ASSESSING STUDENT LEARNING IN FIRST-YEAR QUANTITATIVE COURSES AT BABSON COLLEGE: IMPLEMENTATION AND ANALYSIS

MCKENZIE, Jr. John D. and RYBOLT , William H.
Babson College
USA

This paper reports on the implementation of the experiments we designed to assess how well our first-year students learn statistics and mathematics with electronic quizzes. It describes many issues that arose in our attempts to study the impact of such quizzes. An explanation of these issues should assist others who wish to assess the impact of technology in the classroom. The paper concludes with a preliminary analysis of the data from our experiments in fall semester of 2006 and a description of similar experiments planned for the spring semester of 2007.

INTRODUCTION

Assessing different teaching techniques based upon electronic quizzes seems straight forward and easy to implement. Alas such is not the case as we report on our experience in this paper. This paper is divided into three parts: issues encountered when we conducted experiments in the fall semester of 2006 that were described in the preceding paper, results of a preliminary data analysis from these experiments, and a description of similar experiments planned for the spring semester of 2007.

Some of the issues that we encountered included: quizzes that contain errors, changes in the software generating results than did not satisfy the tolerance built into the quiz, laptops not working, and students needing to leave class early, missing class and wanting to make up a quiz. In a normal academic setting, these issues are not that difficult to handle. But when these occur as part of an experiment, one has to be careful in handling the data so that these extraneous factors do not generate false conclusions. The entire point of this effort is to refine protocols for using electronic quizzes to assess other educational issues. It is our hope that others who plan to conduct similar statistics education experiments will benefit from knowledge of the issues we experienced.

In the following section we discuss some of the issues encountered when we actually implemented the experiment designs from the preceding paper. These issues are separated into five groups: questions, software, hardware, students, and instructors. While no one issue was a source of major difficulty, the collective impact of these numerous issues, was to confound the data from the experiment.

QUIZ QUESTION ISSUES

Our electronic quizzes were implemented in an EDU environment. The EDU quiz questions themselves are sometimes a source of difficulty. The amount of time typically allocated to take a quiz was 20 to 30 minutes for a quiz that should require 10 minutes or less. When a student did not complete the quiz in that time period, he or she occasionally could not get a grade on his or her completed work. This could cause difficulty when they tried to take the quiz a second time.

We had a variety of occasions in which students who lost conductivity ended up with multiple versions of phantom quizzes that they couldn't access and on which they were given grades of zero. Although we notified EDU of this issue, it was never possible to get a better resolution than don't let it happen. The lost of connectivity might be only an instant, but this was still a source of difficulty. Because some students were required to take the quizzes a maximum of three times, even though they had never intentionally quit, they found that they were not allowed to complete the quiz they had started and were not allowed to start a new quiz.

Another source of difficulty was the grading tolerance issue. One needs sufficient tolerance to eliminate incorrect answers, but not overly restrictive tolerance that fails to give credit for correct answers. For example one needs enough accuracy to be sure that a Z critical value is not marked correct when a t critical value is requested. Students might enter too few digits because of carelessness or because their computational software did not give the accuracy

requested. Some of our questions that required that students explicitly calculate certain summary statistics were written based upon an earlier release of Minitab. A newer release of Minitab continues to perform the calculation, but presents the results with less accuracy than in the past. When students use Minitab to do these calculations, they often get the impression that they have made an error. In actuality, it is an issue with the question.

Unfortunately, EDU is a very poor environment in which to write or edit the types of questions we want to use. While we could rewrite questions when issues occur, this is a very time consuming process. We are seeking a more productive environment in which to create new questions and modify existing questions. It is important to be able to customize questions easily and quickly to match the terminology of the textbook, the software package, and the instructor.

As another example, our grading algorithms were designed under the assumption that exact values would be calculated in Minitab and not approximate values retrieved from a table. However, when another instructor used the quiz and encouraged the students to use a table, this created a new issue. We are unsure of how to best handle this situation. This issue is related to the tolerance issue mentioned above.

Because some answers are case-sensitive, another issue occurs when students entered answers with the wrong case. When these are graded as incorrect, this becomes a source of aggravation. The appropriate wording of the questions is a simple source of conflict. No matter how carefully one words a question, other instructors may use a slight variation in the wording. This is especially true in multiple sections in smaller schools. The trouble comes when students are only familiar with an alternative wording. Hence they may not perceive the question correctly and thus receive less credit than for which they are entitled.

