For example, 1 + 3 + 5 = 9 (= 3 ^ 2). What change must be made in the following program to correctly demonstrate this fact for the first few odd numbers:

```
Private Sub btnDisplay_Click(. . .) Handles btnDisplay.Click

        Dim oddNumber, i, j, sum as Integer
        For i = 1 To 9 Step 2                    ' Generate the first few odd numbers
                oddNumber = i
                For j = 1 To oddNumber Step 2    ' Add odd numbers
                        sum += j
                Next
                lstBox.items.Add(sum & "is a perfect square.")
        Next
End Sub
```

(A) Change the Step size to 1 in the first **For** statement
(B) Move **oddNumber = i** inside the second **For** loop
(C) Reset **sum** to zero immediately before the second Next statement
(D) Reset **sum** to zero immediately before the first Next statement

A RESEARCH DIRECTION: HYPOTHESES AND IMPLICATIONS

Bloom's educational taxonomy and an analysis of our own MC and CR test questions have alerted us to the potential usefulness of classifying examination questions by knowledge level for the purpose of greater accuracy in student assessment. They also suggest that a strong relationship may exist between student performance on MC and CR test questions "if the analysis is restricted to comparisons of test results for questions at the same cognitive level." In other words, we hypothesize that, if examination questions at the same level of understanding are asked in different formats, there should be a strong correlation in the assessment scores. This means that researchers should find a greater overlap in cognitive skills between lower level MC questions and lower level CR questions, and a similarly strong relationship between higher (Knowledge level) MC questions and CR questions. Thus, we hypothesize:

> H_1: Given multiple examinations testing the same basic programming skills, one utilizing only constructed response questions and one utilizing only MC questions, the correlation between the scores on the CR examination and the MC exams will be high if all questions are at the same knowledge level.

The rationale for this hypothesis is straightforward: where both MC and CR questions test student understanding of class materials "at the same underlying cognitive skill level," the observer should find the same level of achievement. This does *not* mean that all students will perform equally well on both types of item even when they are at the same level. Rather, it means that those students who do well on

the MC questions will do well on the CR questions, and vice versa—hopefully for the obvious reason that they understand the material. Conversely, we would expect those students who do *not* do well on, say, the MC portion of an examination will also do poorly on the CR portion of a test covering the same material—again for the same reason.

What is less straightforward is a corollary to our hypothesis: MC examinations that indiscriminately mix questions at varying knowledge levels will exhibit less correlation with the scores of CR examinations on the same material. We believe it is this element that has, in fact, been an important confounding factor in those empirical studies which have not controlled for it and subsequently found weak relationships between the student performances on separate-but-otherwise equal MC and CR tests. Finally, we believe it is also possible that MC examinations containing only questions at the highest cognitive levels contained in the mixed-level examination will exhibit the strongest correlations with CR questions. Again, however, this is hypothetical and subject to empirical investigation, but is supported by general investigations in educational psychology (Martinez, 1999).

In the discussions above, we have restricted our hypotheses to our chosen discipline—computer programming. However, given the abstract nature of our thinking, there is good reason to suspect that these relationships apply to MC/CR tests in similar disciplines as well. By "similar disciplines" we mean those that, like computer language programming, call for both strong language skills to parse the problem statement, and strong logic skills to formulate a solution from a possibly infinite set of workable solutions (divergent production). Thus, a second hypothesis is:

> H_2: The relationships between MC and CR suggested above for introductory programming classes will generalize to student tests in subjects requiring the same mix of strong language and logic skills, i.e., to the tests in such disciplines as accounting, finance, production, and simulation.

These are testable hypotheses that should be investigated empirically over a wide range of examinations, subjects, and disciplines. Accordingly, we invite others to join us in conducting field experiments to test them over the next several semesters. The test instruments should be actual examinations administered in college-level classes—for example, programming classes. In our case, one possible test statistic is the correlation of the scores on different portions of examinations of basic programming skills containing different types of both MC and CR questions. However, we feel that a better test statistic are the R^2 values in multiple regression analyses. The experimental design and analysis are described more fully in Kuechler and Simkin (2004).

Implications

We note several important implications of our study. If empirical evidence supports our contention that student performance on MC and CR questions are strongly related when the questions examine same-level knowledge, instructional faculty will have an important justification for using only MC questions on student

examinations—the argument that such questions accurately test the same level of understanding of underlying course concepts. This seems especially significant for basic computer language programming and related instructional domains where structural fidelity issues have traditionally led many instructors to prefer labor-intensive CR tests.

