A Comparative Study of Partial Credit Assessment and Computer-Based Testing

Eddie Fuller, Majorie Darrah, and David Miller
West Virginia University
P.O. Box 6310
Morgantown, WV 26506
ef@math.wvu.edu, mdarrah@math.wvu.edu, and millerd@math.wvu.edu

Introduction

The use of technology in mathematics classrooms has grown in prevalence over the last decade. Beginning with the use of graphing calculators to guide discovery and leading into the use of computer based testing in systems such as WebWork and WEBAssign or course management systems such as Moodle and Blackboard. In most cases, educators are confronted with the inability to provide feedback to students that is targeted towards the work done in solving a problem since they only have access to the final answer presented to the testing system. As a result, many students may feel frustrated by the impersonal nature of the process. Others may be adversely impacted by the limitations placed on them in communicating their answers.

On the other hand, mathematics tests graded by the computer provide a myriad of benefits for the instructor: no grading of hundreds of papers, consistency in the grading, ability to make hundreds of versions of each question to reduce cheating, quick feedback to students, ability to analyze results, etc. The students also realize some of these same benefits: quick feedback, consistency in grading, etc. Historically, many approaches to such systems have been utilized and studied. Most systems to date have focused on multiple-choice testing, especially systems used for standardized testing such as the ACT or SAT. Other systems have also appeared that had the capability to allow alphanumeric input such as the WeBWork online homework system. Symbolic input of answers using traditional mathematical notation has been generally lacking but in one example the University of Georgia at Athens developed and deployed a computer-based testing system in its college algebra and precalculus classes beginning in the mid 1980s. This system allowed for student input to be open ended and had an equation editor that allowed for symbols to be submitted for answers and stored on a centralized server. Hunt (Hunt, 1993) conducted a study of this system and found that this system primarily assessed a student's ability to recreate an algebraic process leading to acceptable answers. She cites Schoenfeld (Schoenfeld, 1988) who comments elsewhere that many learning environments encourage rote memorization and have assessment structures that reward that strategy.

In the current work the authors seek to find an application of the computer based testing environment that avoids the pitfalls observed by Hunt and Schoenfeld while maintaining the strengths of such a system. No matter what the benefits of computerized testing, a concerned teacher is always asking herself questions such as:

1) *Would students do better on tests if they are given partial credit points?*
2) *On average how many points better would students do if they were given partial credit on a test?*
3) *Do computerized tests change the outcome of the students' grades?*
4) *Is there a way I can make it up to the students for not giving partial credit that doesn't involve just giving away points?*

At the heart of these questions, such a teacher is focusing on the assessment of true understanding instead of the demonstration of algorithmic repetition and so we constructed a testing paradigm that would enhance our traditional computer-based term exams with a comprehensive supplemental assessment other than the final exam that would present the students with an opportunity to recover missed partial credit points. The questions in this supplemental quiz were chosen from a large set of possible questions covering all topics in the

course so that students would be more likely to demonstrate understanding of the topics rather than rote memorization of small sections of the course as they tend to do on term exams. We hypothesized that this process would be comparable to an analysis of a student's work and the awarding of partial credit.

In order to explore this concept, it was decided to design an experiment that would try to determine what a student would have received had they been given a pencil/paper test graded with partial credit versus a computerized exam. In the fall of 2007, a study was implemented with all the students enrolled in Introduction to Calculus at West Virginia University. It did not seem fair to select students at random to take pencil/paper test graded with partial credit, while others would take the computerized tests. Although this control group method is widely used, the comparison of these two groups would have several drawbacks, not the least of which is that the students may find it unfair. It also did not seem that we would produce consistent results by giving students some tests on the computer and some tests using pencil/paper graded with partial credit. This would allow comparison of the same student on the different types of testing, but the material for each test would be different, so there is no way to know if the method or the material on the test made the difference in the score. Both of these methods were flawed in their own ways.

The experiment that was designed compared a student against himself/herself taking the same test on the computer and taking the test with pencil/paper graded with partial credit. On first glance this approach seems like it would produce biased results, but the experiment was designed in the following way. Exams were administered online using the WebCT/Blackboard assessment module. Questions ranged from multiple choice to open answer questions with calculated answers as shown in Figure 1.

Each student would take the test on the computer, but they would also have a numbered scrap work sheet where they were asked to write down all their work. In this way the scrap work could be graded as if it were a pencil/paper test to see if there were points that the student would have gained if the test were graded by hand. A random sample of students was selected and their scrap work graded. The scores on the computerized test would be compared against the scores on the scrap work graded version of the test. This way they were comparing each student against himself/herself on the same test taken in both ways.

