3-1-2016

Why Use Multiple Choice Questions with Excess Information?

Jason Bergner
University of Nevada, Reno

Joshua J. Filzen
Boise State University

Mark G. Simkin
University of Nevada, Reno

Publication Information

Bergner, Jason; Filzen, Joshua J.; and Simkin, Mark G. (2016). Why Use Multiple Choice Questions with Excess Information?. *Journal of Accounting Education, 34*, pp. 1-12. doi: 10.1016/j.jaccedu.2015.11.008

Why use multiple choice questions with excess information?

Jason Bergner
Assistant Professor of Accounting
College of Business
University of Nevada, Reno
Mail Stop 0026
Reno, NV 89557
jasonbergner@unr.edu

Joshua J. Filzen
Assistant Professor
Department of Accountancy
Boise State University
1910 University Drive
Boise, ID 83725-1610
joshuafilzen@boisestate.edu

Mark G. Simkin*
Professor of Information Systems
University of Nevada, Reno
Mail Stop 0026
Reno, NV 89557
markgsimkin@yahoo.com

* Corresponding Author: Ph: (775) 784-4840.

ABSTRACT

The examinations administered by accounting instructors, professional certification examiners, textbook writers, and preparatory accounting software all routinely include multiple-choice (MC) questions with excess (yet related) information. Despite their widespread use, little is known about how MC questions with excess information (hereafter MCE questions) affect student test performance. Based on an empirical analysis of the tests of 374 students in two introductory accounting classes at a single university, we found that average performance was lower on MCE questions than non-MCE questions, but was sensitive to the overall difficulty of the tested concept. We also found no significant difference in the power of the two question types to discriminate—both types appeared equally competent in differentiating between low- and high-performing students. Although accounting professors may wish to use MC questions with excess information for a number of other reasons, we found that MCE questions, as used in the present setting, do not appear to better discriminate student understanding relative to non-MCE questions.

Keywords:
Multiple-choice questions
Assessment
Excess information

1. **Introduction**

Many reasons motivate the widespread use of multiple-choice (MC) questions on accounting exams, including convenience (the ability to grade, record, and return tests quickly both in class and online—see, for example, Apostolou et al., 2009), the ability to cover a wide range of material within a single exam (Collier & Mehrens, 1985), and because these questions comprise the majority of the points on the most frequently-taken professional certification exams (AICPA, 2015). In this research note, we study how multiple-choice questions with excess information affect student performance.

We define "multiple-choice questions with excess information" (hereafter MCEs) as questions where the stem contains both the information needed to answer correctly and additional information that, although appearing to be relevant to the question's concept, should be ignored. Accordingly, we refer to MC questions that do not contain any excess information (i.e., the stem contains only the information needed to answer the question) as "non-MCEs." This definition fits the concept of many MCEs found on professional and accounting examinations, yet to our knowledge has not been studied. We formally discuss this concept in the second section of this paper, including contrasting the definition with well-known question flaws such as "window dressing."

There are many reasons why accounting educators might want to include MCEs in their classroom examinations. In this study, we focused on two: "difficulty" and "discrimination.[1]"

[1] Other reasons may include 1) the desire to assess student understanding at higher levels of cognition than the rote recall characteristic of many non-MCE questions, 2) the belief that MCEs might better prepare students for the professional certification examinations they may take once they become accounting professionals, and 3) the belief that MCE questions enjoy better *structural fidelity*—i.e., better reflect the skill sets required to perform accounting tasks on the job. While these are all testable hypotheses, they would require vastly different research designs and are beyond the scope of this paper.

Thus, our study falls primarily in the realm of item analysis within the broader field of psychometrics (Reynolds et al. 2009; Nunnally and Bernstein, 1994).

Reynolds et al. (2009) suggest that item difficulty is an important characteristic of assessment. If test questions are too easy or too hard, the exam may lack *validity* (i.e., be unable to adequately assess knowledge of the material). Thus, one reason MCE questions may be utilized by exam preparers is to adjust the difficulty level of the exam. Our first objective of this study was to determine whether MCEs are truly more difficult questions by examining the percentage of correct responses relative to paired non-MCE questions.

In addition to variation in difficulty, MCE questions might be better discriminators—i.e., might better distinguish between those who understand the material and those who do not (Reynolds et al., 2009). This concept is different from "difficulty," as two tests can have different levels of difficulty, but perform equally well at distinguishing superior students from poorer ones (with a difference only in the level of the score). Thus, how well a test question discriminates is useful to accounting instructors wishing to include high quality questions that separate knowledgeable students from those with imperfect knowledge.[2] Consequently, our second objective was to determine whether MCEs better discriminate among students than paired, non-MCE questions.

We found consistent results using two separate experiments. MCE questions were more difficult for students (measured as percentage of correct responses). However, we found that the difficulty of the tested concept impacts the degree to which excess information affects performance. When examining question discrimination (using the point-biserial correlation), we

[2] McMillan et al. (1989) use the term discrimination to describe the ability of a question to discriminate between students. We emphasize that discrimination is a term that applies to the question, with the result being that the question can better identify students with a higher level of knowledge from those with a lower level of knowledge.

found no significant difference between MCEs and their non-MCE counterparts. In other words, in these two experiments, MCE questions were no better at differentiating superior students from poorer ones, relative to non-MCE questions.

In the next three sections of this paper, we review the literature and develop our hypotheses, describe our research method, and present our results. In the final section, we summarize the study, discuss some limitations, and offer suggestions for future research.