With this implication come two important caveats. One is the need for instructors to carefully construct MC examination questions that test student understanding of the material at an appropriate cognitive level. The other is the need for instructors to content themselves with questions that limit the testing to such levels. As detailed in the section on knowledge levels, we and other authors believe that it is difficult to construct MC questions that go beyond the first three cognitive levels of Bloom's taxonomy, and thus perhaps limit the usefulness of MC questioning to undergraduate classes.

Another important implication of our work concerns the current, often indiscriminate use of the questions in existing test banks. At least for books in many lower division courses, the authors have found that these questions are typically organized by chapter headings and are rarely piercing. Until instructors require publishers to create questions by skill level, perhaps in accord with a uniform code as suggested in our earlier discussions, it seems likely that we will continue to use suboptimal testing procedures.

SUMMARY AND CONCLUSIONS

There are many reasons why many university students and faculty prefer MC test formats over most other kinds of test formats. But do MC questions measure the same level of student understanding as other types of questions—especially constructed response questions? Past studies attempting to answer this question empirically have yielded mixed results.

A potential problem with such research has been their treatment of MC questions as homogenous entities, an assumption supported neither by pedagogical experience nor prior research. We hypothesize that using MC questions to test student understanding of programming concepts can exhibit broad ranges of difficulty, probe different cognitive processes and, therefore, test different levels of understanding. The value of detecting a strong relationship between level-corrected MC and CR tests is high. If such a relationship can be demonstrated, for example, conscientious instructors can have the satisfaction of knowing that simple-to-administer, easy-to-grade MC tests are, in fact, doing the job desired of them—fairly and incisively evaluating student understanding of course concepts.

To explore this hypothesis, the authors operationalized Bloom's taxonomy of educational goals for our specific area of interest, basic computer language learning. This task consisted of interpreting the very broad statements of *evidence of accomplishment* of a knowledge level that were given in the original taxonomy in the context of a first and second course in programming languages for IS students. For undergraduate programming courses only the first three levels of Bloom's six-level taxonomy are readily applicable, given our interpretations of evidence of accomplishment. It is difficult to construct questions that probe the forth level, analysis, due to that format's requirement to choose a definite answer from a small

list of supplied answers. The highest two levels call for analytic skills that we interpret as more appropriate for Masters- and PhD-level studies, at least for the specific area of computer programming languages.

Using the programming language instruction focused taxonomy, we analyzed multiple MC test banks and instructor created MC exams, and categorized the questions. We found that, as hypothesized, MC questions from our domain could be categorized according to the broad range of cognitive levels of the taxonomy (knowledge levels). From this analysis, we abstracted a set of rules describing the construction of types of MC questions for each of the applicable levels of the taxonomy. Next we deconstructed programming constructed response (coding) questions from multiple examinations given in first and second courses in computer languages at our university. The result of each deconstruction was the skill set required for the solution of the question; each skill was then categorized according to the knowledge-level taxonomy we had developed, just as we had categorized the MC questions.

In comparing the results of the MC question categorization and the constructed response skill categorization, we were surprised to find that it is very difficult to compose MC items that can reach the highest knowledge level achieved by the more sophisticated CR questions, the *Application* level. MC questions, by the nature of their construction, preclude multiple production and if code segments are presented for analysis, always allow the answer to be determined by mentally tracing program execution. This is a Comprehension skill, and utilizes different cognitive processes than solution construction (Anderson & Krathwohl, 2001).

Recent studies that have mapped the knowledge levels of Bloom's taxonomy to specific cognitive skills (Anderson & Krathwohl, 2001) indicate that Comprehension-level MC questions have a greater overlap of cognitive skills with constructed response questions than do Knowledge-level MC questions. We are, therefore, encouraged to explore the hypothesis that an MC test can be constructed that will correlate more closely with constructed response performance on basic programming skills examinations than has been shown in previous research.

We are excited about the potential of the research to discover a means of determining basic programming understanding that is as effective as constructed response examination questions, but which has the statistical and grading tractability of MC examinations. Then, too, such demonstrated testing effectiveness speaks to the importance of carefully worded, *categorized* test banks that enable instructors to create such examinations—a commodity that does not, at present, appear to be widely available. [Received: March 2004. Accepted: August 2004.]

REFERENCES

ACM (multiple contributors). Model computing and information systems curricula (http://www.acm.org/education/curricula.html, last accessed July 7, 2004).

Anderson, L. W., & Krathworhl, D. R. (Eds.) (2001). *A taxonomy for learning, teaching, and assessing*. New York, NY: Longman.

Ashburn, R. (1938). An experiment in the essay-type question. *Journal of Experimental Education, 7*(1), 1–3.

