MAXIMIZING FINAL EXAM SCORES IN QUANTITATIVE COURSES

RANDY J. ANDERSON
CALIFORNIA STATE UNIVERSITY, FRESNO

ABSTRACT

College professors teaching quantitative subjects have the opportunity to favorably influence the evaluation outcomes of their students. This can be accomplished by changing testing parameters including testing format, time allotment, and grading method. The thesis of this paper is to challenge the traditional, established approaches to testing and increasing scores for college students in quantitative courses.

I. PROLOGUE

With the increased usage of microcomputers, Internet access, and online classes, the mode of college teaching is undergoing a metamorphosis. The classroom is being stretched beyond its ordinary walls, and students are starting to participate from a far. It is a new day for college instruction, as well. College instructors are no longer constrained to the technology of overhead projectors and chalk dust to expand the horizons of today's students. But how this transformation is made depends upon the vision of the instructor. One of the major concerns that college instructors face is determining the proper testing format for their students with respect to a particular teaching approach. Pure online instruction would dictate online testing. Pure classroom instruction would dictate classroom testing. A hybrid approach could involve a combination of both types of testing. One thing is for sure, any testing format has the potential of suffering the loss of security, which in turn, could compromise the results. Critics of online classes point to this very problem (Colwell and Jenks, 2005). However, as most college instructors know, the same exact problem exists in the traditional classroom setting (Clabaugh and Rozycki, 2005).

Another concern for instructors is how to exhort students to perform at the peak of their abilities on an exam regardless of the mode of instruction. This circumstance can be controlled to some extent by the instructor. If the most advantageous testing format for a particular subject matter can be determined, the resulting outcome could be considered to be maximized, the best of all possible outcomes. Furthermore, if the daily learning and testing environment could be adjusted to fit the best possible outcome on a comprehensive final exam, then the instructor could subtly manipulate the testing format throughout the semester in preparation to achieve the greatest positive outcome on the final exam. This approach would not only be worthy of note on behalf of the instructor, but it certainly would be greatly appreciated by all of the students involved. Some may effectively argue that "testing scores" may not always be indicative of learning, as some students are gifted test-takers but fall woefully short when it comes to knowledge. However, in most if not all quantitative classes where test questions require exact answers derived from a quantitative process, it is less likely that a skilled test-taker would have any real advantage over those students who possess the requisite subject knowledge.

In 1982, an investigatory study of different pedagogical testing approaches was begun. In narrowing the research topic, the thesis of this study was to determine the best instructional approach which would lead to maximizing the student's outcome on a comprehensive final exam for an inferential statistics course. This study has continuously collected testing data from over fifty semesters utilizing a 100 question multiple choice/true false final exam (whose content has, for the most part, remained unchanged) as the control. Although many of the problems scenarios have been continually changed and updated during this study, the final exam has consistently tested the same definitions, concepts, and quantitative methods representing the requisite knowledge needed in inferential statistics. Due to the rather static nature of statistics and stochastic processes, the information, approaches and equations contained in the textbook from which the author used as a graduate student thirty plus years ago (Parsons, 1974) are still valid today. It is more likely that statistical textbooks fall out of favor with their audiences not due to change in its content but rather due to publishers' marketing efforts and the four-color approach of newer books. It is for this reason alone that the final exam administered since 1982 has remained a valid control in this study.

All of the unit exams in the author's statistics classes are unique, with a new exam being written for each statistical unit during each semester. Not only do the unit exams continually exemplify the strong statistical principals being taught, they represent a broad spectrum of everyday applications and scenarios, including business, medicine, sports, leisure activities, and education. In trying to create exams that are interesting, current, and even humorous, the author has attempted to tear down the self-constructed walls of most students' contempt and dislike of statistics. Moreover, past unit exams have been continuously made available as an added resource and study guide for the author's students (at absolutely no monetary gain to the author). If student evaluations are any basis for the students' likes and dislikes, the author has consistently received high marks for his attempt to bring daily applicable scenarios to light with his exam problems.

The accountability and disposal of the final exam over the twenty-five plus years has been unblemished. Due the somewhat obsessive compulsive nature of the author, every exam has been numbered and has been crosschecked with each student's name, both on the exam and the answer sheet. Moreover, all of the exams are first counted and numbered when reproduced and then recounted after the final exam testing period is completed to insure that all exams have been returned. Once the prescribed archival time to keep exams readily available for student viewing has passed, the exams are shredded at the author's home and the discard recycled.

