Partial Credit on Multiple-choice Exams: Does It Help or Hurt Students?

Mariah Moss
MariahMoss5@gmail.com

Vic Matta
Matta@ohio.edu*

Raymond Frost
Frostr@ohio.edu

Janna Chimeli
chimeli@ohio.edu

Analytics & Information Systems Department
College of Business, Ohio University
Athens, OH 45701

Abstract

One criticism of multiple-choice tests is the all or nothing nature of credit on responses. Written assessments, by contrast, allow students to earn partial credit for partial knowledge. This study investigates if a multiple-choice test designed to reward partial credit for partial knowledge would benefit or harm student grades. Surprisingly, the study demonstrates that partial credit multiple-choice exams actually harm student grades. An analysis of thousands of records shows that students who are reasonably confident in their answers lose points by distributing their answer choices. Interestingly, as the semester progressed, students figured this out and modified their behavior accordingly—opting for a more traditional "all in" response pattern. Additionally, the partial credit form may discourage students from preparing for a test since they view the partial credit as a safety net. The only possible advantage of the partial credit exam is that it may encourage students to reflect on their responses. However, we conclude that partial credit multiple-choice exams should be avoided.

Keywords: wagering, diversifying, multiple-choice, test difficulty, gender, confidence

1. INTRODUCTION

Over the last century, multiple-choice testing gained its popularity due to its ability to enable teachers to grade tests with large numbers of questions and/or assess a large number of students efficiently (Schermerhorn, Gardner, and Dresdow 1992). However, multiple-choices test also have problems:

- They do not reward partial knowledge. Conversely, they do not let the instructor know the extent of each student's knowledge.
- They allow students to gain points through guessing.
- They tend not to elicit reflective learning.
- They have a higher likelihood to measure only lower level thinking.

Multiple choice tests are often stressful when made correctly, wherein the correct answer is mixed in with reasonable but incorrect options. It can be difficult to eliminate all choices and the situation genuinely involves varying levels of confidence.

Literature Review

Overtime, many researchers have tried to overcome some of these multiple-choice tests shortcomings by using a variety of scoring methods. The two main scoring methods are number-right, used by traditional multiple-choice tests, and negative marking (Burton 2005). Researchers have long argued over the pros and cons of these two categories (Lord 1975; Rowley and Traub 1977). In number-right scoring, students earn points for every correct answer and zero points for wrong answers. In negative marking tests, students receive points for correct answers and negative points for wrong answers, thereby increasing the tests validity (Bradbard, Parker, and Stone 2004; Warwick, Bush, and Jennings 2010).

One research stream in pedagogy focuses on reflective practice (Abrahams and Singh 2010; Isaias, Issa, and Pena 2014; Lavy and Yadin 2010; Stahl 2011). The idea is that students who reflect on their learning will learn better, make connections, and make their knowledge more permanent; for example, journaling is a reflective practice (Jefferson, Martin, and Owens 2014; Muncy 2014). Research has shown that reflective assessment can help students focus on goals of knowledge building (Yang et al. 2016). Unfortunately, reflection is far less prevalent during assessment of learning (Lavy and Yadin 2010). Tests are far more summative than formative. This can be especially true with multiple-choice tests, which fail to provide a bridge towards further learning. The standard perspective is that if you want reflective testing one should move to essay questions rather than multiple-choice (Hickson and Reed 2011). And while there is some value in that, essay tests do not scale well to large classes.

In order to award partial credit and gain a better insight into what is going on in a students' mind when they are answering a question, a new form of multiple-choice was developed called confidence based testing. In confidence-based testing, a student not only answers a question but also indicates how confident they are. The student has a chance to earn partial credit and the professor gains insight into student learning.

One vehicle for partial credit grading is called hedging (Walker and Thompson 2001). Hedging allows students to answer a question twice, receiving 100% if they answer the question correctly twice, 50% if they answer the question correctly one of the two times, and 0% if neither of their answers were correct. As such, hedging tests allow teachers to better understand how well students are grasping concepts and allows students to earn partial credit.

But if two answers per question is good, perhaps four answers per question is better. A more granular form called the Apperson Form A 1699 was developed by the University of Sydney. It allows for 0%, 25%, 50%, 75%, or 100% credit per question, and is the focus of the current study.

Terminology

In the literature, the concept of combining correct scores with confidence levels to earn partial credit is referred to as wagering and hedging interchangeably. We find these terms to be counterintuitive. In colloquial terms, wagering refers to someone taking a risk. However, in the case of these exams, the bigger risk is going all in on a question. Hedging may be a better word than wagering, but still carries the connotation of risk taking. By contrast, we borrow the term diversifying from finance. The student who diversifies lowers their risk, but also lowers their potential return. The risky strategy is to go "all in" and not diversify, just as the risky strategy in the stock market is to hold only one stock.

