European Journal of Science and Mathematics Education
Vol. 4, No. 4, 2016, 501-505

Introducing new material for a standardized exam: a controlled, prospective, double-blinded test of student learning

Kopl Halperin[1] and William S. Dunbar[2,*]
[1] Science, Montgomery County Public Schools, Rockville, Maryland, USA
[2,] Department of Mathematics, Bard College at Simon's Rock, Great Barrington, Massachusetts, USA
For correspondence: kenhalperin@yahoo.com

Abstract
Do multiple choice unit tests reflect what students have learned during the unit? The day before the administration of a county-mandated multiple choice test, two classes were shown a topic they had not previously seen, and told it would be on the test. One class was shown the same material and told it was not important, and two classes were not shown the material. There was one question on the test concerning this material. The first two classes (n=54) had 83% correct on the question. The next class (n=32) had 81% correct. The last two classes (n=43), who did not see the material, had 63% correct on the question. The results were statistically significant, p=0.045. There was no difference between the class averages on the test overall. The results appear to indicate that one can prepare for multiple choice tests independently of other aspects of learning.

Keywords: Test preparation, standardized testing, prospective, controlled, double-blind

Introduction

Can students learn and retain new material on the day before a standardized, county-wide multiple-choice unit exam? If so, does this say anything about standardized testing, or learning more generally?

The school district of the study is a large suburban school district: 13000 teachers, 152,000 students. The students are 32% white, 27% Hispanic, 21% black, 14% Asian. 35% are on free or reduced meals. In a typical year there are 10,800 8th grade students, all of whom take the same science course: Introduction to Earth Space Systems. The county has standardized, multiple-choice unit exams, which drive the content of the teaching. Teachers have access to the unit exams before they are given. There are also semester exams that cover the same material; these are not seen in advance by the teachers. Within this system, teachers have reasonable freedom to present the material, bearing in mind the topics that will be on the exams. The time pressure to cover all the topics is high. During the school year studied there were 10 snow days, which made covering the material very difficult.

The studied school is one of the 38 middle schools in the district. The student population is highly diverse in ethnic background. The overwhelming majority of the students come from ethnic groups that are considered white or Asian; these two groups generally perform well on standardized tests (NCES, 2011). Fewer than 5% of the students are on free and reduced meals. There are 2.5 FTE eighth grade science teachers, who make every effort to cover the same material during each unit. One of the teachers, not the one in this study, has all of the IEP (individualized educational plan) inclusion classes. Otherwise the 8th grade students in the school year were randomly distributed over the 12 sections. There are no gifted and talented classes. Thus the 131 students in this study are a randomized selection of all the eighth grade students at the school, other than those with IEPs that include the need for a supported classroom. There are interactive whiteboards in every classroom.

These enable the presentation of complicated pre-made lessons, which can then be saved and posted online in the school's online learning community.

The third unit exam of the year had 27 multiple-choice questions, and one short essay question worth four points. There was one topic, tested by one multiple-choice question, that had not been covered in the course material. It was decided not to count that question towards the unit exam score. This question concerned construction for earthquake damage minimization. The exam was worth roughly 15% of the final grade for the quarter.

Exam Review

On the day before the unit exam, there was an exam review. Typically, students are given a number of exam review sheets in the few days before the exam. One of them is made by the county itself; it lists topics the students need to know. Teachers are free to make and distribute further exam review sheets. Three other exam review sheets were made available. Extremely diligent students, perhaps 20% of the student body, actually review these before the exam review day. On the exam review day,one of the reviews was gone over. On this particular day, the exam review used was straightforward. The exam questions themselves were not presented. What was presented was the material that was going to be on the exam. The students had a review sheet. This sheet matched a pre-made exam review whiteboard presentation.

This pre-made exam review had on it a question about the topic under study here: earthquake damage minimization, and this question was similarly covered on the whiteboard presentation. The material on earthquake damage minimization was placed in the middle of the review. The test question itself was number 22 of the 27 questions. The whiteboard presentation simply stated that earthquake damage is minimized if a building is built on solid bedrock; a diagram accompanied the words. Five sections were included in the study: 2nd period (22 students), 4th (32), 5th (32), 6th (17) and 7th (26). In the first two classes (n=54) the review treated the earthquake damage question the same way it treated all other questions. In the 5th period class (n=32), the question was reviewed during the review, but the students were told it would not be on the test. In the last two classes (n=43), the earthquake damage material was not covered: the review skipped over it. The whiteboard presentation was posted on the internet the evening before the test, and this question was answered in the internet version for those who chose to review it.