2. Literature review and hypothesis development

2.1. Background

The basic structure of an MC question includes the question statement (the *stem*), followed by several choices. These choices include the correct response and a number of incorrect responses (*distractors*). Despite the wide spread use of MC questions in educational settings, professors have been found to have difficulty writing high quality MC questions (McMillian et al., 1989).[3] Recent literature on best practices for constructing effective MC questions using this common format is extensive (e.g., Gronlund, 2006; Haladyna, 2004; Haladyna et al., 2002; Osterland, 1998; Baldwin, 1984). The guidelines for constructing effective MC questions generally include (1) putting as much of the wording as possible in the stem in order to avoid repeating it in the choices (i.e., develop focused stems), (2) ensuring that the intended answer is the only correct one, and (3) creating distractors that are plausible to students with less-than-perfect understanding.

Despite this plethora of literature regarding the importance of constructing effective MC questions, the accounting literature is void of studies examining how student performance is

[3] The authors define quality as being of sufficient difficulty and discrimination.

affected by the use of MCE questions and variables pertaining to them. For instance, a search of *Issues in Accounting Education, Journal of Accounting Education,* and *Accounting Education: An International Journal* revealed no articles pertaining to this topic over the last five years. We also searched the *ERIC* database and found articles in other disciplines that are related to this topic, though nothing was found that directly addressed our questions.[4]

2.2. Definition of MC questions with excess information

For this paper, we define MCEs as questions in which the stem contains both the information required to answer the question and additional information that, although relevant to the topic of the question, is not needed to answer it. For example, consider the following questions, which we feel are typical of an introductory accounting examination question both without and with excess information:

> ***Non-MCE:*** *XYZ Company started the year with total assets of $400,000 and total liabilities of $300,000. During the year XYZ recorded $200,000 in revenues, $100,000 in expenses, and dividends of $40,000. The company issued no common stock for the period. What was stockholders' equity at the end of the year? (Four or more choices would follow.)*
>
> ***MCE****: XYZ Company started the year with total assets of $400,000 and total liabilities of $300,000. During the year XYZ recorded $200,000 in revenues, $100,000 in expenses, and dividends of $40,000. The company issued no common stock for the period* ***but purchased a piece of equipment for $60,000****. What was stockholders' equity at the end of the year? (Four or more choices would follow.)*

Although most accounting academics may be familiar with MCEs, it is important to distinguish them from such other terms in the literature as *window dressing, verbose questions,* or *red herrings,* all of which describe stems with information that is both irrelevant to the

[4] One paper is tangentially related to the topic of student performance in response to the color of the exam itself (Meyer and Bagwell, 2012). Other papers in non-accounting disciplines have investigated other variables pertaining to student performance, such as item/answer order (Schroeder et al., 2012) and finding "true" versus "not true" answers (Laprise, 2012).

specific question and unrelated to the concept in it (Haladyna, 2004; Case et al., 1996).[5] In contrast, our focus is on information that test takers cannot obviously exclude. There is a clear theoretical difference between these two concepts. However, we recognize that in practice the relevance of excess information is likely to fall on a continuum in which differences are subjective. While "window dressing" is likely to have a small diversionary effect upon test takers, most instructors are likely to consider their effects minor (Case et al., 1996). An example is:

> ***Window dressing question:*** *ABC Company recorded revenues of $100,000 and expenses of $50,000.* ***The company employs 40 people****. What is net income for the period? (Four or more choices would follow.)*

The number of employees is of no consequence in computing net income, and we would expect even weak accounting students to know this. Consequently, such diversions contrast with our definition of excess information, which refers to information that a student might believe is relevant to the question, but that in fact is not.[6]

3. Hypotheses

Adding excess information to MC questions may increase the difficulty of the question. This seems intuitive, as it forces students to process at least one additional piece of information. Prior research has studied test bank quality, and finds that accounting test banks contain numerous questions that hinder clarity (Hansen and Dexter, 1997; Moncada and Harmon, 2004). To provide further insight on this for our setting in particular, we analyzed a sample chapter's (accrual accounting concepts) test bank from a well-known publisher of an introductory financial

[5] "Window dressing" is also known as "unnecessary information" (Tarrant & Ware, 2008), "superfluous information" (Case & Swanson 2001), or "extra information" (Case et al., 1996).

[6] Our study also contrasts with work on *unfocused stems*, which are questions where the central idea of the question is not included in the stem (Haladyna, 1997; Downing, 2005).

accounting text to better understand how excess information was used in the test bank. Of the 181 MC questions in the chapter, we found what we believe to be 28 MCEs (15.5 percent).

The test bank also attached difficulty levels to each question of "easy," "medium," or "hard." The test bank authors identified their 28 MCE questions as follows: 1 easy, 21 medium, and 6 hard. The percentage of questions in each category that were MCEs were as follows: percentage of easy questions that were MCE is one percent (1 of 91 questions defined as "easy"), percentage of medium questions that were MCE is 27 percent (21 of 79 questions defined as "medium"), and percentage of hard questions that were MCE was 55 percent (6 of 11 questions defined as "hard"). Thus, as the (perceived or author-indicated) difficulty level increases in the test bank, questions are more likely to be MCEs.

Based on our earlier discussion of the intuition of MCEs being more difficult, along with the analysis above, our first hypothesis is (stated in alternative form):

H_1: *MCEs are more difficult than paired MC questions.*

As discussed earlier, educators might also want to include MCEs on exams with the belief that MCEs are better discriminators in determining the relative knowledge of the test takers (an important quality of any examination). A priori, there are at least two reasons to believe MCEs may be more successful at discrimination. First, non-MCE questions could be too easy, thus limiting a test's ability to distinguish between those who know the material well and those who do not. Second, since MCEs contain excess information, students who know the concept better will be more likely to recognize and ignore the excess information. Thus, MCE questions would seem to favor students that understand a concept better. Because there is scant research on this, we state our second hypothesis in null form:

H_2: *MCEs are no more discriminating than paired MC questions.*

4. Methodology

4.1. Setting and design

The setting for the experiments was two introductory financial accounting courses that students typically take in their sophomore (second) year. The two experiments that follow were conducted in two separate sections taught by the same professor. Participants were students from a western, public, land-grant institution. The instructor taught both classes using face-to-face lectures with online homework designed to help students master the subject material.