Prior research has examined diversification in a more constrained range (50-100) which splits the points between two answer choices, allowing students to earn 0%, 50%, or 100% depending on how they answer. Our research uses a different form (described in more detail in the next section), which examines a more precise and confidence-based diversification, allowing for a larger range of categorization (25-50-75-100). Students could earn 0%, 25%, 50%, 75%, or 100% depending on how they answer.

Description of new Scantron form

In order to provide teachers with a deeper understanding of their student's confidence, the University of Sydney created Form A1699 for sale and use by Apperson (see Figure 1 in the appendix). Form A1699 is a diversify (25-50-75-100) form (Apperson.com). Form A1699 benefits the student by allowing them to get partial credit for recognizing the correct answer, and allows the teacher to see which questions are difficult or confusing. After scoring, the professor can create a clarifying lecture tailored to the needs of the class. The insight to the thinking process is the same goal of other confidence testing methods (Eser, Holbrook, and Colbert 2012; Peyton 2010; Swartz 2006). The Apperson A1699 was used in Team-Based Learning literature (Sibley and Ostafichuk, 2015) the only form found that facilitated diversification of answers. It was therefore selected for our research.

The form consists of four lines for each question, each with four choices: A, B, C, and D. This means that the student must choose an answer for each line worth 25% of the student's grade for the question. Points are only given for answer choices which are correct—this allows for partial credit.
By answering the question four times the student reveals their confidence in their answer. If they answer the same letter four times then they are 100% confident (AAAA). If they answer the same letter three times they are 75% confident (AAAB). If they answer the two letters twice then they are 50% confident (AABB). If they answer two letters twice and then two different letters they are 25% confident (BBAC), and if they answer all four different letters then they are 0% confident (ABCD), which is assumed as pure guessing. The mixture of scores and confidence levels resulted in eleven possible combinations. These combinations will be referred to as competence codes and are described below.

Competence Code
This is a combined measure of student confidence and test score. The competence code is derived from the Form A1699, which displays student responses to tests as the score they received on a test and the letters that they bubbled in for each question. We determined confidence by the amount **not** diversified. For instance, if all four responses for Question 1 were the same letter, we considered the confidence as Perfect (P) or 100%. Along these lines, the confidence component included Perfect (P), High (H), Medium (M), Low (L), and None (N). The correctness component was the score they received for the question: 100, 75, 50, 25, or 0. The combination of correctness and confidence resulted in eleven different possible competence codes: P100, P0, H75, H25, H0, M50, M0, L50, L25, L0, and N25. Table 1 (Appendix) displays these codes based on possible answer examples, given that the correct answer in each example is A.

2. HYPOTHESES AND THEORY DEVELOPMENT

It is important to test the validity of a new testing method, as well as its effects on student grades. Previous research has found that the grades of students who use traditional multiple-choice tests, answer-until-correct tests, or short answer tests, are not significantly different from each other (Persky and Pollack 2008). This shows that multiple-choice tests are a valid form of testing. Researchers have also found that partial credit examinations are also a valid and reliable form of assessment (Bradbard, Parker, and Stone 2004; Walker and Thompson 2001). This means that positive partial credit is a valid form of testing. While Form A1699 is a positive partial credit multiple-choice form, it has had no prior research.

Our task is to see whether a diversify (25-50-75-100) form, Form A1699, meets the same standard. In particular, we are concerned with the students who display 75% confidence on a question. As Walker and Thompson (2001) stated "From a pure grade-maximizing standpoint (ignoring risk preferences) one wonders whether it ever pays to hedge." Assuming that a student is able to narrow down their choices to two answers, and is 75% confident in one and 25% confident in the other, then their expected grade on a traditional multiple-choice test would be 100%, while it would be only 75% if the student went with their confidence level. Based on this, we think that this student should "go all in." Table 2 shown in the appendix summarizes this thought process.

We believe that significantly more students that are 75% confident will also score 75%, meaning that they have a competence code of H75. Therefore, by not "going all in" these students lose 25% of the points they would have received on a traditional multiple-choice test. Therefore, we hypothesize:

Hypothesis A-1: Students who are 75% confident on a question will be hurt by diversifying on an exam.

In addition to looking at the grading effects of diversifying on questions in which students were confident, we looked at the grading effects on questions in which students were **not** confident. Students who have low confidence (25%) are able to eliminate only one potential answer. They then face the task of distributing four votes into three choices. The inconsistency between the number of votes and the number of choices allows students to receive a score of 50%, 25%, or 0% on a question as discussed in the section "scoring with 25% confidence". Students have a competence code of L25 will be helped by having the ability to diversify, as it would provide them with 25% when they would have likely received 0% on a traditional multiple-choice test. Students who have a competence code of L50 will be hurt as they would have chosen that answer in a traditional multiple-choice test and would have received 100%. Table 3 shown in the appendix clarifies this thought process.