Students may feel it unfair when they are not given partial credit for answers on electronic quizzes. Some instructors do not use electronic quizzes for this reason.

Students also feel uncomfortable when they are confronted with a slightly ambiguous question. When the person that has written a question is in charge of the class, they know the intention behind the question and can easily clear up the ambiguity. When others who are less familiar with the wording use that question, this becomes a source of concern to both the students and the instructor. If the students don't get the main point behind a question, they may feel alienated when they are asked to retake the question. Ambiguous questions often occur on electronic quizzes. The same thing happens when a different textbook is used from the textbook from the quiz is constructed.

In a typical course situation each instructor has certain preferred ways of constructing questions. Even when trying to agree on common questions for exams, different instructors often feel compelled to reword the question to be consistent with the format and vocabulary that they use. When one is using an electronic question shared among several sections, what is perfectly valid wording to one instructor may be a source of some difficulty to another instructor who prefers slightly different wording. The format involved in the questions is another source of concern as students and instructors often prefer different question formats. They want to use the format with which they are most comfortable. This is an issue that we have tried to address by getting a consensus among existing instructors. But, when a new instructor uses the same questions this again becomes a source of concern.

Of course we attempted to ensure that all questions were well designed and tested. This was made difficult because the answers were randomly generated with multiple branches. It is often not possible to test all possible branches and answers. Occasionally a question, that has worked well, would take an untested branch and fail. Instead of the quiz being a source of learning to the students, it became a source of confusion and frustration. Students became unsure how they should respond. If they believed the quiz grading was correct then their understanding was incorrect. If they thought the quiz grading was incorrect, they became frustrated.

HARDWARE ISSUES

In order for a quiz to work successfully, the physical connection between the student's laptop and the server must be maintained. The software that we are using is such that if conductivity is lost even for a very short period of time the quiz may be handled incorrectly. Unfortunately the Internet cable connecting the classrooms and the server has a bandwidth

limitation of 10 megabits per second. Not infrequently the network becomes saturated and some of the data packets are lost. This causes the quizzes to be corrupted. This issue is even more common when students try to connect to the Internet wirelessly. The wireless connection is quite marginal and data packets are frequently lost.

Even when the students are using a physical cable, the connection is dropped a few percent of the time. This occurs for two reasons. Some of the pins in the Internet jacks are loose and this can cause a momentary loss of signal. A second reason is that some of the students' cables have loose connections and these also can cause a loss of signal.

SOFTWARE ISSUES

There are a number of pieces of software involved in our program. Students run Excel and Minitab in a Windows XP environment. Internet Explorer is employed to access the quizzes. They may be connected through Blackboard into an EDU environment. Finally we have servers running the programs required to administer the quizzes.

There is also an issue with the use of Internet Explorer. Version 7 caused some quizzes to become unusable even thought they worked perfectly in Version 6.

We have encountered a variety of issues associated with software. Blackboard promoted that it could be used as a convenient front end to EDU so our administrators set up some of the sections this way. Other sections connected directly to EDU. Each method created its own set of issues. Access through Blackboard caused major issues partway through the spring semester of 2006. It caused some quizzes to become inaccessible and others to be handled incorrectly. The cause of this issue was never resolved so some sections began bypassing Blackboard and using EDU directly. This generated a new set of issues. Direct connection created difficulties in registering students. When the students self register and create their own passwords, some students have multiple usernames. This creates an administrative nightmare.

STUDENT ISSUES

Another issue was that before a student had finished and graded an electronic quiz, they would lose network connectivity and their work. The system would record that they had taken the quiz but would give them a grade of zero. Also the system often generated phantom quizzes that the student could not see or access. Some students had more than five such phantom quizzes. Because we had an experimental design with a predetermined number of quizzes, how should we deal with these issues and not jeopardize the experiment?

Unfortunately while we are working to minimize the above issues, they have a lingering impact on our experiment and student attitudes towards electronic quizzes. When students perceive that they are not treated fairly by the electronic quizzes, they are less apt to take them seriously.