Different testing approaches for the unit exams and the comprehensive final exam as well as different lecturing methods and different topic introductions were utilized. Even classes sizes were increased (mandated by economic constraints) during the experiment. The authentication in deciding which lecture method, which order to introduce the statistical topics, and which testing methodology was the best approach was, and still is, the average final exam score for the students.

II. INTRODUCTION

How can college instructors teaching quantitative subjects best test students in any classroom (real or virtual) to achieve maximum results? In order to attain fairness, ethical considerations must be considered. Otherwise, the value of the testing instrument could become compromised, and the results could be biased and contaminated. The well known and former testing procedure at the United States military academies was the best example of honorable and ethical testing procedures (Peterson, 1984). The exams were placed at the front of the classroom at the beginning of the day. The students entered the classroom on time, took their exams to their seats, and began the test. At the end of the prescribed time, these students then placed their exams on the front desk and left the classroom. There was no talking, no dawdling, and no cheating. This procedure was fair and honest. The students were held to a certain code of conduct (this was to be followed to the letter), so-called “put on their honor.” They would not cheat nor would they tolerate anyone who did. In a perfect world, this procedure would still be a viable testing technique, inside or outside of the classroom. But even at the military academies, the temptations were too great and corruption followed. The cheating scandals are a matter of public record. Moreover, the widespread use of text messaging and camera cell phones presents even great temptations to cheat for today’s college students.

But in this imperfect world, how can educational professionals react to the pressure and technological conditions, which exist today? What if a testing procedure could be developed that showcases the student’s best work? What if a student could update his/her knowledge during an examination, without this update conflicting with ethics and honor? What if you could change the probability of an event occurring, while the event was still happening? In statistics, this concept is called a “Bayesian” approach to probability (Parsons, 1974). Bayesian statisticians believe that you can update your probabilities based upon prior knowledge gathered during an event for future applications. This is much like riding a “hot streak” while playing blackjack or playing craps. Even though the House will probably win in the end, an individual player can win during the short term.

How can instructors use this philosophy to challenge students, to alter their outcomes on exams and increase their knowledge base? Success in the workplace is not usually based upon a solitary exam, but is based upon a compilation of many “exams.” Workers are allowed to continually update their work, to change when more information is available that could affect the final outcome. Consequently, how can educational professionals use this same approach to adequately assess a student’s potential? Treatment “C” (see Table 1) is an attempt at a “Bayesian” approach to testing.

III. RESEARCH

TABLE 1: HISTORICAL TESTING OUTCOMES IN INFERENTIAL STATISTICS COURSES

Time Frame Historical Data	Semester Exam Format	Average Unit Exam Score	Final Exam Format	Average Class Size	Average Final Exam Score	I.D.
36 semesters: Fall 1982-Fall 1997	100% subjectively graded and no multiple choice format	84	Strictly Multiple choice, 2 hours total, one sitting	34	65	A
3 semesters: Spring 1998-Spring 1999	Combination: 75% subjectively graded and 25% multiple choice format	82	Strictly Multiple choice, 2 hours total, one sitting	43	69	B
3 semesters: Fall 1999-Fall 2000	100% subjectively graded and no multiple choice format	83	Strictly Multiple choice, 1 hour preview and one 2 hour sitting for final: 3 hours total	43	81	C
1 semester: Spring 2000	Combination: 10% subjectively graded and 90% multiple choice format	80	Strictly Multiple choice, one 2 hour sitting for final (in class)	35	74	D
2 semesters: Fall 2001-Spring 2001	Combination: 10% subjectively graded and 90% multiple choice format	79	Strictly Multiple choice, one two hour sitting for final (in class)	35	76	E

TABLE 1 (CONTINUED): HISTORICAL TESTING OUTCOMES IN INFERENTIAL STATISTICS COURSES

Time Frame Historical Data	Semester Exam Format	Average Unit Exam Score	Final Exam Format	Average Class Size	Average Final Exam Score	I.D.
2 semesters: Summer 2003-Summer 2004 Six week classes	Combination: 10% subjectively graded and 90% multiple choice format	80	Strictly Multiple choice, one two hour sitting for final (in class)	28	75	F
1 semester: Fall 2004, large classroom instruction format	Combination: 10% subjectively graded and 90% multiple choice format	76	Strictly Multiple choice, one two hour sitting for final (in class)	41	69	G
Two semesters: Fall 2006-Spring 2007	On Line Testing for two exams; in class testing for the two other exams: 100% multiple choice	78	Strictly Multiple choice, one two hour sitting for final (in class)	32	62	H
Summer 2007 (five week session)	100% multiple choice	83	Strictly Multiple choice, one two hour sitting for final (in class)	32	80	I
Fall 2007	100% multiple choice	79	Strictly Multiple choice, one two hour sitting for final (in class)	40	67	J