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short answer questions in a marking context. *Journal of Marketing Education, 25*(1), 31–36.

Baranchi, A., & Cherkas, B. (2000). Correcting grade deflation caused by multiple-choice scoring. *International Journal of Mathematical Education in Science & Technology, 31*(3), 371–380.

Bar-Hillel, M., & Attali, Y. (2002). Seek whence: Answer sequences and their consequences in key-balanced multiple-choice tests. *The American Statistician, 56*(4), 299–303.

Barnett, S., & Ceci, S. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612–637.

Beard, R. M., & Senior, I. J. (1980). *Motivating students*. London, UK: Routledge & Kegan Paul.

Becker, W. E., & Johnston, C. (1999). The relationship between multiple choice and essay response questions in assessing economics understanding. *Economic Record, 75*(231), 348–357.

Becker, W. E., & Walstad, W. B. (1990). Data loss from pretest to posttest as a sample selection problem. *Review of Economics and Statistics, 72*(1), 184–188.

Becker, W. E., & Watts, M. (2001). Teaching methods in U.S. undergraduate economics courses. *Journal of Economic Education, 32*(3), 269–279.

Bell, R. C., & Hay, J. A. (1987). Differences and biases in English language examination formats. *British Journal of Educational Psychology, 57,* 212–220.

Bennett, R. E., Rock, D. A., & Want, M. (1991). Equivalence of free-repsonse and multiple-choice items. *Journal of Educational Measurement, 28*, 77–92.

Biggs, J. B. (1973). Study behavior and performance in objective and essay formats. *Australian Journal of Education, 17*, 157–167.

Bloom, B., Englehart, M., Furst, E., Hill, W., & Krathwohl, D. (1956). *A taxonomy of educational objectives, Handbook I: Cognitive domain.* New York, NY: David McKay Company.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement, 27*, 165–174.

Boodoo, G. M. (1993). Performance assessments or multiple-choice? *Educational Horizons, 72*, 50–56.

Bridgeman, B. (1991). Essays and multiple-choice tests as predictors of college freshman GPA. *Research in Higher Education, 32*, 319–332.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*, 253–271.

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement, 31*(1), 37–50.

Bridgeman, B., & Morgan, R. (1996). Success in college for students with discrepancies between performance on multiple-choice and essay tests. *Journal of Educational Psychology, 88*(2), 333–340.

Bridgeman, B., & Rock, D. (1993). Relationships among multiple-choice and open-ended analytical questions. *Journal of Educational Measurement, 30*(4), 313–329.

Brown, G., Bull, J., & Pendlebury, M. (1997). *Assessing student learning in higher education*. London, UK: Routledge.

Burton, R., & Miller, D. (2000). Why tests are not as simple as A, B or C. *Times Higher Education Supplement, 1421*, 42.

Bush, M. (2001). A multiple choice test that rewards partial knowledge. *Journal of Further and Higher Education, 25*(2), 157–163.

Carey, J. (1997). Everyone knows that $E = MC^2$ now, who can explain it? *Business Week, 3547*, 66–68.

Chan, N., & Kenedy, P. E. (2002). Are multiple-choice exams easier for economics students? A comparison of multiple choice and equivalent constructed response exam questions. *Southern Economic Journal, 68*(4), 957–971.

Delgado, A. R., & Prieto, G. (2003). The effect of item feedback on multiple-choice test responses. *British Journal of Psychology, 94*(1), 73–85.

Dufresene, R. J., Leonard, W. J., & Gerace, W.J. (2002). Making sense of students' answers to multiple-choice questions. *The Physics Teacher, 40*, 174–180.

Entwistle, A., & Entwistle, N. (1992). Experiences of understanding in revising for degree examinations. *Learning and Instruction, 2*, 1–22.

Epstein, M. L., & Brosvic, G. (2002). Immediate feedback assessment technique: Multiple choice test that behaves like an essay examination. *Psychological Reports, 90*(1), 226.

Epstein, M. L., Epstein, B. B., & Brosvic, G. M. (2001). Immediate feedback during academic testing. *Psychological Reports, 88*(3), 889–895.

Fellenz, M. (2004). Using assessment to support higher level learning: The multiple choice item development assignment. *Assessment and Evaluation in Higher Education, 29*(6), 703–719.

Fenna, D. S. (2004). Assessment of foundation knowledge: Are students confident in their ability? *European Journal of Engineering Education, 2*, 307–313.

Frary, R. B. 'The none-of-the-above' option: An empirical study. *Applied Measurement in Education, 4*(2), 115–124.

Fredricksen, N. (1984). The real test bias. *American Psychologist, 39*, 193–202.