We believe that significantly more students who are 25% confident will have a competence code of L25. This will leave most students helped, as they would have received 0% on a traditional multiple-choice test. Therefore, we hypothesize:

Hypothesis A-2: Students who are 25% confident on a question will be helped by diversifying.

Note, we divided hypotheses into two groups A and B. Group A studies how differences in diversification behavior effect grade outcomes. Group B studies how differences in external factors (gender, test difficulty and semester beginning versus end) affect diversification behavior, Hypothesis B-4 is an exception because it examines whether frequent diversifiers will score lower on traditional multiple-choice tests. Group B hypotheses are discussed below.

The idea behind partial credit Scantron forms is that students will diversify based on their confidence in and knowledge of a subject. However, other factors could influence diversification behavior. One factor that could possibly influence it is gender. Research has started to call attention to the unconscious effects of diversity, including gender, on the validity of testing formats that are used in classrooms (Ghorpade and Lackritz 1998). Previous research is divided on the effects of gender on diversifying behavior. Several researchers have found that females appeared to be more risk seeking then males—that is females tend not to diversify as much. They go all in on their answers (Ben-Shakhar and Sinai 1991; Jack et al. 2009). However, this only accounted for 4.8% of the variance between men and women's diversifying behavior (Jack et al. 2009). Other recent research has found that there is no difference between the diversifying behavior of men and women when measuring difference by diversifying frequency or overall exam grades (Curtis et al. 2013; Klymkowsky et al. 2006). Although the research is divided, Ben-Shakhar and Jack's results showed that gender accounted for very little difference in diversification behavior, so we hypothesize:
Hypothesis B-1: There will be no difference between the diversifying behavior of men and women.

Harder tests and harder questions on tests tend to affect students' confidence levels. For instance, Koku and Qureshi (2004) found that overconfidence increases with the difficulty of exam questions. In addition, their results suggest that overconfidence tends to increase as student performance decreases on an exam. Koku and Qureshi also found that having students give reasons to their answer choices causes the overconfidence to decrease. Therefore, it is possible that causing students to think about how confident they are will reduce confidence. Hence, we hypothesize:
Hypothesis B-2: Students will diversify more on harder tests.

One fear of administrators is that as students become more familiar with a testing format they will start to find shortcuts or rules of thumb that cause the testing format to help the students earn a higher grade. Inconsistent with this fear of administrators, Bradbard, Parker, and Stone (2004) found that when using negative marking partial credit testing that the students' test taking behavior did not change over time. However, no one has looked into whether test-taking changes overtime when a partial credit correct response test format is used. Based on findings of Bradbard et. al., we hypothesize:
Hypothesis B-3: Students will not change their diversification behavior during the two halves of the semester.

A potential problem with using a diversification test format is that it could provide students with a false sense of confidence and a poor development of study habits as they are receiving partial credit for problems that on a traditional test they might have gotten completely wrong. Therefore, those who rely on a crutch from diversifying may score lower on a final exam in which they cannot diversify. While this has not been studied before, it is an important topic, as it could show that using a diversification testing form throughout the year could develop poor study habits in students and result in lower learning as evidenced by a lower score on a traditional multiple-choice final exam. Due to this, we hypothesize:
Hypothesis B-4: Students who diversified more on previous tests will score lower on a traditional multiple-choice exam.

3. METHODOLOGY

We conducted this experiment at a large Midwestern University with ten sections of two sophomore level business core courses. Six of the sections were in Information Analysis and Design and the other four sections were in Quantitative Business Statistics. We gathered data for this experiment from scores of all 17 of the quizzes taken by 874 students, their final grade in the course, and descriptive data such as their team membership, timing of the exams, and their gender.

The diversification forms were Form A1699 from Apperson. Datalink software from Apperson designed for diversification forms graded form A1699. The data from the forms included the student's identification number, the test number, team number, responses and scores. We used Microsoft Excel to collect and compile the data, and SPSS to analyze it. We coded every question on every quiz for every student using the competence code described in section 1.3. After this we tallied the number of

questions that fit each code for each student on each quiz.