TABLE 2: COMPARATIVE P-VALUES (IN PERCENT)

A	A									
B	7.05	B								
C	0.11	0.11	C							
D	2.31	2.12	3.03	D						
E	2.42	2.05	4.13	9.78	E					
F	2.00	2.02	1.35	10.33	10.12	F				
G	9.46	12.30	0.37	2.13	3.52	1.89	G			
H	8.47	4.37	0.09	1.03	1.96	1.00	4.27	H		
I	0.09	1.01	11.85	3.61	5.19	5.23	1.36	0.06	I	
J	11.10	10.70	0.99	6.98	3.00	1.85	11.60	6.68	0.30	J

In developing Table 2, a two sample test of means (using the normal distribution) was used. If a hypothesis test was being conducted, the null hypothesis would be that the means of these two treatments are statistically the same. The alternate hypothesis would be that one of the means is statistically greater than the other. Once the test statistic is calculated, this number is then generally translated into a "p-value," using the "NORMSDIST (Z)" function in EXCEL. A p-value is simply the remaining amount of tail area left in a probability distribution with respect to the test statistic. If the test statistic (the difference between the two mean values) is small, then, the p-value is large meaning there is no significant difference in the outcome of the two treatments being compared. Consequently, final exam scores from the two treatments are statistically the same. If the test statistic is large, then the p-value is small, meaning that the outcome of the two treatments is significantly different. This means that one of the treatments results in significantly higher test scores on the final exam and can easily be determined by checking which mean value is greater. Moreover, it can be inferred that this particular treatment results in better retention of material and more knowledge gained during the semester.

The ten different treatments (labeled "I.D." in Table 1) have been lettered from "A" to "J" for easy identification in Table 1. Each treatment was compared to all of the other treatments and the resulting p-values placed in Table 2. For example, when treatment "G" (10% subjectively graded unit exams with 90% objectively graded unit exams with a final exam which is strictly multiple choice in one two hour setting) is compared to treatment "B" (75% subjectively graded unit exams with 25% objectively graded unit exams with a final exam which is strictly multiple choice in one two hour setting), the corresponding p-value is 12.30%. The conclusion is that neither of these two treatments is distinguishable from each other. The results from either exam would be statistically the same.

In a second example, when "I" is compared to "A," the resulting p-value is 0.09%. This is an extremely small p-value, which can be interpreted that one of the treatments demonstrated a significantly higher final exam score that the other. By looking at the average final exam test scores from Table 1, we can see that "I" from the summer of 2007 has an average final exam score of 80.39% and "A" has an average final exam score of 65%. We can conclude that treatment "I" results in significantly higher final exam scores. The remainder of Table 2 can be interpreted in the same fashion.

IV. METHODOLOGY

Any comprehensive Final Exam could be and should be considered the culminating experience in a college course and used to assess the knowledge gained during the entire course. Such is the case in DS 123, the second semester, senior level statistics class, in the Sid Craig School of Business at California State University, Fresno (CSU, Fresno). This course covers inferential statistics, the application part of business statistics. Historically, when broken into four introductory components or units and tested on these four units separately, DS 123 students have performed quite well. But when asked to answer 100 multiple choice/true false questions on a comprehensive final exam within the allotted timeframe, the outcome has been quite different. In quantitative subjects, this comprehensive exam requires the student to piece together the entire giant puzzle for the whole semester. What this type of exam is really testing is the ability of the student to synthesize the complete material and its complexities in one sitting, rather than being tested piecemeal.

Because the content and concepts on the DS 123 comprehensive final exam have not changed over the past twenty-five years (fifty plus semesters, including summer sessions), the expected value of the outcome (the historical average) on this final exam has been tracked and validated to be 65%, a middle “D” grade. Moreover, the standard deviation of this average has been reasonably small (5.13%), which means that students have historically averaged about the same scores on this exam. In other words, there have not been bimodal outcomes where one group of students have done very well on the final exam countered by a similarly large number who have performed just as poorly.

As can be seen in Table 1, students have not fared well on this final exam. Many reasons could be hypothesized as to why there is such a lowly outcome. One rationale could be that the subject matter is difficult in general and even more difficult when tested in a comprehensive fashion. Or, perhaps, by the end of the semester, most students have lost interest in a class that is perceived by some as just an impediment in their quest for a four-year college degree. Or, the students are just plain tired. All of these justifications could be valid. However, the explanation of “WHY” students scored what they did is not the main thesis of this paper, but rather “HOW” to improve the students’ outcome on the final exam without compromising the exam itself.

