3-2015

E-testing in Graduate Courses: Reflective Practice Case Studies

Martin W. Sivula Ph.D.
Johnson & Wales University - Providence, msivula@jwu.edu

Elizabeth B. Robson J.D.
Johnson & Wales University - Providence, ERobson@jwu.edu

E-testing in Graduate Courses:

Reflective Practice Case Studies

Martin W. Sivula, Ph.D.

Elizabeth B. Robson, J.D.

Johnson & Wales University

College of Management, Graduate Studies

March, 2015

These case studies were from courses in the fall and winter terms 2014-2015.

Abstract:

Do testing and exam conditions make a difference in final exam grades? Do testing "out of class" and "in class" produce different results over the same courses? Several graduate courses (N = 84) were tested under different and varying conditions. The majority of students were international, where English was a second language. In general, the "online" e-testers performed at a higher level than the ""in class"" testers with and without any time restrictions while test taking. Tentative implications might be that online exams (less controls) yield grades which are possibly higher, and may or may not be "grade inflated." On the other hand, possibly less controls in exams yield more learning and higher retention of course content.

Introduction

The assessment of student learning is of high importance no matter what the instructional delivery method is. Traditional assessment methods would be quizzes, tests, and exams given periodically over a term or semester over a given length of time. "A **test** or **examination** (informally, **exam**) is an assessment intended to measure a test-taker's knowledge, skill, aptitude, physical fitness, or classification in many other topics (e.g., beliefs) (Wikipedia, 2014, http://en.wikipedia.org/wiki/Test_%28assessment%29). During the course of several terms and classes two professors experimented with various "e-testing" strategies and their effect on student performance. They removed "in class" controls, such as proctoring, time limits for e-testing groups, and in some cases raised expectations for higher performance of the e-testing groups. We posit that given adequate amount of time and resources students can perform at a higher level through e-testing ("out of class") when compared to the traditional "in class" testing methods. We have divided our case studies into a quantitative analysis (case study one) and qualitative methods (case study two). Case study one collected data from archival faculty grading records, direct observations, and participant

observations. The second case study used documents, grading records, interviews, and direct observations (Yin, 2003).

Traditional Testing

Instructors would design courses where components carry different weights (usually percent) of some final grade. From these tests or exams (traditional or classical methods) of assessment would be proctored by the instructor in a real time, face to face, self-contained classroom. Minor assessments (quizzes) might not be announced or "popped" to the student while "in class". Tests and exams might be announced by the instructor so that students could properly (in their personalized fashion) prepare for the upcoming assessment. These assessments would have a date, time, and place to be administered. Also, there would be a time limit to their completion (50 minutes, 1 hour, 2 hours, etc.). The instructor would "proctor" the quiz, test, and/or exam, possibly answering questions to confusing items, or clarifying items, and act as a controller for any potential academic dishonesty. These types of assessments or tests are referred to as criterion-based where the student is matched against the content covered by the test. The goal of these assessments is to determine whether the student has learned the material. Large scale assessments or "norm"-referenced test match the student against student over a given content. The term large scale refers to a population of test takers over given material and where an individual student lies with respect to the population of test takers (e.g., SAT, GRE). Large scale assessments follow the approximately the same protocols, however, sometimes they are referred to as "high stakes" tests where performance determines a highly valued outcome, e.g., success or failure into a desired college or university. Our research here is limited only to the criterion-based assessments.