Let us summarize the student issues from three perspectives. First, when the program terminates unexpectedly the students can not learn from the question and are unsure what to do on the retake. Second, when students are exposed to many different ways of taking the quizzes they may become confused by all of the options and they lose sight of what they are trying to learn. They are called upon to decide when to retake a quiz, how and when they should get help, and whether to simply retake and hope for an easier quiz.

A third perspective concerns grading. The loss of partial credit may generate a sense of unfairness. We hope is that allowing students to retake the quiz multiple times is an alternative to partial credit. If they get a portion of a question wrong, they are expected to retake the quiz and correct their errors. One of us feels that this is a superior alternative to giving partial credit for partial mastery. But if an instructor also feels strongly that partial credit is an important grading method, this may cause students not to treat electronic quizzes and mastery as the positive learning experience they are designed to be.

INSTRUCTOR ISSUES

The issues mentioned above cause a certain amount of anxiety among instructors. They impact an instructor's willingness to use the electronic quizzes. There are also other issues unique to instructors. It is common in the United States to judge the teaching ability of an instructor by

student opinion surveys at the end of the semester. Many instructors feel that if they use electronic quizzes to master the course material, they will be rated less favorably. In addition, instructors also know that many of the technical issues are beyond their control and do not believe that the benefits that may accrue from using the electronic quizzes outweigh the disadvantages.

The complaints generated by any of the above issues must be handled on an individual basis. They have the potential to defeat the whole purpose of the electronic quizzes. That is to save time and to enhance student learning through repetition. No matter how careful we attempt to design the questions for the chosen environment there were always issues.

Simple human errors during the experiment were another issue. On one occasion an instructor gave the wrong type of quiz in one section, and on another occasion an instructor forgot to give the paper quiz. How should an instructor deal with these difficulties? Should data from mistakes be deleted or modified for inclusion in the analysis. Our small sample size makes this issue more critical. An alternative was to wipe out the incorrect data and through e-mails give the section the correct quiz type after recognizing the mistake, but this may lead to other issues of concern. We are working on a central database to better track and identify such issues as they occur.

We had hoped to recruit more faculty members to our experiments. Unfortunately some saw no upside and many downsides. They were unsure that the students would learn more. They believed that the amount of time that they could spend teaching the material would be reduced if quizzes were given at the end of each class. They also believed that the student workload (and anxiety) would be increased. This comes back to the attitude of the instructor. What was the upside of trying new experiences with the students as opposed to simply doing what had worked previously?

Some instructors, who initially volunteered to participate, quit at the first sign of difficulty. It was not worth their time to engage in an experiment with known issues. Others were quite robust in their attitude because of long range benefits that can accrue from performing objective learning experiments. We feel that it is important to recruit more such individuals so that we can objectively evaluate the merits of different approaches.

There is also the inevitable issue of scale. Something that may work well with one instructor often runs into the difficulty when there is more than one instructor. When there are multiple instructors, there is a need to publicize and track issues as they occur. There is also a need to agree on the solutions and how to implement them.

Let us summarize some of the key points. We gave a series of quizzes with different treatments. Each quiz was graded on a 10-point scale. Students were given 10 to 20 minutes to complete the quiz and took it one, two, or three times. When a quiz did not grade correctly, students were generally given an opportunity to retake. In one extreme case, after several unsuccessful attempts to fix the quiz, we decided that the data were so confounded that it was best to treat the quiz results as missing data. When any issue arose the instructor had the option of manually altering the grade. There were a number of factors that confounded the data. These factors included simple loss of network connections, quiz ambiguity, numerical accuracy, and quiz answer correctness. In the next section, we talk about our preliminary results.

PRELIMINARY RESULTS

A preliminary analysis of the data from the two experiments with paired sections did not indicate that the type of quiz had a statistically significant effect on student learning. We used a multiple regression model with covariates to come to this conclusion.

All of our experiments suffer from small sample sizes. Because of a variety of conflicting constraints, we were limited to sections of approximately 30 students in each treatment. An analysis suggests that the size of the effect we can measure must be 10% or more. However, as a result of attending numerous conferences and examining the literature, we are aware that typical results in educational experiments involve at most single-digit percentage effects. Therefore one of us believes that while we may have a very legitimate, but small effect, our sample sizes may prevent us from verifying this hypothesis.