Greene, B. (1997). Verbal abilities, gender, and the introductory economics course: A new look at an old assumption. *Journal of Economic Education, 28*(1), 13–30.

Grimes, P. W., & Nelson, P. S. (1998). The social issues pedagogy vs. the traditional principles of economics: An empirical examination. *The American Economist, 42*(1), 56–64.

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1), 37–50.

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *Journal of Experimental Education, 62*(2), 143–157.

Heckman, J. M. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153–161.

Hobson, A., & Ghoshal, D. (1996). Flexible scoring for multiple-choice exams. *The Physics Teacher, 34*(5), 284.

Holder, W. W., & Mills, C. N. (2001). Pencils down, computer up: The new CPA exam. *Journal of Accountancy, 191*(3), 57–60.

Kastberg, S. (2003). Using Bloom's taxonomy as a framework for classroom assessment. *Mathematics Teacher, 96*(6), 402–405.

Katz, I. R., Bennett, R. E., & Berger, A. E. (2000). Effects of response format on difficulty of sat mathematics items: It's not the strategy. *Journal of Educational Measurement, 37*, 39–57.

Kennedy, P. E., & Walstad, W. B. (1997). Combining multiple-choice and constructed-response test scores: An economist's view. *Applied Measurement in Education, 10*, 359–375.

Kniveton, B. H. (1996). A correlational analysis of multiple-choice and essay assessment measures. *Research in Education, 56*, 73–84.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice, 41*(4), 212–218.

Kreig, R. G., & Uyar, B. (2001). Student performance in business and economics statistics: Does exam structure matter? *Journal of Economics and Finance, 25*(2), 229–241.

Kuechler, W., & Simkin, M. (2003). How well do multiple choice tests evaluate student understanding in computer programming classes. *Journal of Information Systems Education, 14*(4), 389–399.

Larsen, J. (1997). Application of cognitive, affective and behavioral theories to measure learning outcomes in management training. PhD dissertation, University of South Florida, Tampa, FL, (unpublished).

Lipscomb, J. W., Jr. (1985). Is Bloom's taxonomy better than intuitive judgment for classifying test questions. *Education, 106*(1), 102–107.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement, 31*(3), 234–250.

Lumsden, K. G., & Scott, A. (1987). The economics student reexamined: Male-female difference in comprehension. *Journal of Economic Education, 18*(4), 365–375.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207–218.

Maxwell, N. L., & Lopus, J. S. (1995). A cost effectiveness analysis of large and small classes in the university. *Educational Evaluation and Policy Analysis, 17*, 167–178.

Mayer, R. (2002). A taxonomy for computer-based assessment of problem solving. *Computers in Human Behavior, 18*, 623–632.

Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1995). *Sex differences on constructed-response and multiple-choice sections of Advanced Placement Examinations: Three exploratory studies*. New York, NY: College Entrance Examination Board.

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In E. B. Randy & C. W. William (Eds.), *Construction versus choice in cognitive measurement*. Hillsdale, NJ: Lawrence Erlbaum, 61–73.

O'Neil, P. B. (1993). Essay versus multiple choice exams: An experiment in the principles of macroeconomics course. *American Economist, 45*(1), (Spring), 62–70.

Paul, H. (1982). The impact of outside employment on student achievement in macroeconomic principles. *Journal of Economic Education,* (Summer), 516.

Paxton, M. (2000). A linguistic rerspective on multiple choice questioning. *Assessment & Evaluation in Higher Education, 25*(2), 109–120.

Pleeter, S., & Way, P. K. (1993). *Economics in the News*, (2nd ed.). Reading, MA: Addison-Wesley.

Ramsden, P. (1988). Studying learning; improving teaching. In P. Ramsden (Ed.), *Improving learning: New perspectives*. London, UK: Kogan Page, 13–31.

Ray, M. A., & Grimes, P. W. (1992). Performance and attitudinal effects of a computer laboratory in the principles of economics course. *Social Science Computer Review, 10*(1), 42–58.

Reeves, M. (1990). An application of Bloom's taxonomy to the teaching of business ethics. *Journal of Business Ethics, 9*, 609–161.

Rogers, W. T., & Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement, 59*(2), 234–247.

Rubin, D. (1985). Memorability as a measure of processing: A unit analysis of prose and list learning. *Journal of Experimental Psychology, 114*(2), 213–238.

Saunders, P. (1991). *Test of understanding in college economics, third editions, examiner's manual*. New York: Joint Council on Economic Education.