While the overarching goals of both experiments were the same, the research designs were different. In the first experiment, all students received the same examination (i.e., there was only one version of the exam). The exam content was typical of past exams given to introductory accounting students and was based on the content covered to that point in the course. Each concept in the exam was tested using a matched pair of questions, one MCE and a non-MCE counterpart that was subjectively matched by the authors. This design permitted the comparison of each student's performance on MCE versus non-MCE for each pair of questions, allowing each student to serve as his/her own control.

A weakness with this design is that it requires the researchers to create question pairs that cover the same concepts but are not identical. Otherwise, test questions that are identical, save for the presence of excess information, may enable attentive test takers to recognize such matching and create noise within the experiment (i.e., differences in students' performances on MCE v. non-MCE questions would be correlated with their recognition of similarities between question pairs). In turn, such interactive effects could potentially distort test results and therefore cloud interpretations of student performance. Thus, in our first experiment, we chose to use matching MCE questions that tested the same concept as non-MCE questions but were not identical.

While pairing questions that are not identical allows a student to serve as his/her own control and addresses the issue of students recognizing question pairs on the exam, the subjective nature of pairing "similar-but-not-identical" questions creates an issue about whether a given MCE question is a suitable conceptual equivalent for its non-MCE counterpart. We recognize that differences in content, scope, or even wording can affect student performance. Thus, the first experiment has strengths and weaknesses. While each student answered both questions in each pair (the strength), the questions comprising the pair are not identical (the weakness). We addressed this weakness with a second experiment.

The second experiment used identical questions within a pair, except for the excess information. That is, the MCE and non-MCE questions contained identical wording, with the MCE question having additional information.[7] All of the pairs of questions were then placed into two different versions of the quiz, with each version containing one question out of each pair. Further, the total number of MCE and non-MCE questions was evenly distributed across the two versions. Thus, each test version contained half non-MCE questions and half MCE questions. Half the class took one version of the test and half the class took the other version. This design effectively holds the questions constant, but no student's quiz contained both questions from any pair.

4.2.

Experiment 1

In the first experiment, 206 undergraduate students took an in-class interim examination, one of four interim exams given in the semester. As explained above, the authors created a 40-

[7] Both questions in the pair included a fifth distractor (i.e. choice), "none of the above", to operationalize exact pairing.

question exam (consistent with all other exams) which included 15 question pairs (15 MCEs and a non-MCE counterpart).

To avoid potential *order bias*, the questions were randomly distributed in the examination.[8] Although participants took the exam as part of their course grade, they could opt out of having their data included in the study per the instructions on the cover sheet.[9] All students completed the exam within the allotted 75 minutes.[10] There was no penalty for guessing, and the vast majority of students answered all questions.

4.3. Experiment 2

The second experiment comprised 168 undergraduate students in a twice-weekly financial accounting class, as in Experiment 1.[11] For this experiment, the test instrument was a scheduled quiz covering the same material as the first experiment. In this experiment there were 20 pairs of questions (40 total questions) split evenly between two versions of the quiz. The key difference of this experiment is that no single version contained a complete pair. Students were told the quiz was extra credit, with a point value approximately equal to one percent of their grade. *After* taking the quiz, students were informed that points would be awarded on an all-or-nothing basis contingent on an appropriate participation level. Thus, before the quiz students were not specifically aware of how the points would be allocated. In addition, attendance was not

[8] *Order bias* refers to the idea that the order in which information is presented can affect performance/decisions. For instance, if a non-MCE question was always presented first, students may more easily pick out the excess information from the related MCE question. Thus, presentation order, if not randomized, could potentially bias the results. See Balch (1989) and Richiutte (1992), respectively, for deeper discussions of this issue in education and accounting contexts.

[9] No students chose to have their data excluded from the study.

[10] One student had extra time due to a disability, but the inclusion of this student's responses did not qualitatively affect the results.

[11] The class had 201 students enrolled, yielding an 84 percent participation rate. 176 students technically participated, which translates to a rate of 88 percent. However, six observations were excluded due to scores at or below the expected value of guessing. One observation was excluded due to arriving after the experiment had begun. Finally, one observation was excluded from analysis due to that student having a documented disability, but additional time was not given to the student (due to time constraints). This did not affect that student's point allocation.

mandatory, so we believe students in attendance had an incentive to participate with effort, as evidenced by 97 percent of participants earning scores greater than the expected value of guessing.

The instructor administered the quiz with prior notice to the students. Although participants were required to provide identifying information to allow for a credit award, they could choose to opt out of this study per the instructions on the cover sheet.[12] All students completed the quiz within the allotted 40 minutes. There was no penalty for guessing, and the vast majority of students answered all of the questions.

5. Results

5.1. Descriptive statistics

A total of 206 students participated in our first experiment. Of these 117 (56.8%) were male and 89 (43.2%) were female. The composition of intended majors for these participants was: accounting: 35 (17.5%), other business majors: 133 (64.6%), nonbusiness majors: 37 (18.0%).