As the university classroom is increasingly moving towards the faceless virtual classroom, the necessity of “online” delivery of students’ performance evaluations (testing) is becoming a reality. Once the best face-to-face classroom evaluation method is determined, then its online counterpart can be developed into a reasonable parallel evaluation tool to successfully monitor the students’ progress in a particular college course, especially if quantitative in nature.

One concern is the fact that in many quantitative related college classes, students have been evaluated subjectively during the regular unit exams (where partial credit can be and is generally given) and then they are asked to perform at the same level of competence on a final exam which is completely objectively graded (True/False and/or multiple choice). It could be argued that, since the format of an exam can have a direct bearing on the outcome, a change in this format could change (either enhance or detract) the outcome of the final exam without affecting its intrinsic value.

Consequently, ten years ago (see Table 1: Spring 1998 "B"), the format of the regular unit exams was changed to incorporate a multiple-choice component. While the first two unit exams of the semester were subjectively graded, the last two exams were converted into a multiple-choice approach. Translating equivalent objective questions from their subjective counterparts for inferential statistics was a challenge. To elicit similar knowledge and expected outcomes, questions needed to be phrased in such a fashion that students still needed to effectively work the problems. But a mistake at any juncture of working a problem containing ten questions on an objectively graded exam would compound the error by causing further incorrect answers on subsequent questions (despite the fact that the procedure was being followed correctly). As a result, the questions on the objectively graded portion of these exams were restructured to reflect "what if" scenarios which did not specifically rely on the prior question's outcome to determine the current question's answer. During the Spring 2000 (treatment "D") semester, almost all portions of the four unit exams had been converted to a multiple-choice format, leaving a small portion (one to two problems per exam) to be subjectively graded. Since this time, the four unit exams during the regular semester have been a mixture of subjectively graded and objectively graded (as indicated in Table 1).

As shown, the outcome on the final exam can be improved. It is possible to conclude that prior knowledge of the final exam format has increased scores. The final exam average score jumped from 65% (Table 1, treatment "A") to 69% (Table 1, treatment "B"). The respective p-value (7.05%) shows that this change is not very significant in a statistical sense. What is important is that the average student was still earning a less than stellar mid-to-high "D" on a comprehensive final. Consequently, in order to help students perform better on exams which in turn hopefully demonstrates mastery of the subject matter, more modifications to both the unit exams and final exam were considered.

In the fall of 1999 (Table 1, treatment "C"), a newly devised approach for the final exam was tested with heralded success. Prior to this semester, the final exam was given in its entirety during a one sitting, two-hour exam. Because the average student's poor performance on the final exam was still a great concern to the professor, a novel approach to administering the same final exam was brought to the attention of the department chair and school dean for approval. One thought that guided this change was that, since most employees in the workforce very rarely complete a task in one sitting, the students might benefit from the same approach to problem solving at the collegiate level.

Students were given the opportunity to take the same final exam in three one-hour sittings (Monday-Wednesday-Friday) with a day off in between sittings. This "Bayesian" approach to problem solving would allow students to view the exam in its entirety the first day in order to contemplate its level of complexity and completeness.

Students then could spend time between testing sessions focusing on the topical areas in which they were weak or unfamiliar and study where needed. When the test reconvened, students then could apply their rededicated knowledge into improving their outcomes.

With all other things being equal (like a normally distributed student population and all biorhythms peaking at the same time), this approach garnered extremely positive results. The average score on this exam jumped to an incredible 81%. When viewing the respective p-values of this treatment with the others, the results of treatment “C” are statistically significant and better than all of the other treatments except for the summer of 2007 (“I”). In a statistical sense, this increase in average score in treatment “C” over the past and future final exam experiences was significant. Furthermore, it is felt that the final exam itself was not compromised, but the ability for the average student to demonstrate his/her knowledge and ability was showcased to a greater extent, resulting in higher final exam scores.

This new final exam testing procedure showed great promise as students were scoring higher, which in turn was reflected in their final grades for DS 123. Unfortunately, it took only one student in one class during the Spring 2000 (treatment “D”) semester to change the author’s approach for the Bayesian testing procedure. Due to the sophistication of programmable calculators, this student was able to input the entire final exam during a preview period on his calculator, which could have invalidated the results of the final exam for all students. Fortunately for the author, this student was generous to a fault as he shared the exam’s problems with other students by publishing his ill-gotten gain on the internet. Luckily, one brave and honest student brought this incident to the professor’s attention and it was quickly rectified. The professor rewrote the entire final exam overnight. New problems addressing the same statistical concepts were given to the students the next day, much to the chagrin of those involved in the cheating scandal. This unpleasant attempt by just one student has increased the author’s vigilance to keep the final exam secure. It seems that the professor was too trusting and did not take into consideration the temptation for students to cheat (which unfortunately happens), much like at the United States military academies. The questions were posted on the internet for any obliging student to copy and complete outside of class. This sad occurrence not only compromised this exam, but it invalidated this Bayesian testing procedure and the professor’s trusting nature. Consequently, the professor reverted back to the normal final exam approach of one sitting of two hours.