Diversification behavior
We used this variable to indicate the extent to which each student diversified. We calculated it by averaging the product of two variables: number of times each competence code was used in a quiz for a student (Num_c) and the percentage that each student diversified in that quiz. We treated diversification as a converse of confidence. Therefore, when Confidence (C) varied from Perfect (100% or 1) to None (0% or 0), diversification, calculated as (1 – C), varied from 0 to 1. Table 4 (see appendix) provides an example calculation of diversification behavior.

$$Db(C) = \left(\frac{\sum_{C=0}^{1}((1-C)*Num_c)}{\sum_{c=0}^{1} Num_c}\right) \times 100 \quad \text{..... (1)}$$

$Db(C)$ is diversification behavior
C is confidence
Num_c is the number of times the competence code was used in the quiz

The remaining data needed in order to analyze our hypotheses included quiz score, final exam score, gender, and test date. We captured the test date and quiz score data in the Form A1699, collected the final exam scores from the gradebook and matched them with students via their identification numbers. We determined the gender of the students using pictures provided by the school's database. With this data, we analyzed each of the hypotheses and received the results as described below.

4. ANALYSIS AND RESULTS

Hypotheses A
Hypotheses A-1: Students who are 75% confident on a question will be hurt by diversifying.

We coded student's information and answers to each quiz as described in the methodology section. The data included 6800 quiz scores, with multiple questions in each quiz. We then focused on students that were 75% confident and therefore coded as H75, H25, and H0. Students coded as H75 who were hurt because they earned only 75% credit instead of 100% on a no diversification quiz. In comparison, students coded H25 were helped because they were 25% correct and would have score 0. H0 got the answer wrong, which meant that diversification had no impact on their score; they were neither helped nor hurt, and were therefore excluded from our analysis. This classification yielded 1082 items for analysis (Table 5 appendix).

We analyzed this data using a chi-square goodness of fit test (Table 6) with 745 items of H75 and 337 items of H25. The result supported Hypothesis A-1 ($p<.000$) suggesting that confident students (75%) are hurt by diversifying.

Hypothesis A-2: Students who are 25% confident on a question will be helped by diversifying.
Conversely, less confident students (L) who earned 50% credit (L50) were hurt as they earned only 50% credit instead of 100% credit, on a no diversification quiz. Students who earned 25% credit (i.e. L25) benefitted as they would have earned 0%. As in Hypothesis A-1, students who were coded as L0 were excluded. Table 7 (appendix) shows 204 items for this analysis.

We analyzed this data using a chi-square goodness of fit test with 89 items of L50 and 115 items of L25. The results (Table 8) did not support hypothesis A-2 ($p=.0687$), that less confident students are neither being hurt nor helped by using diversifying forms.

Assumptions
We evaluated two pre-test assumptions prior to the analysis of all Hypotheses B: equal variances and normally distributed data. To detect the equality of variance, we used Levene's test for each variable. A significant result ($\alpha<.05$) would lead to a rejection of equality of variance. Table 9 shows the results of our data for each Construct.

Next, we tested the assumption of normality using Kolmogorov-Smirnov test. The result was significant ($p<.05$) requiring us to reject the assumption of normality (Table 10 - appendix). All variables failed both assumptions, except gender, which only failed the normality test. Therefore, we used the Mann-Whitney, to test mean differences for all hypotheses B.

Hypotheses B
Hypothesis B-1: There will be no difference between the diversifying behavior of men and women.

To examine diversification behavior based on gender, the data sample of 6800 samples was prepared as follows. First, we removed 177 samples due to subjects using the wrong identification number. As the results show (Table 11), there is no significant difference between the diversification behavior of men and women ($p=.102$).

Hypothesis B-2: Students will diversify more on harder tests.

Koku and Qureshi (2004) found that overconfidence increases with the difficulty of the question. Based on this, we categorized difficulty of tests based on the number of P0's students had received. In our study, harder tests had more P0s. We calculated the average number of P0 competence codes for each quiz across all students. Then using this score, we divided the quizzes into two groups based on the median. We removed 509 samples due to incorrect identifiers appearing on both levels of difficult tests and used 6291 data points.

As the results show (Table 12 - appendix), there is a significant difference between the diversification behavior of students on easier and more difficult tests (p=.000). Tests that are less difficult (coded 1) have a significantly lower amount of diversification behavior displayed by students.

Hypothesis B-3: Students will not change their diversifying behavior during the two halves of the semester.

We used 6289 data points after removing records with non-matching identification numbers.

As the results of the Mann-Whitney test (Table 13 - appendix) show a significant difference between the diversification behavior of students between the two halves of the semester (p<.05). Tests that are in the second half of the semester have a lower amount of diversification behavior displayed by students (p=.000).

Hypothesis B-4: Students who diversified more on previous tests will score lower on a traditional multiple-choice exam.

Using the measures of diversification behavior per student per quiz calculated earlier, we divided a sample into two groups along the median. The sample size included 6641 scores after removing items that did not include identification numbers. As before, we used Mann-Whitney to examine it (Table 14 - appendix).