A total of 168 students participated in our second experiment. Of these, 91 (54.2%) were male and 77 (45.8%) were female. The composition of intended majors for these participants was: accounting: 27 (16.1%), other business majors: 111 (66.1%), nonbusiness majors: 30 (17.9%).

5.2. Evidence on H_1

Experiment 1 consisted of 15 non-MCE questions, and 15 matched MCE questions. In this experiment, students performed better on the non-MCE questions compared to the MCE

[12] No students chose to opt out of the study.

questions (60.7% vs. 48.0%, $t = 12.77$, $df = 205$, $p < 0.001$). [13] Our analysis for Experiment 2 produced similar results, with students performing better on the non-MCE questions (62.0% v. 40.8%, $t = 13.48$, $df = 167$, $p < 0.001$). Overall, the results support H_1: on average MCE questions are more difficult relative to paired non-MCE questions.[14]

We also examined the detailed results to better understand this relationship. Table 1 reports detailed results for Experiment 1 by question pair (15 pairs), sorted by non-MCE question scores from high to low. Question numbers range from 4-43, with gaps. We included three demographic questions as the first three questions, and included 10 additional questions not related to the experiment in the examination to increase breadth.[15] Other than the demographic questions, the placement of the test questions was randomized in the test. The table reports average percentage scores by pair as well as a question-level discrimination measure, the point-biserial correlation (rpbs).[16]

[INSERT TABLE 1 ABOUT HERE]

Figure 1 expresses Table 1 graphically by presenting percentage scores in order from easiest to most difficult non-MCE average score. In addition, for each non-MCE, Figure 1 displays the average score earned for its paired MCE score. For this graph, the insertion of excess information consistently reduces the average when the non-MCE score is about 60% or

[13] The statistical tests for H1 and H2 are untabulated for brevity.
[14] MCE questions were observed to be more difficult on average when the sample was broken down by gender and by intended major in all categories for both experiments, suggesting the results are not driven by one of these groups in particular (untabulated).
[15] Generally, these questions were also more difficult to pair with an MCE "partner." For example, a question asking the definition of a term.
[16] The point-biserial correlation (rpbs), as applied in the present context, measures how well higher performing students did on a particular question compared to lower performing students. Student performance on all other questions, save the question under analysis, is computed and compared via a correlation to the question under analysis. Positive correlations denote that higher-achieving students did better on the question under analysis than lower-achieving students. Negative correlations suggest the opposite, and a correlation of zero indicates no discrimination. Ideally, exam questions should have large positive discrimination scores. Scores below 0.10 are generally considered to be unacceptable (Prometric, 2015).

higher. Below this percentage, however, Figure 1 shows a more unstable pattern. Stated another way, the observed differences in test performance are strongest when non-MCE questions are sufficiently easy. We believe this finding is intuitive: as a question becomes more difficult, the effect of adding excess information becomes less relevant. For this graph, the insertion of excess information consistently reduces the average when the non-MCE score is about 60% or higher.

[INSERT FIGURE 1 ABOUT HERE]

Table 2 reports the same statistics for Experiment 2 (20 pairs), again sorted by non-MCE question scores from high to low. Question numbers range from 4 to 20, as three demographic questions were included in this experiment as well. The pairs were distributed evenly across two versions, such that each version had 10 non-MCEs and 10 MCEs (neither version contained a complete pair). Average percentage scores by question and the point-biserial correlation are included in the table.

[INSERT TABLE 2 ABOUT HERE]

Figure 2 presents a graph of percentage scores for Experiment 2 (from Table 2) sorted in descending order of non-MCE average score, with the paired MCE score overlaid. This figure depicts a pattern that is similar to that observed in Experiment 1. Again, as question difficulty on non-MCE questions increases, the differences in average scores between the pairs becomes less predictable. These results again suggest that adding excess information to an MC question only increases difficulty when the underlying non-MCE question is sufficiently easy. As with Experiment 1, this graph shows that adding excess information consistently reduces the average when the non-MCE score is approximately 60% or higher.

[INSERT FIGURE 2 ABOUT HERE]

Visually comparing Figure 1 to Figure 2, the pattern of difficulty across question pairs appears generally more volatile in Figure 1. We attribute this to the nature of the two experiments. Specifically, the more abstract pairing of questions in Experiment 1 may account for this more volatile pattern relative to the identical pairing in Experiment 2.

Additionally, we examined what might be driving the differences observed to try to get a sense of the types of questions or concepts where excess information matters most. Specifically, we sorted the questions by performance difference between each MC and its paired MCE and examined the individual questions to identify any observable trends. The most obvious trend was already discussed above; overall question difficulty affects whether excess information has a dramatic impact on performance. Beyond that, we noticed the questions with the largest performance differences appeared to be questions about changes in accounts and account classifications (including the accounting equation). Questions where performance differences were smallest included adjusting entry concepts and normal balances. In Experiment 2, there was a trend towards end of exam questions having lower performance differences, perhaps due to test fatigue. Specifically, the five questions with the lowest performance differences all came from the last half of the exam.

Overall, our evidence suggests that MCE questions are more difficult than non-MCE questions. However, both experiments confirm that as the average score for non-MCE questions decreases (below about 60% based on our results), this pattern becomes less predictable. This suggests that the effects of adding excess information to MC questions are strongest (in terms of lowering test performance) when the non-MCE questions are relatively easy.

5.3. Evidence on H_2

In this section we explore whether MCE questions are better discriminators relative to non-MCE questions, utilizing the point-biserial correlation (rpbs) measure (Prometric, 2015; Baldwin, 1984). The point-biserial correlation is a special type of correlation (as opposed to the oft-used Pearson correlation coefficient) that is often used in situations involving a dichotomous variable (Reynolds et al., 2009). This is the case with many tests in which student answers are either correct or incorrect. We took the point-biserial discrimination scores from Tables 1 and 2, and utilized Fisher's r-to-z transformation to first transform the point-biserial correlation coefficients in order to allow for comparisons between group levels (Kenny, 1987).

