2015-05-12

Student-weighted Multiple Choice Tests

Ling, Joseph

Student-weighted multiple-choice tests

Joseph Ling (jling@ucalgary.ca)
Michael Cavers (mcavers@ucalgary.ca)
Department of Mathematics and Statistics
University of Calgary
Calgary, AB T2N 1N4 Canada

Abstract. Large-enrolment courses routinely administer multiple choice tests. Well-thought-out multiple choice tests are excellent assessment instruments, but their format allows for guessing. Furthermore, the opportunity for students to highlight their levels of mastery of material is not necessarily present. Frary (1989) reviews multiple choice methods that attempt to capture student knowledge in each question. We discuss one such format of multiple choice testing known as Confidence Weighting (Echternacht, 1972). In this model, students are given a certain freedom to assign relative weights to individual questions representing their belief in the correctness of their response. The purpose of this is to allow students to demonstrate their level of confidence on their true knowledge of the material. Thus, students with identical responses to the questions may receive different overall scores depending on their indicated degree of confidence for each question. We experimented using this method in two first year Calculus courses during three semesters in 2014. We describe our experience along with issues encountered while using this model.

Introduction.

Multiple-choice tests involving one choice out of four (or five) options is extensively used in the testing of mathematics in undergraduate studies. Advantages include being adaptable to measure various learning outcomes, the ability to cover a broad range of topics, efficient and reliable scoring tools, and flexibility in choosing distractors that may provide feedback on student misconceptions. Disadvantages are that writing good questions and distractors is time consuming, the questions tend to test students on factual recall rather than higher level skills, they don't allow for testing of mathematical ideas, and performance on such tests could also rely on student characteristics not related to the academic content (e.g., reading skills, arithmetic skills, deductive reasoning). Many resources are available in constructing good multiple-choice questions, for example, see Shank (2010).

Here we discuss a non-conventional multiple-choice model: the student-weighted model. In the student-weighted model, a confidence weighting procedure, the examinee indicates an answer choice and their certainty of its correctness. Echternacht (1972) summarizes relevant studies to confidence weighting and states that confidence testing can be used to discourage guessing as the expected score is maximized only if the examinee reveals the true degree of certainty in responding. We first highlight some previous studies on confidence testing and then discuss the study we undertook at the University of Calgary where we used the student-weighted model in two of our first year calculus courses.

Literature Review.

In the 1930s, various studies using confidence testing methods were performed in an effort

to maximize the amount of information that could be obtained from multiple-choice tests.

Hevner (1932) applied confidence testing to music appreciation but used true/false tests. In these tests examinees chose one of two pieces of music as being more musical and indicated their degree of confidence in their answers on a three-point scale. Examinees were unaware of the scoring methods used in the study and four such methods were considered for differences in reliability. The scoring methods include the number of correct answers, the number correct minus the number incorrect, a weighted-correct answer score (correct answers count as three if the highest confidence was indicated, one if the lowest confidence was indicated and two otherwise), and a weighted-correct minus weighted-incorrect score. Using the Spearman-Brown formula, Hevner (1932) found that the weighted-correct scoring method resulted in the highest reliability.

A similar study with true/false tests was completed by Soderquist (1936), however, the subjects knew the scoring system to be used in advance unlike in Hevner's (1932) study. Wiley and Trimble (1936) extended the confidence testing procedures used at that time and analyzed whether or not there were personality factors in the confidence testing method. Gritten and Johnson (1941) used the method of confidence testing but used a five-point scale for confidence levels.

An interesting study we would like to highlight was completed by Dressel and Schmid (1953) who tested four non-conventional multiple-choice models. In their study, they used five sections of a course in physical science containing about 90 students each. During the first hour of the test, each of the five sections wrote a common conventional multiple-choice test. During the second hour, each section was given a different type of test:

- a *conventional* multiple-choice test with five options and one correct answer;
- a *free-choice* test where examinees marked as many answers as needed in order to be sure they had not omitted the correct answer (each question had exactly one correct answer);
- a *degree-of-certainty* test where examinees indicate how certain they are of their answer as being the correct one (each question had exactly one correct answer);
- a *multiple-answer* test where any number of the options may be correct and the examinee is to mark each correct alternative; and
- a *two-answer* test where exactly two of the five options is known to be correct.