The question on design for earthquakes was not counted for the student grade. Two classes reviewed the material and were told it was important. One class reviewed it and was told it was not important. Two classes did not review it. No matter what they were told, it was not counted towards the grade. No one was told whether the question, or any other question, would appear on the exam. Nor was anyone told that there would be a question that was not counted.

Exam Results

The 54 students in periods 1 and 4 were told that this material was as important as any other in the review. Of the 54, 45 got the question correct and 9 got it wrong; 83% were right. The 32 students in period 5 were shown the correct answer but told it was unimportant. Of the 32, 26 (81%) had the correct answer. The 43 students in periods 6 and 7 were not shown the material; the review skipped over it. Of the 43, 27 had the right answer and 16 did not; 63% were right. On control is the randomization of students into the sections. Another control is that the averages on the 26 multiple-choice questions that did count for the score were: 1st and 4th periods, average 22.07 (85%); 5th period average 20.66 (79%); 6th and 7th periods, average 21.37 (81%). The confidence intervals of the overall averages overlapped. Various statistical tests failed to find a statistically significant difference

between the overall averages on the test. Kruskal-Wallis rank sum test gave p=0.174; ANOVA gave p=0.127.

We performed a chi-squared test for the three groups' scores on the earthquake minimization question: Pearson's chi-squared test, df=2, chi-squared = 6.184, p = 0.0454. For completeness, we performed a chi-squared test on the results for 1st and 4th periods on the one question, vs. the results for 6th and 7th periods. Total n = 97. Fisher's exact test gives p=0.033. Chi-square with Yates correction gives 4.26 with 1 degree of freedom, p=0.039. The results are clearly statistically significant: 1st and 4th periods did in fact "learn" the correct answer to the question.

Retention

After three weeks had elapsed, the students were asked a question on the topic of earthquake damage minimization to see if they had retained the information. The question used was from the New York State Regents exam. It was modified to have different distractors.

The results for retention followed the same trend as the results for the exam. There were a few absent students on the day of the retention question. 1st and 4th periods showed 52 students getting the correct answer and only 1 being incorrect: 98% correct. 6th and 7th showed 34 correct and 7 incorrect: 83% correct. Clearly this is not exactly retention, as more people were correct three weeks after the exam than during it. Nevertheless, the Fisher's exact test gives p=0.0097, the Chi-squared with Yates correction gives p=0.0199. 5th period had 26 correct answers and 5 incorrect, 84% correct.

Analysis

The results are statistically significant. Classes of students who were shown this material during exam review did better on the related exam question than students who were not shown that information. The question was multiple-choice, one correct answer and three distractors. The information was given on the day before the test. Furthermore, the classes that were shown this information scored better on a single-question test of their retention of the information, given three weeks after the exam. This question was multiple-choice, one correct answer and two distractors. The simple analysis that follows is that students can be shown information and tested on it immediately thereafter and will do well on the test. Furthermore, students will retain that information for use on a later test.

That 63% of the students who had no exposure to the material chose the right answer is not surprising. The question could be answered fairly well using common sense or prior knowledge, and students who reviewed the whiteboard review overnight saw the material. However, students who were exposed to the material in class did significantly better on the question, which means that common sense alone is not as good as common sense and a short exposure to the concepts.

Any further analysis is speculative, and will require further research. Nevertheless, it is, perhaps, worth speculating on the significance of this work. First, the students in all classes did better on the retention question, three weeks after the exam and its review, than they did on the exam question on the same topic. There are a number of possible reasons for this that come to mind. One, the retention question was a single question. An exam worth 30 points, twenty-seven multiple-choice questions and a short essay, is a considerably different situation than a single question worth 1 point. Also, the retention question allowed for a greater possibility of cheating than the exam did. It appears from the answer sheets that a few people changed their answer to it when they saw what others had answered. Finally, the retention question might simply have been worded in a way that worked better for these students.