Overall, we found no statistical difference in average discrimination scores between non-MCE and MCE questions (0.38 v. 0.43, $t = 0.40$, $df = 14$, $p = 0.68$) when examining all 15 question pairs from that experiment. However, as additional analysis, we take into consideration the guidelines for identifying unacceptable questions set forth by Prometric (Prometric, 2015). The guidelines suggest questions that are (1) too easy (more than 95% get the question correct), (2) too difficult (less than 20% get the question correct), or (3) are poor discriminators (point-biserial correlation of less than 0.10) should be identified as unacceptable and be ignored. Following these guidelines, we identified two pairs of questions as "unacceptable."

After excluding these two pairs from the analysis we again found no statistical difference between non-MCE and MCE questions using this subsample (0.40 v. 0.45, $t = 1.439$, $df = 13$, $p = 0.17$). Finally, we repeated this analysis a third time using the subsample of only questions where the non-MCE questions were consistently less difficult relative to their MCE counterparts (i.e., where at least 60% of the students answered the non-MCE correctly). Using this subsample, again there is no statistical difference in discrimination (0.37 v. 0.43, $t = 0.631$, $df = 6$, $p = 0.53$).

We conducted an analogous analysis for Experiment 2. Again, we find no statistical difference in average discrimination scores between non-MCE and MCE questions for the entire sample of 20 matched pairs (0.31 v. 0.41, $t = 0.774$, $df = 19$, $p = 0.44$). As with Experiment 1, we also performed the analysis for subsamples of pairs where the questions would be deemed acceptable under Prometric standards and where non-MCE questions were consistently less difficult relative to MCE questions. Consistent with our earlier findings, we found no statistical difference for the either the former subsample (0.33 v. 0.34, $t = 0.122$, df = 17, $p = 0.90$) or the latter (0.29 v. 0.34, $t = 1.089$, df = 9, $p = 0.30$).

Overall, our evidence suggests that no statistically significant discrimination gains (or losses) result from including excess information in MC questions. What this means is that instructors seeking to better differentiate between stronger or weaker students are not likely to find MCEs to be particularly effective tools for accomplishing these goals. Thus, at least within the confines of our sample data, "better discrimination" does not appear to be a good reason to include excess information in MC test questions. We speculate that this may be attributable to student confusion about the underlying concepts tested by the particular questions of our sample tests, unfortunate priming in the composition of our test questions (Kahneman, 2012), or perhaps a natural outcome when test questions are sufficiently difficult (McMillan et al., 1989). Clearly, more research is required here.

6. Summary, limitations, and suggestions for future research

Despite their widespread use, there is little research on how MCE questions affect accounting student performance. To address this deficiency, we conducted two different experiments across two semesters of a large introductory accounting course. The two experiments produced consistent results. Not surprisingly, we found that MCE questions lower

exam performance relative to paired non-MCE questions. However, further investigation revealed that this effect seems to apply only if the underlying concept being tested is sufficiently easy. In other words, as performance on non-MCE questions decreases, the observed pattern of MCE questions being more difficult deteriorates.

In addition, we found no statistical difference between the average discriminability of MCE questions relative to non-MCE questions. In other words, MCE questions appear to be no better than non-MCE questions in their ability to identify higher (or lower) achieving students. In a nutshell, our results suggest that accounting tests that contain MCEs are likely to require more generous class curves than their non-MCE counterparts, but are *not* likely to produce assessment metrics that better discriminate among the abilities of the students who take them.

We note that, despite the use of two experiments to alleviate certain disadvantages of various experimental design choices, a number of factors limit our findings. One concern is that the individuals in each of our samples consisted of the students taking only one accounting course, taught by one instructor, and at one university. Because these were sections of an accounting class required of all business majors and both of comparatively large class size, we believe that our sample consisted of a wide cross section of typical college-level students. In addition, we conducted two different experiments during two different semesters, which further mitigates this concern. Nonetheless, we recognize the possibility that these two classes were somehow not representative of alternate student bodies at other schools.

A second concern is that we conducted our experiments as an exploratory study that did not control for many of the factors that might also affect test performance and therefore our study results. The subject domain, the number of test questions, the perceived (as opposed to actual) difficulty of the questions, and even the quality of our questions' phrasing are among the many

variables that can also influence student performance. Similarly, we did not account for general intelligence, test-taking savvy, degree of risk aversion, or maturity of our test takers. Although our sample sizes in each experiment were relatively large and thus we don't expect significant bias in these areas, we encourage further study.

A third limitation is that our study focused entirely on MC questions—not fill-in-the-blank, matching, computational, or any number of additional test formats in which "excess information" can also play a role. Again, the rationale for our focus on MC questions is the many academic and certification tests that rely entirely, or in large part, on MC formats for their assessment tasks. But we also recognize that this focus also limits our findings to a particular question format.

A fourth concern is the possibility of *demand effects* in our experiment—i.e., changes in student performance or behavior attributable to the experiment itself. We feel such effects were unlikely here because the students expected the instructor to be present during the class periods where the experiments were administered and the instructor did not make any class-wide announcements once the experiments began. However, because the instructor did respond to the few students who individually asked questions during the experiments, we must consider the possibility of a small demand effect and we therefore mention it as a possible caveat for our results.