Grade Inflation

Grade inflation is said to occur when higher grades are given when lower grades would have been given in the past. It has been an ongoing academic discussion for decades in the United States as well in Canada, England, and Wales (Wikipedia, 2014). According to Rojstaczer (2014), since about 1960 private institutions calculated GPAs have increased at an average rate of .15 on the 4 point scale per decade. In 1960 at a selective, private college or university an average grade might have been C+, and now decades later the average at the same school might be B+ or even higher. However, Rojstaczer also claims that the inflation rate in public universities and colleges is less and much less at selective engineering schools such as MIT (Rojstaczer, 2014, http://www.gradeinflation.com/). Neili (2014) states that in the Vietnam era, the student unrest in that era might be the starting point of grade inflation. Faculty lost their nerve to maintain high standards and gave in to students wanting more of a "voice" on curriculum, course requirements, and ultimately grades. Also, weak administrations trying to make students happier, less violent, and less prone to demonstrate made for student contentment and in some cases higher grades. Another contributor to grade inflation is student evaluation of faculty. Faculty members in the early stages of their careers need positive (high) student evaluations for gaining rank and tenure. So some students might rate a faculty member's performance high because the faculty member is an easy grader. Or "if you give me the grades I want, then I'll give you the evaluations you want." In many cases being a tough grader is only going to impact a very few (usually outstanding) students. The rise of one's self-esteem as has spread grade inflation since the 1960s where having high self-esteem and a high opinion of oneself is a contributor to self-confidence and high achievement. This formula results in students feeling good about themselves and consequently they'll come to love real

learning and become more productive citizens. This attitude possibly might be a maybe major contributor to grade inflation especially in American student population. The last and final factor is the competition for placement into selective graduate schools where the students need to get higher grades and higher GPAs to enter the most prestigious graduate schools. Therefore, they are getting inflated grades at the undergraduate level. When applying to graduate schools if one institution gives high grades and another institution gives low grades then one institution might look inferior to the other. So if one college gave a "C average" as a grading standard, and another one gave a "B or B+ as a grading standard" then the one that gave the "C" might feel inferior to the one they gave the B+. The only way that grade inflation might be reduced is for colleges and universities to set realistic standards at the regional or national level, but this will probably not happen in the near future (Neili, 2014).

E-Assessment

Recent advances in technology hardware and applications are influencing testing and assessment. Several decades ago computer-based testing could be performed on a single PC (standalone) given the proper application software or subsequently connected to a computer network. Nowadays, there are various electronic communication devices (IPads, IPhone, Smartphones, tablets, computers, etc.) on which one can receive and send tests/exams and other performance-based data. Test security, academic dishonesty, and comparative results in all testing environments are among some of the issues currently being researched.

"In its broadest sense, **e-assessment** is the use of information technology for any assessment-related activity. This definition embraces a wide range of student activity ranging from the use of a word processor to on-screen testing (http://en.wikipedia.org/wiki/E-assessment, 2015)."

Wikipedia (E-assessment, 2015) states the advantages over pencil and paper tests include:

1. "lower long-term costs
2. instant feedback to students
3. greater flexibility with respect to location and timing
4. improved reliability (machine marking is much more reliable than human marking)
5. improved impartiality (machine marking does not 'know' the students so does not favor nor make allowances for minor errors)
6. greater storage efficiency - tens of thousands of answer scripts can be stored on a server compared to the physical space required for paper scripts
7. enhanced question styles which incorporate interactivity and multimedia."

When dealing with online testing with unlimited resources and unlimited time, students have at their disposal a variety of tools that they can use in a multitude of ways. However, test design has to come into play here. At the graduate level of education the cognitive level of the exams should be higher. Focus should be on analysis of content (breaking down) which requires critical thinking, synthesis, putting things together asking new questions which implies creative thinking, and evaluation, judging

content for its worth or merit. At the lower levels, knowledge (simple recall), comprehension (which tests for understanding), and application (which tests for knowledge in new situations might also be used for graduate assessments but to a lesser degree (Bloom's taxonomy, 2015, http://en.wikipedia.org/wiki/Bloom's_taxonomy). The major courseware packages provide for multiple test formats such as fill in, multiple choice, essay, matching, short answer, and true-false. By working with Bloom's Taxonomy one can match the cognitive level with a question type and then design the content of the question. The procedure can be done with a "pencil and paper" test, but it is more labor intensive.