The results show that there is a significant difference (p=.000) between the final exam scores of students with high and low average diversification behavior. Students with high average diversification behavior (coded 1) have a lower score on non-diversifying traditional multiple-choice tests.

5. DISCUSSION

In the last century, multiple-choice forms have gained popularity as a way to allow teachers to grade a large number of questions and/or a large number of students efficiently. However, traditional multiple-choice tests have deficiencies. These deficiencies include allowing students to gain points from guessing, measuring lower level thinking unless well designed, and do not allow teachers to understand their students' thought processes or confidence for each question. Overtime, people have developed various alternatives to traditional multiple-choice tests overcome these shortcomings. However, many of these forms have not undergone research or testing to determine if they are helping or hurting students and teachers. The Apperson Form A1699 allows students to receive partial credit in increments of 25% each time they chose the correct answer. It not only helps lower guessing as students can divide their answers, but also helps teachers understand what questions or topics students find confusing. Questions that have had more diversification can be reviewed and clarified by the teacher during class time, enabling students to more accurately communicate which topics they understand the least. While this idea sounds good in theory, previous research with other multiple-choice forms has shown that issues may affect this type of testing.

While several studies (Bradbard, Parker, and Stone 2004; Persky and Pollack 2008; Walker and Thompson 2001) have found that multiple-choice and partial credit exams are valid and reliable testing methods when compared to traditional testing methods, we wanted to ensure that the Form A1699 meets the same standard. One unique aspect of this form is that it shows students' confidence between four different, levels of confidence. This is imperative as Walker and Thompson (2001) wondered from a purely grade maximizing standpoint whether it paid to diversify. This question is especially important at the 75% confidence level (AAAB if answer is A) as students may become risk averse and underestimate their confidence (Table 2).

Due to the belief of Walker and Thompson that from a grade maximizing stand point diversifying does not benefit the student, we predicted that more students would be hurt than helped by being able to diversify. The data supports this presumed logic. Students are not able to tell the difference between being 75% confident and 100% confident and as such, are missing points they would have received on a traditional multiple-choice test, thereby negatively affecting their grade. This supports Walker and Thompson's theory that it might not pay to diversify.

One of our goals was to examine if it paid to diversify at a 25% confidence. Students who did not choose the right answer when distributing

the fourth vote are helped by being able to diversify as they received a score of 25% (bottom row in Table 3), even though on a traditional multiple-choice they would have received 0%. Students who score 50% are hurt by being able to diversify (top row in Table 3), as they would have chosen the right answer on a traditional multiple-choice test and received 100%.

We predicted that there would be significantly more people helped than hurt, as someone who was more confident in one answer would have eliminated more distractors. However, the data revealed that people were neither helped nor hurt by being able to diversify when they were 25% confident. We believe these students have little knowledge of the correct answer, as evidenced by being able to eliminate only one option out of four. The extra choice is more often randomly placed than placed out of knowledge of content. This means that when students are 25% confident (AABC) and they have to place a fourth vote on one of the three already selected multiple-choice alternatives (the double A), as they do with Form A1699, they guess. This is a problem because while this diversifying form does not affect the grades of students who have a low level of confidence, it does provide false data about what the students understand. This false data causes problems for a teacher who would use this information to devise their lesson plans.

In addition to finding that students are hurt by diversifying forms when they are 75% confident, we tested external factors to examine what else could be influencing the scores. We examined four external factors: (i) gender, (ii) difficulty of the test, (iii) progression through the semester, and (iv) if students who diversify more throughout the semester score lower on final exams.
In regard to the first external factor, our results indicated that gender does not influence diversification behavior. Even though some early studies (Ben-Shakhar and Sinai 1991; Jack et al. 2009) suggested opposing results in spite of their low effect size (only 4.8% of the variance being accounted for), recent studies (Curtis et al. 2013; Klymkowsky et al. 2006) support our finding. One possible explanation for the discrepancies between findings could be differences in culture. Jack, Liu, Chiu & Shymansky's study (2009) and Ben-Shakhar & Sinai's study (1991) were conducted outside the United States, while Curtis, Lind, Boscardin, & Dellinges's study (2013), Klymkosky, Taylor, Spindler, & Garvin-Doxas's study (2006), and our study were conducted in the United States.

When determining whether difficulty of a test affects the diversification behavior of students we found that it does increase the diversification behavior. The combined logic of the studies of Koku & Qureshi (2004) and Curtis et al, (2013) support this result. Koku & Qureshi found that as difficulty increases so does overconfidence; however, having students stop and think about the reasons behind their confidence causes this overconfidence to decrease. Curtis et al, discovered that students find complex questions more difficult than factual questions and as such tend to be more confident on factual questions. These findings show that when students are forced to think about their confidence and the reasons behind their answers (which tends to happen with complex questions) they tend to reduce their confidence. Similarly, we found that students have higher diversification behavior on difficult exams; possibly, due to the fact that forms with diversification force students to think about their confidence in their answers.