Finally, we note that although our student populations were sizable, our question samples were not. For example, our first experiment used 15 pairs of questions, while our second experiment only used 20 pairs. We admit the possibility that what explains the instability in the observed paths of the student answers in Figures 1 and 2 had less to do with our student population, and more to do with our choice of questions. This makes us wonder whether

different questions would result in more consistent patterns of student performance or would better discriminate among students. These concerns speak to the larger issue of which MCEs make good test questions (however this is defined). Again, more study is required to answer this.

Because research into the effects of MCE questions is limited, we encourage further research and propose a research agenda for consideration. Some promising areas that were not tested in this paper include other reasons why questions with excess information may be useful, such as testing whether MCEs have more structural fidelity (i.e. are better able to mimic future job conditions than non-MCE questions) and testing whether including MCEs in the accounting courses better prepares students for certification exams. Studying the effects of MCEs in other settings to explore how well results generalize is also an important area for future work. This includes experimenting in different courses (including courses taught at higher class levels), different topics or disciplines, as well as different question types (e.g., constructed response or fill-in-the-blank questions). We also recognize that other instructors have different testing styles, and that replicating our study with additional controls or variations may provide further insights.

Finally, a deeper understanding of how students process excess information, rather than a focus on assessment (such as in this paper) would also be an interesting area for more research. For example, at present, we are not able to explain *why* some MC questions result in consistently better performance than their MCE counterparts, while others do not. Overall, little is known about the effects of MCEs, despite their widespread use. We consider our work as a first step in more fully understanding the effects of MCEs and look forward to future work in this area.

Acknowledgements

The authors are indebted to the hard work of graduate assistants Greg Hill, Joanna Kardys-Stone, Lisa Rosen, Ross Granahan, Shaojie Wu, and Andrew Jenkins for assistance with various aspects of this study. We thank Mark Jackson, David Stout (the editor), an anonymous associate editor, and two anonymous reviewers for helpful comments and suggestions.

References

AICPA, (2015). Uniform CPA examination FAQ's: Content, structure, and delivery. Downloaded from *http://www.aicpa.org/BecomeACPA/CPAExam/ForCandidates/FAQ/Pages/computer_faqs_2.aspx, October 2015.*

Apostolou, B., Blue, M. A., & Daigle, R. (2009). Student perceptions about computerized testing in introductory managerial accounting. *Journal of Accounting Education, 27*(2), 59-70.

Balch, W.R. (1989). Item order affects performance on multiple-choice exams. *Teaching of Psychology, 16*(2), 75-7.

Baldwin, B. A. (1984). The role of difficulty and discrimination in constructing multiple-choice examinations: With guidelines for practical application. *Journal of Accounting Education, 2*(1), 19-28.

Case, S. M., & Swanson, D. B. (2001). *Constructing Written Test Questions for the Basic and Clinical Sciences* (3rd ed.). Philadelphia, PA: National Board of Medical Examiners.

Case, S. M., Swanson, D. B., & Becker, D. F. (1996). Verbosity, Window Dressing, and Red Herrings: Do They Make a Better Test Item? *Academic Medicine, 71*(10), S28-S30.

Collier, H. W., & Mehrens, W. A. (1985). Using multiple-choice test items to improve classroom testing of professional accounting students. *Journal of Accounting Education, 3*(2), 41-51.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education, 10*, 133-43.

Gronlund, N. E. (2006). *Assessment of Student Achievement.* Pearson Education, Inc.

Haladyna, T.M. (1997). *Writing Test Items to Evaluate Higher Order Thinking*. Needham Heights, MA: Allyn & Bacon.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items.* New York, NY: Routledge press.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-34.

Hansen, J. D., & Dexter, L. (1997). Quality multiple-choice test questions: item-writing. *Journal of Education for Business,* 73(2), 94-97.

Kahneman, D. (2012). The marvels of priming. In *Thinking fast and slow* (pp. 52-58). New York: Farrar, Straus and Giroux.

Kenny, D. (1987). *Statistics for the behavioral and social sciences.* New York, NY: Little Brown.

Laprise, S.L. (2012). Afraid not: Student performance versus perception based on exam question format. *College Teaching*, 60: 31-6.

McMillan, J. R., G. A. Mundrake, and S. A. McGuire (1989). Multiple-choice tests for the business school—idealism versus reality. *The Delta Pi Epsilon Journal* 31(4), 174-181.

Meyer, M.J. and J. Bagwell. (2012). The non-impact of paper color on exam performance. *Issues in Accounting Education* 27(3): 691-706.

Moncada, S. M., & Harmon, M. (2004). Test item quality: an assessment of accounting test banks. *Journal of Accounting & Finance Research*, 12(4), 28-39.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. McGraw-Hill, Inc.

Osterlind, S. J. (1998). *Constructing Test Items: Multiple-Choice, Constructed-Response, Performance, and Other Formats.* Kluwer Academic Publishers.

Prometric. (2015). https://www.prometric.com/en-us/news-and-resources/reference-materials/pages/Internal-Psychometric-Guidelines-for-Classical-Test-Theory.aspx, May 2015.

Reynolds, C. R., Livingston, R. B., & Willson, V. (2009). *Measurement and Assessment in Education*. Upper Saddle River, NJ: Pearson Education.

Richiutte, D.N. (1992). Working-paper order effects and auditors' going-concern decisions. *The Accounting Review, 67*(1), 46-58.

Schroeder, J., K.L. Murphy, T.A. Holme. (2012). Investigating factors that influence item performance on ACS exams. *Journal of Chemistry Education*, 89(3): 346-50.

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high stakes nursing assessments. *Medical Education*, 42, 198-206.