Case Study One – Static Group Comparison – Research Methods Course

The purpose of this case study was to determine if different testing conditions affect final exam grades over the same content. Specifically two hypotheses were addressed:

H0: $\sigma 1 = \sigma 2$

Ha: $\sigma 1 < \sigma 2$

Where σ1 represents the standard deviation of scores of the online class and σ2 represents the standard deviation of the "in class" testing group (represented here by the standard deviations √ (σ) (σ)). It is hypothesized that with virtually unlimited resources over a four day period, the standard deviation of the online testing group would be lower than the proctored "in class" group.

This hypotheses addresses whether the two means of the different groups are statistically different.

H0: $\mu 1 = \mu 2$

Ha: $\mu 1 \neq \mu 2$

The alternative research hypothesis favors that the means of the two groups will be statistically different. One group would clearly out preform the other, but not knowing which one, where $\mu 1$ = the online group, and $\mu 2$ = the "in class" group (traditional).

Method

The testing sample was comprised of two graduate research methods sections (face to face "in class", n=20 and outside of class-online, n=21). The normal university scheduling procedure was used and no special treatment or random assignment was applied for both groups of students. The assumption here is that both groups are equal to begin with and there are no significant differences between them. The static group took the final examination "in class" within a maximum time period of two hours. The online testing group had a maximum of four days to complete the exam with unlimited resources at their disposal. This group was instructed to do the exam alone and not with any other person or human assistance. The instrument was a comprehensive final exam consisting of twenty-two short answer essay questions which had multiple sub-parts per question. The content validity was established from the course textbook, class notes, and Blackboard class materials. Since it was an exam of essay type, reliability could not be estimated with any real degree of confidence. Each item of the exam was scored

on a six point scale, 5=excellent; 4=good; 3=average; 2=below average; 1=poor; 0=blank. Both groups where given a brief topic review "in class" prior to the exam. The online testing group used the Blackboard assessment tool. The instructor set the testing parameters on Blackboard so students could go in and out of the test, leave it, come back, and had unlimited attempts over the four day period. The "in class" testing group had a maximum two hour testing period with the instructor acting as the test proctor. At the end of the four day period all online testing students had completed the exam, not one was marked late. The instructor then graded both classes in a timely fashion. Both groups were given a total percent for the exam. Statistical analysis was then performed with means and standard deviations calculated for both groups, t-tests, equal variance estimates, confidence intervals, and Cohen's d statistic to estimate effect sizes.

Results

The online (n = 19), "out of class" exam group's M = 90.18 and SD = 6.92 with the Mdn= 92. The "in class" (n = 20) proctored exam's group M = 80.10 with an SD = 11.18 with a Mdn = 85. The coefficient of variation was calculated for both groups. The online testing group had a CV = 7.6%, compared to the "in class" coefficient of variation (CV = 13.9%). We note here that the variation is almost double that on the online testing group. Using an F-test for equal variances, and the above hypothesis F = .38 and p = .023, so the null hypothesis was rejected in favor of the alternative research hypothesis that the online group variance in test scores would be lower. The second hypothesis that the test scores would be statistically different is also supported. Using a separate variance estimate t-test, the results were as follows: t= 3.40 and df=31.9 with a p = .001. Therefore the second null hypothesis is also rejected in favor of the alternative research hypothesis of the means being statistically different ($p < .01$). The two sample t-interval is: 95% CI (4.04, 16.11) and note here that zero is not in this interval. Lastly the effect size was calculated using Cohen's d = 1.084. So as far as the exam is concerned the online testing group performed approximately 34% better than the "in class" group within the given conditions.

Limitations

This was not a true experimental design, and there was no random assignment to testing groups, therefore you cannot attribute any statistical differences to the treatment or intervention of "e-testing." The statistical tests only suggest that there is a difference from mere chance occurrence with the observed data. We can only say that the two groups are statistical different with the given data, and what makes this difference (e-testers) is only speculative, given the testing conditions.

