EVALUATING PEDAGOGICAL TECHNIQUES IN INTRODUCTORY STATISTICS: PROFICIENCY GRADING AND ASSIGNMENT RESUBMISSION

POSNER, Michael A.
Villanova University
USA

How do innovative pedagogical techniques improve learning and mastery of introductory statistics? This research study examines proficiency grading and assignment resubmission and compares them to traditional statistical teaching methodology in two introductory statistics classes. The control class received traditional numeric grades, while the experimental class received grades on a three-tiered proficiency ranking and the opportunity to resubmit assignments to increase their proficiency score. Students in the control class scored higher on a common final exam (although not statistically significant), and believed the material was better taught, while students in the experimental class claimed to have learned more and were more satisfied with the grading in the course. Future research will expand data gathering and improve the research design.

BACKGROUND

In a typical classroom, the conscientious statistics instructor imparts onto the students the daily topic which she has carefully prepared with real-life examples to bring the class to life. After enthusiastically professing about study design, biases, or the choice of statistical analytic techniques, she turns to the class and asks "Do you understand? Are there any questions?" Students lift their heads up from their notebooks and softly nod in affirmation. This is a typical statistics classroom (and many unlucky students have the instructors who use naked data, which lacks a real-life context, and lack enthusiasm in their presentations). This method fails to truly assess what students have learned. How do we enhance student learning? How do we encourage students to assess their own work and learn from their mistakes?

In the Spring semester of 2006, I had the fortune of taking a seminar on formative assessment with Victor Donnay and Dylan Wiliam, a gifted instructor and expert in the area (Wiliam & Black, 2006). This seminar (and the research presented here) was sponsored by the Math Science Partnership of Greater Philadelphia (funded through the National Science Foundation's Math Science Partnership grants). This seminar brought together high school teachers and college professors from around the Greater Philadelphia area. Pedagogical techniques were discussed and demonstrated to assess students. Formative assessment involves lecturing based on the current understanding of the students, learned through these assessments, with various alternatives available depending on the level of student comprehension. Through this seminar, we were encouraged to try out assessment techniques in our classrooms.

In addition to this seminar, I attended a lecture from Peggy Baker of the Young Women's Leadership Charter School of Chicago (www.ywlcs.org). The YWLCS is a successful charter school that uses innovative assessment techniques. (Farrington and Small, 2006) Course outcomes are defined for each class and students are evaluated using a three-tiered system – high performance, proficient, or not yet proficient – rather than traditional numeric letter grades. Students who are not yet proficient in a subject matter are allowed to demonstrate proficiency at a later date. Criteria are set for grade advancement such that freshman need 70% proficiency to advance to sophomore year, sophomores need 75% proficiency (in both freshman and sophomore outcomes) to advanced to junior year, juniors need 80% proficiency to advance to senior year, and seniors need 85% proficiency to graduate. In 2005, YWLCS graduated 43 seniors, 100% of whom were accepted to college. (www.ywlcs.org, accessed on March 25, 2007).

METHODS

My research set out to evaluate the effectiveness of assessment methods similar to that done at the YWLCS in the college setting.

In the Fall 2006, I taught two sections of the introductory statistics class at Villanova University. Classes were held back-to-back for seventy-five minutes each on Monday and Wednesday afternoons. In both classes, a set of eighteen course objectives were handed out at the beginning of the semester along with the course syllabus. The control class was taught using traditional assessment techniques, including homework assignments due weekly where numeric scores were used to evaluate the students and midterm exams using the same traditional evaluation scheme. In the experimental class, homework assignments and quizzes were due weekly and no exams were given. Evaluation followed the YWLCS model where *high performance* (*HP*), *proficient* (*P*), or *not proficient* (*NP*) scores were given on each course objective, typically two per assignments, and separately for homework assignments and quizzes. Students who received a score of *not proficient* on any course objective were allowed to resubmit related sections of their homework assignment the following week or retake a quiz with similar material a week later to improve their grade to *proficient*. A score of *high performance* was only offered to students with a flawless (or close to it) assignment that was handed in on time, to offer incentives for students who completed assignments on time. These differences are summarized in table 1 (below).