Background of Problem

There have been many research studies that have examined various aspects of using computers in a course. These studies have varied from looking at using computers to do homework and administer examinations, to examining students' attitudes towards using computers (Hunt, 1993). The researchers in this study examined business calculus student's performance on computerized exams versus their performance on paper and pencil (traditional) exams.

Bennett et al. (2008) studied the NAEP mathematics examination scores for eighth graders when the exam was administered via the computer versus paper and pencil. The results showed that there was a significant difference between students' score on a paper and pencil exam compared to the same exam given on a computer. Puhan et al. (2007) found no significant difference on a certification test when it was administered on the computer versus paper and pencil. Kim and Huynh (2007) also found no significant difference at the item-level, subtest-level, and whole test-level for computerized versus paper and pencil exams for end-of-course examinations in the subject areas of algebra and biology. Poggio et al. (2005) study looked at computerized testing versus paper and pencil testing for 7th graders in mathematics. They found that there was no significant difference between the paper and pencil exams versus computerized exams. Cole et al. (2002) examined computer-based examinations compared to paper-based examinations on the Force Concept Inventory (FCI). The authors showed there were no appreciable differences on FCI scores or FCI items based on testing methods. However there were differences in FCI scores with respect to gender and time of administration. Bugbee (1996) examined the research showing that paper and pencil testing was equivalent to computerized testing and stated five general

conclusions. Ashton et all (2006) worked on the Project for Assessments in Scotland using Information Technology (PASS-IT) that designed a system that mimicked partial credit. In this system, students could receive full credit on a problem if they knew how to do the problem. However, if the student did not know how to work completely through a problem or was unsure that they could work through the problem completely, could choose the option where they supply intermediate steps as they work through the problem. In this way students could earn partial credit on the problem but usual not full credit. This project shows promise on designing a system that would incorporate partial credit on computerized examinations.

The researchers in this study have designed another way to award partial credit to students on examinations in a business calculus course via a super (bonus) quiz at the end of the semester. This study examines whether computerized examinations with partial credit earned through the super quiz is equivalent to students earning partial credit on a paper and pencil examination.

Outline of Study

Participants. The target population for this study was the students enrolled in an Introduction to Calculus course at West Virginia University in the fall semester of 2007. There were six sections of this course offered by three different instructors with 80 students enrolled in each section for a total of 480 students. This is a required course for business, social science, forestry, or pre-pharmacy majors.

Instruments. The exams for the course were created in the WebCT/Blackboard course management system. They were a combination of multiple choice, short answer (computed questions), and fill in the blank (sample questions are provided in Appendix A). Most questions required some calculation work, but some did not. Students had 55 minutes in the computer lab to complete each exam. The exams were timed so no student received more time. Exams contained a varying number of questions: Exam 1 – 16 questions, Exam 2 – 16 questions, Exam 3 – 11 questions, Exam 4 – 11 questions.

In order to facilitate students showing work on scrap paper, the instructors prepared scrap paper that was numbered and divided into sections for each question. Students were reminded at the beginning of the exam to show all work on their scrap paper and all students were required to turn in the scrap paper before leaving the examination area (computer lab).

Procedure. All students in the course were informed of the study and were asked to fill out a consent form for their data to be used in the study. Most students completed and signed the form. Thirty students were randomly selected from each section (180 students were selected in all), the names were matched to the students who filled out the consent forms to ensure that the students chosen had consented to participate. All students completed and handed in scrap work for each exam.

For several reasons by the end of the semester the sample had decreased to 97 students (20% of the overall students enrolled in the course). The reasons for this loss were (1) common attrition in the course, (2) randomly selected students did not consent for their data to be used, (3) students who did not complete all four exams during the regular scheduled times were not used, and (4) one section had to be dropped because for one exam the scrap work was not usable. The number from each section was Section 2 – 17, Section 3 – 20, Section 4 – 21, Section 5 – 16, and Section 6 – 23.

In order to provide consistent grading, one graduate student was hired to do all the grading of the scrap work. Two professors prepared the grading rubrics for the graduate student and she graded all papers using these rubrics. She not only graded the scrap work, but she also looked at each individual exam to make sure which of the problems they already had correct for which they already received full credit. The grader was able to obtain a score for the computerized exam and then a score if they would have received some partial credit for the ones that they did not

receive credit for on the computerized exam. This meant that the scrap work grade was always greater than or equal to the computerized exam.

Analysis of the data

Table 1 below summarizes the averages for the sample group.