Dressel and Schmid (1953) found that students completed the conventional test the fastest, followed by degree-of-certainty, free-choice, two-answer, and multiple-answer. From most to least reliable are the multiple-answer, two-answer, degree-of-certainty, conventional, free-choice.

Frary (1989) reviews multiple choice methods used mainly in the 1970s and 1980s that attempt to capture student knowledge in each question. Confidence weighting methods can be useful but one major obstacle is the existence of personality factors.

Contributions.

In this study we discuss the results of using the student-weighted scoring method in two first year calculus courses at the University of Calgary. The method was used by students on the multiple-choice section of the midterms and final examination. Each question had exactly one correct answer out five options. After each question, the student was asked to assign a relative weight of 1, 2, or 3 to the question. Students were told that they can adjust the weights of the questions based on their level of confidence for each question (i.e., use “Relative Weight = 3” when you feel very confident, and use “Relative Weight = 1” when you do not feel confident). A default weight of 2 was assigned when students left the weighting part of the question blank. To calculate the multiple-choice score for each student, the total sum of the relative weights for each correct answer was divided by the total sum of the relative weights assigned to all questions. During each course, a survey was conducted to solicit student feedback about the method as described below.

Surveys.

We used the student-weighted scoring method in three semesters in two of our first year calculus courses – Winter 2014 (one section of Math 249 and one section of Math 251), Spring 2014 (Math 249) and Fall 2014 (Five sections of Math 249). We conducted a survey in each case but at different points in the semester. We conducted the survey for the Winter 2014 and Spring 2014 semesters only after the semesters have ended, and in the former case, months after the semester had ended. On the other hand, we conducted the survey for the Fall 2014 semester about two weeks after the second midterm test. The main purpose of the survey was for us to learn about the students' perspectives on their experience with our scoring method in order to help us improve both our teaching and future students' learning experience.

We asked the same set of questions in each of the three surveys. There are six questions in total. In questions 1 to 4, the students were asked to tell us how much they agreed or disagreed with a given statement by choosing one of five options: Strong agree, Agree, Neutral, Disagree or Strongly disagree. Below are questions 1 to 4.

Question 1. Assigning relative weights to multiple choice questions was confusing.

Question 2. Assigning relative weights to multiple choice questions was beneficial to my learning.

Question 3. Assigning relative weights to multiple choice questions was beneficial to my grade.

Question 4. I prefer the traditional multiple choice model over the one used in this course where we were to assign relative weights.

Question 5 also involved five choices, but it was focused on students' feelings about the relative emphasis placed on multiple-choice questions and written-answer questions.

Question 5. Which of the following compositions of written questions (W) and multiple choice questions (MC) would you feel most comfortable with? 100% W and 0% MC, 70% W and 30% MC, 50% W and 50% MC, 30% W and 70% MC, 0% W and 100% MC.

Question 6 gave students an opportunity to freely express their opinions.

Question 6. Feel free to write down any comments.

The response rates were low. In Winter 2014, 23 out of 197 students in Math 249 responded and 8 out of 129 students in Math 251 did. In Spring 2014, 2 out of 72 students in Math 249 responded. In Fall 2014, 73 out of 646 in Math 249 responded. Because of the low response rates, we hesitate to draw any definite conclusions about the general student perceptions on the scoring method. We noticed a few recurring comments, a few conflicting comments and comments that do not seem to address our express concerns. The last group of comments is a sign that we might not have phrased our survey questions in the best possible way. It also suggests that we should exercise special cautions when interpreting students' answers in general. The survey questions will need to be completely redesigned for the purpose of a more rigorous scientific study in the future.

Effects on Course Performance.

Questions 2 and 3 in the Survey measure students' perception on learning and course performance (measured by the individual students' grades and the class average marks). With

regard to course performance, we have actual data from tests and examinations that allow us to compare the students' marks based on the conventional (uniform-weight) scoring method and the student-weighted scoring method. In Winter 2014, the effect on the class average on (the multiple-section of) midterm test 1, midterm test, and the (completely multiple-choice) final examination is as follows:

EFFECT ON CLASS AVERAGE

	Conventional MC	Weighted MC	Effect
Midterm test 1	63.15%	68.50%	+5.35%
Midterm test 2	65.82%	71.38%	+5.56%
Final Examination	62.48%	69.15%	+6.67%

Students whose mark is affected by 5 percentage points or more would more likely than not receive a different letter grade than if the conventional scoring method were used. Also, a difference of 3 percentage points in some cases would also affect the student's letter grade. The following tables show the percentage of students in this category. We have separated the "beneficial" cases from the "detrimental" cases.