Case Study Two – Instructor Narrative – Global Economics

I have had an advantage in teaching the same Global Economics course in both an online (GEO) and an in-class (GEI) format. The first term teaching GEO, I saw significant grade inflation at both the assessment and the final grade level in the online class. I was not entirely sure, however, that this was only the result of the online testing conditions. This same term, I was also teaching GEI, an in-class section of Global Economics with 35 students, compared to the 8 in my online section. The test/exams in GEO included the answering of essay questions requiring critical thinking and detailed analysis whereas the exams in GEI had a combination of multiple choice questions and short-answer essays. Both GEO and GEI two hours to complete the exams, but, obviously, the GEO group had access to resources that

the GEI students did not. Additionally, with only eight students in GEO compared to thirty five students in GEI, there was a significant possibility that those eight students received more of my personal attention through our weekly online discussions and interactions than that of my traditional GEI group. Initially, I changed the final exam in GEO to counter the grade inflation I had seen with the midterm exam. Requiring a minimum of three relevant sources, increasing the scope of the analysis for the case, and raising expectations for grammar, punctuation, and spelling appeared to be the most rational adjustments. This appeared to counter the grade inflation I had seen at midterm. I ran into a problem in testing my theory, however, because the GEO class did not immediately run again due to scheduling.

What I did instead was to incorporate the tools that I suspected were working to the benefit of the GEO students into my GEI sections. That following term, I taught sections of GEI with both of those sections having 35 students. I used that opportunity to utilize the additional supports for students I had included in GEO the term before. Including weekly online discussion boards to continue the discussions we started "in class" insured that all students had the opportunity to be heard. Adding narrated PowerPoints to weekly materials provided opportunities for students to review important concepts as many times as they needed. Uploading video recordings of myself emphasizing key "take-aways" from the weekly readings and classroom discussions reminded students that I was available throughout the week for guidance and feedback. *I realized that by increasing opportunities for interactions throughout the week, I was also creating an environment whereby I could increase my expectations for assessment.* The only variation between the two sections was that I gave one section, GEI-1 an in-class final exam and the other section, GEI-2 the same online final exam I had used the prior term with my GEO class. GEI-1 students taking the in-class exam were given two hours to complete the case study analysis. They were unable to use any additional material or the internet. The students in GEI-2 were given an online version of that same case study analysis. They had one week to complete the exam which had to then be uploaded through Turnitin before uploading through the online submission tool. They had a rubric to follow detailing the expectations for different grades and had unlimited access to additional resources. *The grade distribution was similar in both sections.* It was clear, however, that both sets of students were capable of performing at the higher level with the additional online support and that the more rigorous online assessment could curtail the presumed grade inflation.

Discussion

Online testing seems to have positive results for both case studies. What about academic dishonesty and cheating, many instructors believe that in the online environment cheating is more pervasive. However, in a large public university study the frequency of cheating in the online environment was no more pervasive than in the traditional, in-class testing environment. If fact the opposite was found true, the typical form of cheating (e.g., panic cheating) was less likely to occur in online classes (Grijavla, Kerkviet, & Nowell, n.d.). So what type of controls do we need in an online exam? Many of believe that there is not one answer, but it depends on the course, content, test construction, time variation (unlimited or fixed) and the student. No matter how we design the test whether online or "in class", there will always be someone that will be dishonest, so we should always review in our own minds what assessment means (Everson, 2011). Current literature and thinking on grade inflation does not seem to

explain our findings. In our case studies here the two hour time limit for both the online and "in class" tests does not seem to be a factor. However, given that the online students know how to navigate the e-resources quickly might have given them some advantages, consequently receiving higher grades than the traditional "in class" group.

We did arrive at mixed findings where the research methods class "out of class" e-testing group performed higher on the final exam than their "in class" counter parts. The global economics final exam scores (in class and out of class e-testers) performed similarly on the final exam grade distribution. We can only conclude that given more online resources and time students might perform at a higher level.