Exam	Comp. Average	Partial Credit Average	Average Diff.	Max Diff.
1	74.49	80.35	3.86	17
2	66.52	67.76	3.21	13
3	65.49	71.60	6.10	29
4	72.02	76.04	4.03	14.6

Table 1: Comparison of Test Grades and Scrap Work Grades

Using each student's score on the computerized exam compared to their score with points added from the graded scrap work, a paired t-test was run on the student data. For all four exams there was a significant difference between the computerized exam grade and the grade on the partial credit graded scrap work at the $\alpha = 0.01$ level.

The total points the student would have earned on the 4 tests in partial credit was then compared with the score for the Super Quiz. Table 2 summarized this data. Although at first glance, the average amount of partial credit points and the average on the Super Quiz seem to be similar, a two sample z-test comparison show that the two numbers are significantly different at the $\alpha = 0.01$ level. A paired t-test was run on the student data and again there was a significant different between what the student earned in partial credit and what the student earned on the Super Quiz.

	Partial Credit	Super Quiz	Difference
Average	17.17	23.42	-6.24
St. Dev	9.78	7.03	13.75

Table 2: Comparison of Partial Credit Points and Super Quiz Points

After further investigation it was determined that for a group of 55 of the 89 students the Super Quiz was indeed a good substitute for the partial credit points that they would have earned, however in some cases these scores on the two varied significantly.

Conclusions

The data indicates that the students in this course would have gotten higher grades on the test if they would have had their work graded using partial credit. The question remains, *how much better would they have done?* Do the bonus points made available through the Super Quiz at the end of the semester make up for the missed partial credit points on the four exams? The motivating notion for such a supplemental assessment as the Super Quiz was that students on average would gain perhaps 6-8 points on an exam due to partial credit. This study indicates that our initial estimate of partial credit effects was high for most exams, and that the Super Quiz overcompensated for the lack of partial credit. Specifically, it was observed that students would have received an average of 17.17 points over four tests, which was approximately 4.3 points per exam. Students benefited from the Super Quiz by an average of 23.42 points suggesting a reduction in the number of points assigned to it. Nevertheless, this methodology has the advantage of giving students the opportunity to demonstrate understanding in the computer-based testing environment in the case that a prior opportunity failed to assess their actual mastery level. Ultimately, this is the goal of assessment and we feel that this system gives students strong psychological support when encountering computer-based testing by providing a clear alternative to the partial credit they have come to expect. It is of interest to know whether there is a demographic that tends to benefit more from this arrangement than others. We will also investigate alternatives to the Super Quiz that will compensate for the lack of partial credit in the

course. It is clear from the data collected that the number of points awarded for the bonus assessment does not need to be in the range currently used. In future work, we plan to revisit this study by revising the scoring of the Super Quiz and assessing student affect simultaneously to better measure this psychological impact.

Acknowledgements

References

Ashton, H.S., Beevers, C.E., Korabinski, A.A., Youngson, M.A. (2006). *Incorporating Partial Credit in Computer-aided Assessment of Mathematics in Secondary Education.* British Journal of Educational Technology. Vol. 37, Issue 1, pg 93 – 119.

Bennett, R.L., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., and Yan, F. (2008). *Does it Matter if I Take My Mathematics Test on Computer? A second Empirical Study of Mode Effects in NAEP.* The Journal of Technology, Learning, and Assessment. Vol. 6, No. 9. Available from http:www.jtla.org

Bugbee, Alan C. (1996). *The equivalence of Paper-and-pencil and Computer-based Testing.* The Journal of Research and Computing in Education, Vol. 28, Issue 3.

Hunt, A. (1993). *The Effects of Computerized Testing on the Performance and Attitudes of College Pre-calculus Students (Mathematics Attitudes).* Doctoral Dissertation, University of Georgia, DAI, 54, no. 04A, 1273.

Kim, D.-H., & Huynh, H. (2007) *Comparability of Computer and Paper-and-Pencil Versions of Algebra and Biology Assessments.* Journal of Teaching, Learning, and Assessment, 6(4). Retrieved November 18, 2008 from http://www.jtla.org.

Poggio, J., Glanapp, D.R., Yang, X., & Poggio, A.J. (2005). *A Comparative Evaluation of Score Results from Computerized and Paper and Pencil Mathematics Testing in a Large Scale State Assessment Program.* Journal of Technology, Learning, and Assessment, 3(6). Available from http:www.jtla.org

Puhan, P., Boughton, K., & Kim, S. (2007). *Examining Differences in Examinee Performance in Paper and Pencil and Computerized Testing. Journal of Technology, Learning, and Assessment,* 6(3). Retrieved on November 18, 2008 from http://www.jtla.org

Shoenfeld, A. H. (1988). *When Good Teaching Leads to Bad Results: The Disasters of Well-taught Mathematics Courses.* Educational Psychologist, 23, 145-166.