Table 1
Comparison of teaching methods between classes

	Control	Experimental
Class Time	Mon/Wed 3:00-4:15pm	Mon/Wed 1:30-2:45pm
Instructor	Posner	Posner
Grading	Numeric Scores	Proficiency (3 categories)
Weekly Homework	Yes	Yes
Weekly Quizzes	No	Yes
Two Exams	Yes	No
Resubmissions Allowed	No	Yes
Final Exam	Common Exam	
Pre-class Survey	Yes	Yes
Post-class Survey	Yes	Yes
Follow-up Interviews	No	Yes

Statistical design of experiment theory requires randomization of the groups. This would require students to be randomized into a section of the course. Academic freedom prohibits this from occurring in the college setting. This restriction forces limitations on educational interventions. Therefore, in addition to a primary analysis of all students, analysis of only first year students will be performed as well. First year students are put into classes by the University using methods that, we assume, are as random as we can come by.

Informed consent was obtained from students from both classes on the first day of class (this offered a nice teaching opportunity to discuss the purpose of informed consent forms and research design). All students agreed to participate in the study. Institutional Review Board approval was obtained to gather data from the registrar on past experiences, student grades in other courses during the semester, as well as some follow-up data.

Surveys were given to the students at the beginning of the course and follow-up surveys were given at the completion of the course. Demographic information gathered included gender, race, and class year. Also gathered was expected major and minor, number of math courses currently taken, and attitudes towards mathematics. Post-intervention surveys included additional demographic questions including high school experiences, attitudes towards mathematics, time spent on the course, number of office hours attended, highlights and lowlights from the course, and evaluation of the assessment.

In addition to evaluating students on the proficient/non-proficient/high performance scale, students were asked to evaluate their own work on this scale as well.

A common final exam was given in both classes. The primary outcome was score on this exam. Additional outcomes included (anonymous) course evaluations and survey data. Future outcomes will include grades in subsequent mathematics classes and whether this class had an impact in later college and professional decisions.

In the experimental class, student self-evaluations will be compared to professor evaluations on each course objective. In addition, some survey outcomes were targeted specifically to the experimental class.

Prior to the study being designed, a power calculation was done. A full letter grade (10 point) increase on the final exam would clearly show effectiveness of the method. A standard deviation of exam scores was estimated at 14 points from two sections of the same course taught the previous year. Originally, a second instructor was supposed to participate in the intervention, resulting in 40 freshman students in each arm of the study. These numbers resulted in a 88% power to detect such a difference. However, in reducing this to only my class with an estimated 15 freshman per class, the power dropped to an inadequate 47%. Further data gathering is in the planning stages to increase the sample size.

RESULTS

Table 2 presents information on the 47 students in the two classes (26 in the experimental class and 21 in the control class). One student from the control group dropped the class and was excluded from all analyses. Percentages represent non-missing responses to questions.

Table 2
Demographics of students

	Control	Experimental
Number of Students	21 (45%)	26 (55%)
Female	14 (67%)	16 (62%)
Class Year		
First Year	14 (67%)	13 (50%)
Sophomore	1 (5%)	7 (27%)
Junior	1 (5%)	4 (15%)
Senior	4 (19%)	2 (8%)
Part-Time	1 (5%)	0 (0%)
Caucasian	16 (76%)	17 (65%)
Taking other Math this year	14 (67%)	16 (62%)
Enjoy Math by Hand	8 (38%)	11 (42%)
Like Math Formulas	10 (48%)	10 (38%)
Understand the Need for Math	4 (19%)	10 (38%)

The primary outcome was final exam score. The experimental class scored an average of 74 on the final exam while the control class scored 83 (not significant, and in opposite direction than expected). Among the first year students, the mean scores were 72 and 81, respectively, again not significant and in the opposite direction than expected.