NUMBER (AND PERCENTAGE) OF "BENEFICIAL" CASES

	Mark on MC increased by 5 percentage points or more	Mark on MC increased by 3 percentage points or more
Midterm Test 1	269/723 = 51%	478/723 = 66%
Midterm Test 2	328/657 = 50%	454/657 = 69%
Final Examination	384/603 = 64%	450/603 = 75%

NUMBER (AND PERCENTAGE) OF "DETRIMENTAL" CASES

	Mark on MC increased by 5 percentage points or more	Mark on MC increased by 3 percentage points or more
Midterm Test 1	9/723 = 1.2%	25/723 = 3.5%
Midterm Test 2	11/657 = 1.7%	25/657 = 3.8%
Final Examination	2/603 = 0.3%	6/603 = 1.0%

These results indicate that course performance as a whole is better when we use the student-weighted scoring method than if we had used the conventional scoring method where all multiple-choice questions are weighted the same. However, how to interpret this fact in terms of student learning is a much more complex task with many challenging issues of contention.

Issues for Future Investigation.

What at first motivated the second author to devise the student-weighted scoring method were his concern to minimize the impacts of 1) students' careless mistakes by providing incentive to students to check their work more carefully, and 2) students' guessing the correct answer by very gently encouraging them to lower the weight of work done by guessing. When we implemented this scoring method and started to reflect on our experiences and to gather feedback from students, colleagues, and participants in our workshops, questions began to surface. Suspected pros and cons

as well as a few issues of contention kept being brought up. During the workshop, participants had an opportunity to discuss four questions we considered controversial. The first question concerns students' learning experience. If you were a student writing a student-weighted multiple-choice test, would you feel more worried or less? The second question concerns how well students' course performance reflect their knowledge of the subject. In every semester we used the student-weighted scoring method, students' test marks and course grades were positively affected. So, has our method made the test results look better than they were? The third question concerns what it is that we should consider more important to be warranted to be used both as part of the teaching and learning experience and as the measure of course performance. What is more important: Student's confidence in their knowledge or the percentage of material they understand? The fourth question concerns the possibility of students gaming the system. Could a clever students get a passing mark simply by being confident about which topics they understand?

These are only a few among a long and growing list of issues that are worthy of or even demand careful investigation. Here are some issues on our list. How do students perceive the scoring method in relation to their learning experience and course performance? How do these perceptions correlate with actual course performance? How do students think their knowledge and their confidence in their knowledge are related? To what degree does student perception change as the semester progresses? Do and how do students take steps to build up their confidence in their knowledge? To what degree are students' confidence and course performance correlated? Are there, and what are the long term impact on students' learning and performance in subsequent courses? The list goes on. We do not have answers to many of these questions, which have evoked different and sometimes opposing opinions from our peer teachers and educators. It is our hope that our workshop has raised the awareness of general possibilities and specific attempts to modify common assessment practices to enhance learning experience and to improve the quality of our assessment tools. We hope that more teachers and educators will be inspired by our workshop to explore this topic in depth.

Acknowledgements.

We would like express thanks to those who participated in our workshop at the "Design for Learning: Fostering Deep Learning, Engagement and Critical Thinking" conference.

References.

Dressel, P. L., & Schmid, J. (1953). Some modifications of the multiple-choice item. *Educational and Psychological Measurement*, *13*, 574-595.

Echternacht, G. J. (1972). The use of confidence testing in objective tests. *Review of Educational Research*, *42(2)*, 217-236.

Frary, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education*, *2(1)*, 79-96.

Gritten, F., & Johnson, D. M. (1941). Individual differences in judging multiple-choice questions. *Journal of Educational Psychology*, *32*, 423-430.

Hevner, K. A. (1932). A method of correcting for guessing in true-false tests and empirical evidence in support of it. *Journal of Social Psychology*, *3*, 359-362.

Shank, P. (2010). *Create Better Multiple-Choice Questions*. The American Society for Training &

Development, Vol. 27, Issue 1009.

Soderquist, H. O. (1936). A new method of weighting scores in a true-false test. *Journal of Educational Research*, 30, 290-292.

Wiley, L. N., & Trimble, O. C. (1936). The ordinary objective test as a possible criterion of certain personality traits, *School and Society*, *43*, 446-448.