This is especially true when the effect is compounded by the numerous difficulties that we have presented above. One reason for similar experiments is the hope that if we carefully follow the design we will be able to better test our hypotheses.

One of us has the following exploratory evidence that the use of electronic quizzes enhances learning, but may be more short-term than we originally imagined. In a set of exploratory experiments in two sections, he gave an electronic quiz with three retakes to one section and the other section had only a single paper quiz. In the very next class, a similar quiz was given. The electronic-quiz section scored almost two points higher on a 10-point scale than the paper-quiz section. Unfortunately this difference did not seem to persist over the entire semester.

One of us believes that the benefits are real, but the limitations on our experiments prevent us from detecting the difference in learning. Our hope is that by performing similar experiments in the spring semester of 2007 with better attention to some of the issues discussed above that we will be able to quantify the value of electronics quizzes.

SPRING SEMESTER OF 2007 EXPERIMENTS

Aware of the numerous issues that we encountered during the fall semester of 2006, we hope to eliminate many of them during the spring semester of 2007 when we undertake three more experiments. In these experiments, we will be using six sections of our introductory applied probability and statistics course taught by three different instructors. Again, each instructor will teach two sections of the course on the same day during adjacent time periods when possible. We will use an experimental block crossover design similar to the ones employed in the fall semester of 2006.

One major difference will be that one of three paired sections will explore the implications of the "just-in-time" approach. Before each class, students will be assigned several short answer questions on the material that they are to read for that class. The other section will only be assigned the same reading. The lectures for the two classes will be on the same topic. The quiz at the end of each of the two sections will be the same. Again, we will employ a quiz designed to require five to 10 minutes, but the students will be given 25 minutes. All students will take the quiz once in class. If they have time, they may take it two more times in class or they may take it twice outside of class. In any case they are requested to take it a total of three times before 8:00 am two days after that class. The purpose of the additional day is to remove some of the stress that we are told that students might be under.

The second of the paired sections we will conduct an experiment very similar to the one employed during the fall semester of 2006. The only differences will be the additional day for outside of class quizzes and roughly weekly, instead of daily, quizzes.

In the third of the paired sections, the instructor who performed the paired section experiment for the introduction to quantitative methods students will use the block crossover design to test the effect of the electronic quizzes and the effect of repetition, again with the additional time for outside of class quizzes.

We expect to be able to present two set of analyses of our experiments at the 2007 IASE Satellite conference, our experiments for the fall semester of 2006 and the spring semester of 2007. This work has been supported by a grant from the Davis Educational Foundation.

REFERENCES

Aieta, J., & Rybolt, W. (2004). Technology and Assessment in Quantitative Methods at Babson College. *Proceedings of the 17th Annual International Conference on Technology in Collegiate Mathematics*. New Orleans, LA: Addison-Wesley & Prentice Hall.

Bell, D. (2007). Electronic Student Assessment Systems. M. Hvidsten and B. Yoshiwara (Organizers). Panel presented at Joint Mathematics Meetings. New Orleans, LA: Mathematical Association of America.

Gonzalez, J., et al (2006). Formal Assessment of an Innovative Web-Based Tool Designed to Improvement Student Performance in Statistics. Paper presented at 7th International Conference on Teaching Statistics. Salvador, Bahia, Brazil: International Association for Statistical Education.

Heizer, J., & Render, B. (2007). Automated Homework and Exam Grading in the OM Course: Benefits to Faculty, Benefits to Students. Paper presented at 37th Annual Meeting of the Decision Sciences Institute. San Antonio, TX: Decision Sciences Institute.

Johnson, G. (2006). Optional online quizzes: College student use and relationship to achievement. In *Canadian Journal of Learning and Technology*, Vol. 32(1) Winter 2006. Etobicoke, ON, Canada: Association for Media and Technology in Education in Canada.

Kletskin, I. (2007). Maple TA: Making Automated Assessment A Reality. Paper presented at 19th Annual International Conference on Technology in Collegiate Mathematics. Boston, MA: Addison-Wesley & Prentice Hall.

Whiting, D., & Scott, D. (2006). YSTATTEST: A System for Automated Online Test Creation and Correction. Paper presented at 7th International Conference on Teaching Statistics. Salvador, Bahia, Brazil: International Association for Statistical Education.