A second outcome included the subjective experience of the grading methods. Four students (15% of the experimental class) specifically listed the grading scheme as a highlight of the class. 24 experimental students (92%) said they were strongly satisfied with the new grading system of getting HP, P, or NP instead of letter grades (1 student (4%) said s/he was moderately satisfied and 1 student (4%) didn't answer the question). 75% of students in the control class said there were satisfied that they could work at their own pace while 96% of experimental students said they were satisfied with this (p=0.07). 80% of experimental students vs. 55% of control students said they were very satisfied with the fact that were able to ask the professor for help (p=0.07). 92% of the experimental group said there were satisfied with the fact that they were involved in their own assessment. The experimental class was more likely (Jonkhere-Terpstra

p=0.04) than the control class to say that they were satisfied with the interactive support via the internet, although it is unclear why this was the case since no additional information was posted on the internet for these students.

Some differences were found on course evaluations. Course evaluations are ordinal scales from 1 to 5, with 5 being the highest score. Students in the experimental group thought the goals of the class were clearer (4.9 average compared to 4.7 in the control group, not significant), perhaps due to frequent referencing to course objectives. They also believed the instructor interacted better with students (4.87 vs. 4.56, p=0.08), that the grading was more fair (4.57 vs. 4.00, p=0.07), and felt like they learned more (4.13 vs. 3.67, p=0.12). Students felt that explanations were clearer in the control class (4.72 vs. 4.48), which happened second. From this, we can see that it may be the clearer explanations in the subsequent class that may have led to students in this class performing better on the final exam. One would expect that this would bias the result above that students in the experimental class felt like they learned more towards the null. Thus, in reality this effect might be stronger once this bias is taken into account. Selected differences between groups are summarized in Table 3.

Table 3
Selected Differences Between Groups

	Control	Experiment	p-value
Liked Working at Own Pace	75%	96%	>0.1
Could Ask Prof for Help	55%	80%	0.07
Goals of Course were Clear	4.7	4.9	>0.1
Prof Interacted Well with Students	4.6	4.9	0.08
Grading was Fair	4.0	4.6	0.07
Felt Like They Learned a Lot	3.7	4.1	>0.1
Explanations were Clear	4.7	4.5	>0.01

DISCUSSION

Prior to beginning this intervention, I expected to encounter resistance from students who were comfortable at navigating established systems of student evaluation. To my surprise and pleasure, students embraced this new method. I overheard discussions by students using the language of *NP*s, *P*s, and *HP*s and retaking quizzes and homework assignments. I noticed that some students from the experimental section seemed to be attending office hours more than those in the control section, in particular within the time period when they could resubmit this homework assignments.

Overall, students in the control group outperformed students in the experimental group, although the differences were not statistically significant. Students in the experimental group engaged with the new evaluation methods and seemed to appreciate them. The most striking difference between groups on the evaluations was that students in the experimental group felt like they could ask the professor for help and that the grading was fair.

Next fall, I will continue data gathering. I will be teaching the same two class times and will switch the order of which class is control, so as to wash out the effect of different class times, according to statistical theory, assuming no year-class time interaction. In addition, both classes will be offered homework assignments and exams, rather than separate grading schemes where one class was offered quizzes. Lastly, through the use of online computer software, I will be offering extended resubmission of homework assignments for the experimental class that does not limit students to a one week time period to resubmit. In future work, I hope to include additional faculty at Villanova and at other universities in the research.

REFERENCES

http://www.ywlcs.org/toppages/ywlcsglance.html, March 25, 2007

Farrington, C. H. & Small, M. H. (2006). *Removing Structural Barriers to Academic Achievement in High Schools: An Innovative Model*. American Educational Research Association Annual Meeting.

William, D. & Black, P. (2006). *Inside the Black Box: Raising Standards Through Classroom Assessment.* NFER Nelson Publishing Co Ltd.