Figure 1.
Percentage correct for non-MCE and MCE questions, Experiment 1

Figure 1 presents a graph of the percentage of correct responses from the 15 question pairs from Experiment 1 sorted from easiest to most difficult for non-MCEs.

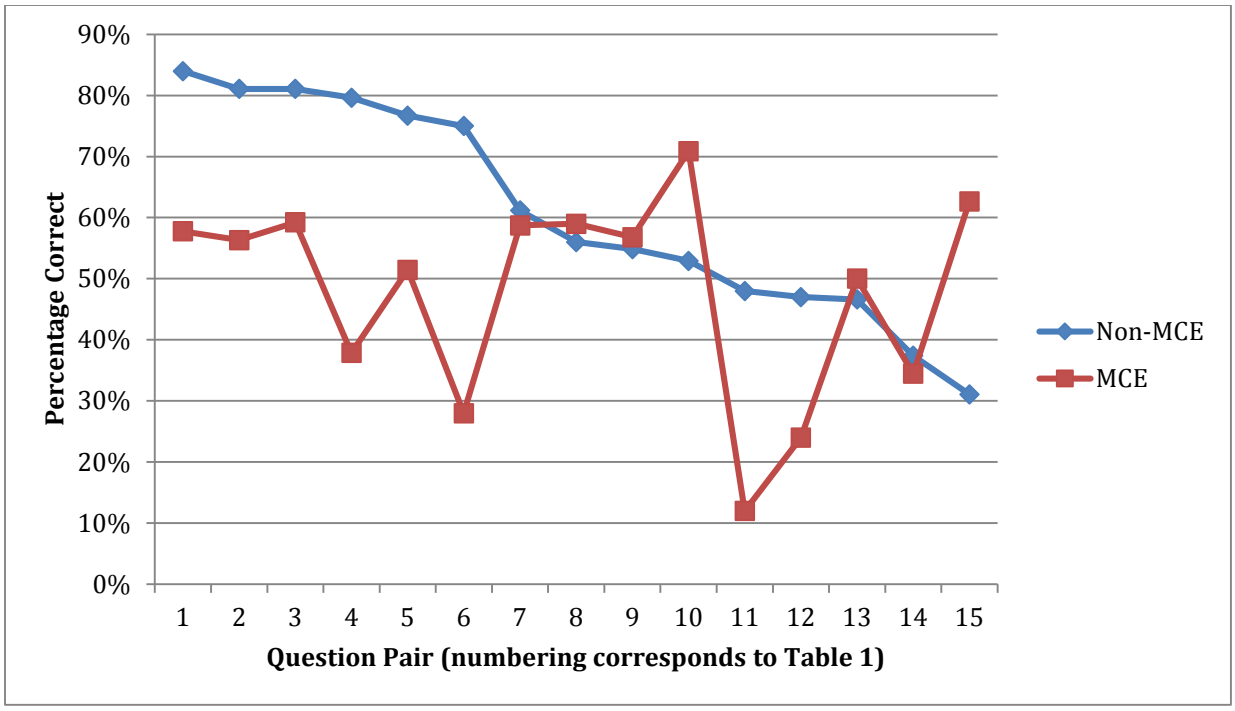

Figure 2.
Percentage correct for non-MCE and MCE questions, Experiment 2

Figure 2 presents a graph of the percentage of correct responses from the 20 question pairs from Experiment 2 sorted from easiest to most difficult for non-MCEs.