In looking where in the progression of the semester the test is given, our research shows that students diversify less as the semester progresses. However, previous research has shown that student's test taking behavior does not change over time (Bradbard, Parker, and Stone 2004). One possible reason for this discrepancy in results may be due to the fact that students begin to realize that when they are 75% confident they are being hurt by diversifying and as such they should "go all in". Future studies should further examine this discrepancy.

Lastly, we compared the final exam grades of students who had high and low diversification behavior to determine if students were using the diversification form as a crutch and not preparing for the final exam throughout the year. Our findings conclude that this is in fact happening. Students who had high diversification behavior score lower on final exams in which they cannot diversify. This shows that there may be a danger to student learning in allowing them to use partial credit multiple-choice forms.

One of the limitations of this research is that this research was conducted with quantitative courses that required students to think objectively. Therefore, its generalizability to more subjective and qualitative courses is untested.

Future recommendations

Multiple-choice forms have the ability to provide efficient feedback to students and partial credit multiple-choice forms have allowed teachers to maintain. this efficiency particularly in large classes. Our results indicate that any benefits of partial credit multiple-choice testing are out-

weighed by the negative consequences. For instance, the more confident students are hurt by permitting them to diversify (Hypothesis A-1). These are typically students who believe that they know their content, which is an expected outcome of working harder. It does not help the less confident students (A-2) Further, while students diversify more on harder tests, those who diversify more tend to score lower on non-diversifying exams, than their complements (Hypothesis B-4).

Because of this, we believe that partial credit multiple-choice testing should be avoided. Instructors should just use better questions in the traditional multiple-choice format. Unfortunately, most test banks provided by publishers tend to focus on regurgitation of facts. Therefore, the onus is on instructors to write their own tests with higher level questions. For example, case based questions are an excellent way to test application of knowledge.

6. BIBLIOGRAPHY

Abrahams, Alan S., and Tirna Singh. (2010). "An active, reflective learning cycle for e-commerce classes: Learning about e-commerce by doing and teaching." *Journal of Information Systems Education* 21 (4):383.

Ben-Shakhar, Gershon, and Yakov Sinai. (1991). "Gender Differences in Multiple-Choice Tests: The Role of Differential Guessing Tendencies." 23.

Bradbard, David A., Darrell F. Parker, and Gary L. Stone. (2004). "An Alternate Multiple-Choice Scoring Procedure in a Macroeconomics Course." *Decision Sciences Journal of Innovative Education* 2 (1):11-26.

Burton, Richard F. (2005). "Multiple-choice and true/false tests: myths and misapprehensions." *Assessment & Evaluation in Higher Education* 30 (1):65-72. doi: 10.1080/0260293042000243904.

Curtis, Donald A., Samuel L. Lind, Christy K. Boscardin, and Mark Dellinges. (2013). "Does student confidence on multiple-choice question assessments provide useful information?" *Medical Education* 47 (6):578-584. doi: 10.1111/medu.12147.

Eser, Zekeriya, Mary E. Holbrook, and Jan Colbert. (2012). "Confidence Based Marking: Implementation and Feedback Measures." *Journal of Higher Education Theory & Practice* 12 (1):27-38.

Ghorpade, Jai, and James R. Lackritz. (1998). "Equal Opportunity in the Classroom: Test Construction in a Diversity-Sensitive Environment." *Journal of Management Education* 22 (4):452-71.

Hickson, Stephen, and Bob Reed. (2011). "More evidence on the use of constructed-response questions in principles of economics classes." *International Review of Economics Education* 10 (2):28-49.

Isaias, Pedro, Tomayess Issa, and Nuno Pena. (2014). "Promoting Higher Order Thinking Skills via IPTEACES e-Learning Framework in the Learning of Information Systems Units." *Journal of Information Systems Education* 25 (1):45.

Jack, Brady Michael, Chia-Ju Liu, Hoan-Lin Chiu, and James A. Shymansky. (2009). Confidence Wagering during Mathematics and Science Testing. Online Submission.

Jefferson, Jonathan K., Ira H. Martin, and Jake Owens. 2014. "Leader development through reading and reflection." *Journal of Leadership Studies* 8 (2):67-75.

Klymkowsky, Michael W., Linda B. Taylor, Shana R. Spindler, and R. Kathy Garvin-Doxas. (2006). "Two-Dimensional, Implicit Confidence Tests as a Tool for Recognizing Student Misconceptions." *Journal of College Science Teaching* 36 (3-):44-48.