Saunders, P., & Walstad, W. B. (1998). Research on teaching college economics. In *Teaching undergraduate economics: A handbook for instructors*. Boston, MA: Irwin/McGraw Hill, 141–166.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examinations versus assignment essay. *Higher Education, 35*, 453–472.

Siegfried, J. J., & Fels, R. (1998). Research on teaching college economics: A survey. *Journal of Economic Literature, 17*(3), 923–969.

Siegfried, J. J., & Walstad, W. B. (1998). Research on teaching college economics. In W. B. Walstad, & P. Saunders (Eds.), *Teaching undergraduate economics: A handbook for instructors*. Boston, MA: Irwin/McGraw Hill, 141–166.

Siegfried, J. J., & Kennedy, P. E. (1995). Does pedagogy vary with class size in introductory economics? *American Economic Review,* (Papers and Proceedings) *85*, 347–351.

Snow, R. E. (1993). Construct validity and constructed-response tests. In E. B. Randy, & C. W. William (Eds.), *Construction versus choice in cognitive measurement*. Hillsdale, NJ: Lawrence Erlbaum, 45–60.

Snyder, A. (2003). The new CPA exam: Meeting todays' challenges. *Journal of Accountancy, 196*(6), 11–120.

Soper, J. C., & Walstad, W. B. (1983). On measuring economic attitudes. *Journal of Economic Education, 14*(4), 4–17.

Soloso, R., & Biersdorff, K. (1975). Recall under conditions of cululative cues. *The Journal of General Psychology, 93*, 233, 246.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361–370.

Sykes, R. C., & Yen, W. M. (2000). The scaling of mixed-item format tests with one-parameter and two-parameter partial credit model. *Journal of Educational Measurement, 37*, 221–244.

Tate, R. (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*, 329–346.

Tombaugh, T., & Hubley, A. (2001). Rates of forgetting on three measures of verbal learning. *Journal of the International Neuropsychological Society, 7*(1), 79–91.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In E. B. Randy & C. W. William (Eds.), *Construction versus choice in cognitive measurement*. Hillsdale, NJ: Lawrence Erlbaum, 29–44.

Traub, R. E., & MacRury, K. (1990). Multiple-choice vs. free-response in the testing of scholastic achievement. *Tests And Trends, 8*, 128–159.

Tuckman, B. W. (1993). The essay test: a look at the advantages and disadvantages. *NASSP-Bulletin, 77*(555), 20–26.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*, 103–118.

Walstad, W. B. (1998). Multiple choice tests for the economics course. In B. W. William & S. Phillip (Eds.), *Teaching undergraduate economics: A handbook for instructors*. New York, NY: McGraw-Hill, 287–304.

Walstad, W. B., & Becker, W. E. (1994). Achievement differences on multiple choice and essay tests in economics *American Economic Review, 84*(2), 193–196.

Walstad, W. B., & Robson, D. (1997). Differential item functioning and male-female differences on multiplechoice tests in economics. *Journal of Economic Education, 28*, 155–171.

Welsh, A. L., & Saunders, P. (1998). Essay questions and tests. In B. W. William & S. Phillip (Eds.), *Teaching undergraduate economics: A handbook for instructors*. New York, NY: McGraw-Hill, 305–318.

Wesolowsky, G. O. (2000). Detecting excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics, 27*(7), 909–921.

Zeidner, M. (1987). Essay versus multiple-choice type classroom exams: The student's perspective. *Journal of Educational Research, 80*(6), (July/August), 352–358.

Zimmerman, D. W., & Williams, R. H. (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. *Applied Psychological Measurement, 27*(5), 357–371.

Mark G. Simkin is a professor of Computer Information Systems at the University of Nevada, Reno. He earned his MBA and PhD degrees from the University of California, Berkeley. His research in internet law, end-user computing, computer education, and computer crime appears in over 100 academic journal articles and fifteen books, including *Decision Sciences, The Journal of Accountancy, Communications of the ACM, and Communications of the Association for Information Systems*. Two of his most recent books are *Applications Programming in Visual Basic 5* (Half Moon Bay: Scott/Jones, 1998) and *Core Concepts of Accounting Information Systems* (New York: John Wiley and Sons, 2005). Professor Simkin has been a member of the Decision Sciences Institute since 1975.

William L. Kuechler is an associate professor of Information Systems at the University of Nevada, Reno. He holds a BS in Electrical Engineering from Drexel University and a PhD in Computer Information Systems from Georgia State University. His current research areas benefit from his twenty-year career in business software systems development and include 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, Information Technology and Management, Journal of Information Systems Education,* the proceedings of WITS, HICSS and other international conferences and journals. Dr. Kuechler is a member of AIS and ACM.