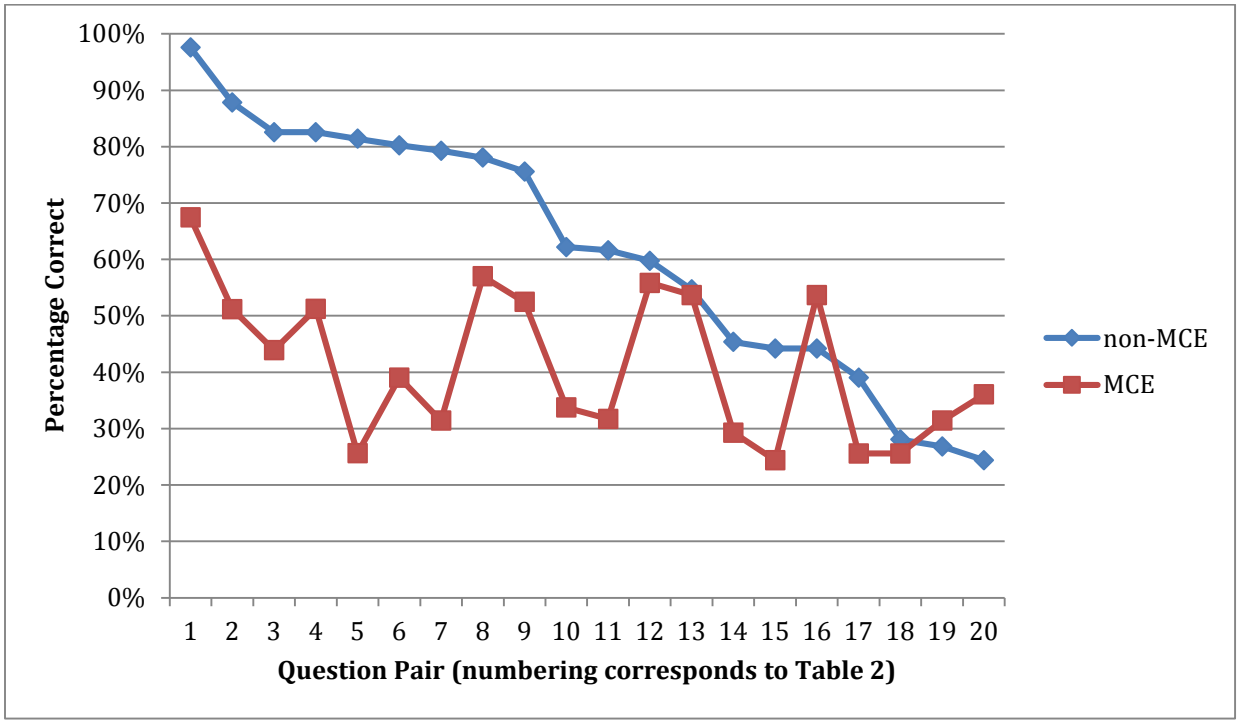

Table 1.
Results for question pairs, Experiment 1

This table presents the 15 MCE and Non-MCE question pairs, including their numbering on the actual exam for Experiment 1 sorted by non-MCE question scores from high to low. Also included is the percentage of correct responses by question, the point-biserial as a measure of how well the question discriminates, and differences between percentage correct by pair with t-statistics and p-values.

	non-MCE question				MCE question						
Pair	**Question Number**	**Percentage Correct**	**Point-biserial**	**N**	**Question Number**	**Percentage Correct**	**Point-biserial**	**N**	**Difference**	**T-statistic**	**P-value**
1	11	83.9%	0.42	206	28	57.8%	0.41	206	26.1%	6.5	0.000
2	29	81.6%	0.25	206	24	59.2%	0.38	206	22.4%	5.1	0.000
3	18	81.1%	0.37	206	4	56.3%	0.37	206	24.8%	6.1	0.000
4	33	79.6%	0.34	206	37	37.9%	0.36	206	41.7%	10.6	0.000
5	27	76.7%	0.43	206	9	51.5%	0.54	206	25.2%	6.2	0.000
6	13	75.7%	0.20	206	26	28.1%	0.29	206	47.6%	10.8	0.000
7	42	61.2%	0.45	206	8	58.7%	0.48	206	2.5%	0.7	0.516
8	40	56.8%	0.19	206	34	59.2%	0.32	206	-2.4%	-0.6	0.575
9	21	54.9%	0.40	206	35	56.8%	0.49	206	-1.9%	-0.3	0.757
10	7	52.4%	0.51	206	6	70.9%	0.47	206	-18.5%	-5.0	0.000
11	15	47.8%	0.22	206	5	12.1%	0.14	206	35.7%	8.7	0.000
12	10	47.4%	0.53	206	16	24.2%	0.38	206	23.2%	5.3	0.000
13	36	46.1%	0.40	206	43	50.0%	0.55	206	-3.9%	-1.0	0.333
14	19	37.3%	0.29	206	38	34.5%	0.40	206	2.8%	0.6	0.546
15	12	31.1%	0.42	206	17	62.3%	0.63	206	-31.2%	-7.5	0.000

Table 2.
Results for question pairs, Experiment 2

This table presents the 20 MCE and Non-MCE question pairs, including their numbering on the actual exam for Experiment 2 sorted by non-MCE question scores from high to low. Also included is the percentage of correct responses by question, the point-biserial as a measure of how well the question discriminates, and differences between percentage correct by pair with t-statistics and p-values.

	non-MCE question				MCE question						
Pair	**Question Number**	**Percentage Correct**	**Point-biserial**	**N**	**Question Number**	**Percentage Correct**	**Point-biserial**	**N**	**Difference**	**T-statistic**	**P-value**
1	13	97.5%	0.16	82	13MCE	67.4%	0.26	86	30.1%	5.5	0.000
2	17	87.8%	0.31	82	17MCE	51.2%	0.41	86	36.6%	5.6	0.000
3	16	82.6%	0.42	86	16MCE	51.2%	0.30	82	31.4%	4.6	0.000
4	4	82.5%	0.12	86	4MCE	43.9%	0.25	82	38.6%	5.7	0.000
5	18	81.4%	0.19	86	18MCE	25.6%	0.19	82	55.8%	8.7	0.000
6	8	80.2%	0.22	86	8MCE	39.0%	0.41	82	41.2%	6.0	0.000
7	11	79.2%	0.33	82	11MCE	31.3%	0.34	86	47.9%	7.1	0.000
8	19	78.0%	0.30	82	19MCE	56.9%	0.32	86	21.1%	3.0	0.003
9	12	75.6%	0.35	86	12MCE	52.4%	0.44	82	23.2%	3.2	0.002
10	7	62.2%	0.32	82	7MCE	33.7%	0.15	86	28.5%	3.8	0.000
11	6	61.6%	0.24	86	6MCE	31.7%	0.42	82	29.9%	4.0	0.000
12	9	59.8%	0.35	82	9MCE	55.8%	0.45	86	4.0%	0.5	0.608
13	14	54.7%	0.56	86	14MCE	53.7%	0.43	82	1.0%	0.1	0.898
14	10	45.3%	0.41	86	10MCE	29.3%	0.32	82	16.0%	2.2	0.032
15	22	44.3%	0.22	86	22MCE	53.7%	0.17	82	-9.4%	-1.2	0.222
16	20	44.2%	0.35	86	20MCE	24.4%	0.12	82	19.8%	2.7	0.007
17	5	39.0%	0.19	82	5MCE	25.5%	0.35	86	13.5%	1.9	0.063
18	21	28.0%	0.03	82	21MCE	25.6%	0.20	86	2.4%	0.4	0.720
19	23	26.8%	0.41	82	23MCE	31.4%	0.24	86	-4.6%	-0.6	0.518
20	15	24.0%	0.32	82	15MCE	36.0%	0.37	86	-12.0%	-1.6	0.102