Koku, Paul Sergius, and Anique Ahmed Qureshi. (2004). "Overconfidence and the Performance of Business Students on Examinations." *Journal of Education for Business* 79 (4):217-224.

Lavy, Ilana, and Aharon Yadin. (2010). "Team-based peer review as a form of formative assessment-The case of a systems analysis and design workshop." *Journal of Information Systems Education* 21 (1):85.

Lord, Frederic M. (1975). "Formula Scoring and Number-Right Scoring." 7.

Muncy, James A. (2014). "Blogging for reflection: The use of online journals to engage students in reflective learning." *Marketing Education Review* 24 (2):101-114.

Persky, Adam M., and Gary M. Pollack. (2008). "Using Answer-Until-Correct Examinations to Provide Immediate Feedback to Students in a Pharmacokinetics Course." *American Journal of Pharmaceutical Education* 72 (4):1-7.

Peyton, Vicki. (2010). "Using multiple option multiple-choice exam formats in a secondary level science classroom." *The International Journal of Educational and Psychological Assessment* 6 (1):87-103.

Rowley, Glenn L., and R. E. Traub. (1977). "Formula scoring, number-right scoring, and test-taking strategy." *Journal of Educational Measurement* 14:15-22.

Schermerhorn, John R., William L. Gardner, and Sally A. Dresdow. (1992). "Success Profiles for Student Examination Performance in a Large-Lecture Management Course: An Empirical Study." *Journal of Management Education* 16 (4):430.

Sibley, Jim and Pete Ostafichuk (2015). "Getting Started with Team-Based Learning", Stylus Publishing, LLC.

Stahl, Bernd Carsten. (2011). "Teaching ethical reflexivity in information systems: How to equip students to deal with moral and ethical issues of emerging information and communication technologies." *Journal of Information Systems Education* 22 (3):253.

Swartz, Stephen M. (2006). "Acceptance and Accuracy of Multiple-choice, Confidence-Level, and Essay Question Formats for Graduate Students." *Journal of Education for Business* 81 (4):215-220.

Walker, Douglas M., and John S. Thompson. (2001). "A Note on Multiple-choice Exams, with Respect to Students' Risk Preference and Confidence." *Assessment & Evaluation in Higher Education* 26 (3):261-267. doi: 10.1080/02602930120052413.

Warwick, Jon, Martin Bush, and Sylvia Jennings. (2010). "Analysis and evaluation of liberal (free-choice) multiple-choice tests." *Innovation in Teaching and Learning in Information and Computer Sciences* 9 (2):1-12.

Yang, Yuqin, Jan van Aalst, Carol K. K. Chan, and Wen Tian. (2017). "Reflective assessment in knowledge building by students with low academic achievement." *International Journal of Computer-Supported Collaborative Learning* 11 (3):281-311.

7. APPENDIX

Figure 1. Apperson Form A1699

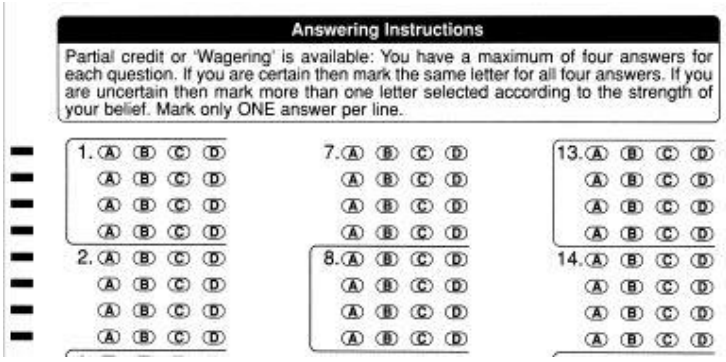

Table 1. Example of the Eleven Possible Competence Codes

Answer	Confidence Level	Scores	Competence Code
AAAA	P = 100%	100%	P100
BBBB	P = 100%	0%	P0
AAAB	H = 75%	75%	H75
BBBA	H = 75%	25%	H25
BBBC	H = 75%	0%	H0
AABB	M = 50%	50%	M50
BBCC	M = 50%	0%	M0
AABC	L = 25%	50%	L50
BBAC	L = 25%	25%	L25
BBCD	L = 25%	0%	L0
ABCD	N = 0%	25%	N25

Table 2. Example of Classification of a 75% Confident Student's Answers

Answer (Correct Answer "A")	Score	Confidence	Competence Code	Helped or Hurt
AAAC	75%	75%	H75	Hurt –Would have received 100% on traditional multiple-choice
BBBA	25%	75%	H25	Helped – Would have received 0% on traditional multiple-choice