Does this research say anything about learning itself? The other 26 questions on the exam covered topics that had been taught with a variety of techniques over the course of two months. Yet students in periods 1 and 4 did just as well on this question, which was covered only by a single slide on a review whiteboard presentation, as they did on most of the remaining questions. That is, 83% of the students in those two classes got the correct answer on this question, and 85% was the average score on the remaining 26 multiple-choice questions. If the goal is to do well on a multiple-choice exam, then this method of learning (going over the material the day before the exam) might work as well as any.

Suggestions for Further Research

These results appear to raise questions about the entire raison d'etre of standardized, multiple-choice testing. Students generally like multiple-choice tests (Sommer and Sommer, 2009) even if that does not mean they do better on them than on other tests (Tasdemire, 2010). It is shown here that students can perform well on a single question on a standardized multiple-choice exam after having been exposed to the material covered by that question for less than five minutes. Further research on the applicability of this result to other populations and subject areas would be useful. These study groups were of sufficient size for significance, but it would be great to attempt to replicate the results over larger populations. The same methodology could be applied to exam questions in social studies, language, math, and in any country or culture.

It would also be useful to apply this methodology to an entire exam. Prepare an exam of 20+ questions on topics that have not been covered at all. Expose half of the students to those topics on the day before the test. Leave the other half of the students unexposed to that material. It is indicated by this research that the students exposed to the topics just for one period will outperform the control group, in a statistically significant way. These results are consistent with the findings of Turner and Williams (2007) that multiple-choice test scores can be improved simply by improving students' vocabulary. That is, "earthquake damage minimization" could be considered to be a vocabulary phrase.

This research indicates that a standardized, multiple-choice exam does not measure what students have learned during the unit, the material covered during the previous six or eight weeks. Such a test measures, instead, the ability of students to retain the answers to what they think (or have been told) will be on the exam. This accords well with the results found by Haynie (2003) that students focus on what they think will be on the test. This material, in this research, was told to them on the day before the exam. They then retain those answers, at least for three weeks.

Is this the kind of "learning" that society wishes to promote? To our minds, the real goal of education is measured by the success of the students in their lives, a goal which cannot be measured until many years have elapsed. It is quite possible that standardized, multiple-choice exams are highly correlated to success in the job market (Tanner, 2003). If so, this research opens up the possibility of promoting the success of many students, by recognizing that this success can be tweaked more by good test preparation on the day before the test than by some other aspects of teaching. It is also possible that success in the job market relies on an ability to think through problems. If so, this research indicates that standardized, multiple-choice tests are not useful in promoting success. This is a position also taken by Chesbro (Chesbro, 2010). Further research could move in many directions.

We recognize that these are pretty large questions for a paper concerning itself with the answers given by students on a single question on an exam. Nevertheless, the results are too strongly significant to disregard.

References

Chesbro, Robert (2010). Strategies for the Meaningful Evaluation of Multiple-Choice Assessments. *Science Scope*, *34(2)* , 12-15.

Haynie, W. J., III (2003). Effects of Take-Home Tests and Study Questions on Retention Learning in Technology Education. *Journal of Technology Education, 14(2),* 6-18.

National Center for Education Statistics: The Nation's Report Card: Science 2009 (2011). *National Assessment of Educational Progress at Grades 4, 8, and 12.* NCES 2011, 451.

Sommer, Robert; and Barbara A. Sommer (2009). The Dreaded Essay Exam. *Teaching of Psychology, 36(3),* 197-199.

Tanner, David E. (2003). Multiple-Choice Items: Pariah, Panacea or Neither of the Above? *American Secondary Education, 31(2),* 27-36.

Tasdemir, Mehmet (2010). A Comparison of Multiple-Choice Tests and True-False Tests Used in Evaluating Student Progress. *Journal of Instructional Psychology, 37(3)*, 258-266.

Turner, Haley and Robert L. Williams (2007). Vocabulary Development and Performance on Multiple-Choice Exams in Large Entry-level Courses. *Journal of College Reading and Learning, 37(2)*, 64-81.