Knowledge retention after a course is completed is whether "in class" or online testers have different retention rates? In a video training course delivered both online and "in class" it was predicted that loss of original knowledge gained would be 15%. When the groups were tested 10, 20, and 40 weeks after initial course completion their knowledge loss was in the area of 14% to 16%, right in line with the predicted value of 15%. Knowledge retention might be a latent or hidden outcome of instruction. However, when the assessment takes place is of key importance, perhaps immediately after a course may not be the best indication of knowledge acquisition (Wisher, 2001). Repeated retrieval of correct information leads toward long term retention. Students self-testing themselves is of great benefit to learned material. However, once the material is learned, many students will drop that material from further learning and testing. But repeated retrieval practice is a great way to enhance long term retention, even of already 'learned' material (Karpicke, 2007). This is where the e-testing environment can really shine because of the different modes of assessments possible, question formats, screen layouts, and software design. The professional publishing companies have assisted academics by providing numerous software applications together with their published textbooks. By using software from the publishing companies you can create a "question pool" of say two hundred questions where the student is given the correct response, the rationale for the correct response, and the topic/section from which it came. From this a "sample" is selected for an e-test. During our courses this has worked very well for us, the students seem well prepared and for the most part do "not panic" cheat from what we have observed. We do not know if there is a difference on knowledge retention with e-testers and in class testers from this research. However, we do believe that repeated practice with online e-tools enhances the learning and might lead to greater knowledge retention.

In our case studies presented here we gave the students various modes of the same material (printed, oral, graphic, audio, etc.) due to language differences. We might present written material, YouTube videos, and a discussion forum all around the same content topic. We have learned that providing videos (short duration 5-10 minutes) that can be played over and over again on very difficult topics provides great reinforcement for retention. Where possible, we also provide self-tests for the students chapter by chapter, topic by topic. Many times these are provided by the textbook's publishing company to accompany a given book. During the course of the term all groups where subjected to these types of e-interventions which also might have affected the final exam scores, possibly moderating the effect of the e-testing environment.

Conclusion

In conclusion we think that e-testing has a bright future. The positives (convenience, grading, anytime, anywhere, resource availability) outweigh the negatives such as academic dishonesty, technical problems, etc. The “in class” proctoring which might have an effect on “panic cheating” is also not present in the e-testing environment. The e-testing would be external to the face to face “in class” sessions, and truly enrich the human to human interaction within the traditional classroom. In fact, instructors might have more time in a traditional face to face class to focus on course content and not waste valuable class time proctoring “in class” tests.

References

Bloom's Taxonomy. (2015). Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Bloom's_taxonomy

E-Assessment. (2014). Retrieved from (http://en.wikipedia.org/wiki/E-assessment

Everson, M. (2011, March). Academic Honesty in the Online Environment. *e-Learn Magazine*.

Grijavla, T. K. (n.d.). *Academic Honesty and Online Courses.*

Karpicke, J. &. (2007). Repeated retrieval during learning is the key. *Journal of Memory and Language*, 151-162.

McGraw Hill CTB. (2015). Retrieved from http://www.ctb.com/ctb.com/control/topicShowAction

Nieli, R. (2014). *GRADE INFLATION—WHY PRINCETON THREW IN THE TOWEL.* Princeton. Minding the Campus Reforming our Universities. Retrieved from http://www.mindingthecampus.com/2014/10/grade-inflation-why-princeton-threw-in-the-towel/

Rojstaczer. (2014). Retrieved from gradeinflation: http://www.gradeinflation.com/

Test (assessment). (2014). Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Test_%28assessment%29)

Wisher, R. C. (2001). Knowledge retention as a latent outcome measure in ditance learning. *American Journal of Distance Education, 15*(3), 20-35.

Yin, R. (2003). *Case Study Research Design and Methods.* Thousand Oaks: Sage.