Table 3. Example of Classification of a 25% Confident Student's Answers

Answer (Correct Answer "A")	Score	Confidence	Competence Code	Helped or Hurt
AABC	50%	25%	L50	Hurt –Would have received 100% on traditional multiple-choice
ABCC	25%	25%	L25	Helped – Would have received 0% on traditional multiple-choice

Table 4. Example of Calculation of Diversification Behavior for a student on an eight-question quiz

Competence Code	P100	P0	H75	H25	H0	M50	M0	L50	L25	L0	N25
Num_c	4	0	1	0	0	0	1	1	1	0	0
Confidence (C)	1	1	.75	.75	.75	.5	.5	.25	.25	.25	0
Diversn (1-C)	0	0	.25	.25	.25	.5	.5	.75	.75	.75	1
Numerator	4*0 = 0	0*0 = 0	1*.25 = .25	0*.25 = 0	0*.25 = 0	0*.5 = 0	1*.5 = .5	1*.75 = .75	1*.75 = .75	0*.75 = 0	0*1 = 0
Diversification Behavior	$28.125 = \left(\frac{(0+0+.25+0+0+0+.5+.75+.75+0+0)}{(4+0+1+0+0+0+1+1+1+0+0)}\right) * 100$										

Table 5. Example of the Three Possible 75% Confidence Classifications

Possible Answer (A is correct)	Confidence	Score	Code	Hurt or Helped
AAAB	75% = H	75%	H75	Hurt, should have not diversified would have received 100%
BBBA	75% = H	25%	H25	Helped, would have received 0% without diversifying
BBBC	75% = H	0%	H0	Neither, was incorrect and would have been incorrect without diversifying

Table 6. Chi-Square: 75% Confident Students

75% Confident	Expected	Observed	χ^{2}*
Hurt (H75)	541	745	.000
Helped (H25)	541	337	

Note. * = α = .05

Table 7. Example of the Three Possible 75% Confidence Classifications

Possible Answer (A is correct)	Confidence	Score	Code	Hurt or Helped
AABC	25% = L	50%	L50	Hurt, should have not diversified would have received 100%
BCCA	25% = L	25%	L25	Helped, would have received 0% without diversifying
BBCD	25% = L	0%	L0	Neither, was incorrect and would have been incorrect without diversifying

Table 8. Chi-Square: 25% Confident Students

25% Confident	Expected	Observed	χ^{2}*
Hurt (L50)	102	89	.069
Helped (L25)	102	115	

Note. * = α = .05

Table 9. Results of Levene's Test

Constructs	F	Sig.
Gender	1.841	.175
Test Difficulty	37.202	.000
Test Date	25.985	.000
Final Exam Scores	160.560	.000

Note. * = α = .05

Table 10. Results of Kolmogorov-Smirnov Test for Normality for each Construct

Kolmogorov-Smirnov	Groups	Statistic	df	Sig.*
Gender	Men	.327	6623	.000
	Women	.378	6623	.000
Test Difficulty	Difficult	.270	3124	.000
	Easy	.384	3167	.000
Test Date	Early	.285	3215	.000
	Late	.370	3074	.000
Diversification Behavior	Low	.328	6593	.000
	High	.075	6593	.000

Note. * = α = .05

Table 11. Mann-Whitney Results of Diversification Behavior Categorized by Gender

Div. Beh.	N	Mean Rank	Sum of Ranks	Mann-Whitney	Z	Sig.*
Female	2839	3351.54	9515015.50	5259140.500	-1.637	.102
Male	3784	3282.34	12420360.50			
Total	6623					

Note. * = α = .0

Table 12. Mann-Whitney Results of Diversification Behavior Categorized by Test Difficulty

Div. Beh.	N	Mean Rank	Sum of Ranks	Mann-Whitney	Z	Sig.*
Difficult	3124	3411.01	10655991.00	4118967.00	-12.934	.000
Easy	3167	2884.59	9135495.00			
Total	6291					

Note. * = α = .05

Table 13. Mann-Whitney Results of Diversification Behavior Categorized by Test Date

Div. Beh.	N	Mean Rank	Sum of Ranks	Mann-Whitney	Z	Sig.*
1st Half of Semester	3215	3360.82	10805043.50	4247586.50	-10.841	.000
2nd Half of Semester	3074	2919.28	8973861.50			
Total	6289					

Note. * = α = .05

Table 14. Mann-Whitney Results of Final Exam Categorized by Diversification Behavior

Final Exam	N	Mean Rank	Sum of Ranks	Mann-Whitney	Z	Sig.*
Low Div. Beh.	3255	3378.47	10996923.00	4737747.000	-7193	.000
High Div. Beh.	3206	3081.28	9878568.00			
Total	6461					

Note. * = α = .05