Since the return to the two hour, one sitting final exam evaluation, more variables have been introduced into the analysis. At CSU, Fresno, one of the main attractions for potential students is the small classroom atmosphere. The author taught a six-week summer school session in both the 2003 and 2004 summers (treatment “F”) with an average of about 25 students per class meeting daily for 85 minutes. The compressed teaching schedule resulted in a respectable final exam average of 75%. The corresponding p-values illustrate that this format (meeting daily) has great promise when trying to increase comprehension and final exam scores. However, due to the 2004 California state budget constraints, the classroom size was stretched to full capacity. Consequently, students now fill a 50 seat classroom as compared to 35 seats in the past (treatments “G.” “H,” “J”) and, unfortunately, final exam scores have returned to mediocrity.

Most recently, internet availability for college classes has become a reality through the use of Blackboard (abbreviated, “Bb”). Not only can instructors post lectures and ancillary material online for the students’ benefit, instructors can test students as well. The posting of material for DS 123 has been a much appreciated resource for students. However, the testing of DS 123 students using Blackboard was an abject failure. For two semesters, starting in the Fall 2006 semester (treatment “H”), the first two unit exams were given online. Many different reasons might explain the failure of students to perform at historical averages, even on the easier unit exams (the mode of the exam, asking student to perform calculations without more difficult formulae given, the Bb system disallowing for multiple entries, lost internet connections, students’ unfamiliarity with online testing, etc.). The fact is that the average final exam grade (61%) was the lowest in the entire experiment has made abandoning of Bb testing an easy choice.

One last modification was implemented during the summer of 2007 (treatment “I”). Summer sessions are usually scheduled to meet five times a week for six weeks. The professor, with the students’ approval, decided to fast track the DS123 class. The students agreed to stay for a longer period of time (the number of semester minutes for this class was not compromised or diminished) for four days a week for four weeks. The comprehensive final exam was administered at the start of the fifth week with a two hour sitting. The outcome from this approach was overwhelmingly positive. Perhaps it was the caliber of students during the abbreviated summer session or perhaps it was this particular treatment. But the results speak for themselves. More time in a classroom on consecutive days resulted in significantly higher final exam scores.

During the Fall 2007 semester (treatment “J”), the DS 123 classes were again structured with strictly objectively graded exams, meeting twice a week for 75 minutes each time for a normal fifteen week semester. Unfortunately, the results of the final exam reverted back to the norm, with the students averaging in middle “D” grade range. Obviously, as demonstrated in both summer school approaches (treatments “F” and “I”), the most promising approach (barring the Bayesian trial, treatment “C”) is for students to spend more time in the classroom on consecutive days (still meeting the entire 2250 minutes per semester). Moreover, the p-values support this claim. Being creative in the classroom should not be limited to any one department on a college campus. It should be, however, the challenge for every college instructor.

V. CONCLUSION

Final exam scores have improved over the historical average, in spite of the changes in testing treatments. However, the Fall 2004 (treatment “G”) semester’s outcome on the final exam (with its larger classroom sizes) seems to suggest that students have reverted back to the historical average. It might be easy to suggest that both the format of the semester unit exams and the innovative final exam testing approach had a positive effect on the outcome. But an innovative approach during the Summer 2007 (treatment “I”) session with the usual two hour, one sitting final exam trial was extremely successful and should be fully explored by repeating treatment “I.” Although this summer class was a pilot study, the fast track compressed meeting schedule offers very promising results.

The thesis of this paper is to challenge the traditional, established approaches to testing and increasing scores for college students in quantitative courses. The very nature of the environment of today’s educational system is trending toward classrooms without walls. Compressed scheduling is already being used by many junior colleges. And, when you combine this with an online approach, the results will hopefully be very respectable. Not too far into the distant future will students no longer congregate into classrooms to be lectured and tested. Distance education and satellite campuses will require professors to react globally to educational circumstances. Certainly, ethics and honor must be considered on behalf of both the students and the professor when testing in a non-traditional setting. Testing processes can be experimental but should always be a valid evaluation of a student’s knowledge and ability. The “WHY” of testing students will be debated, deliberated, and dissected for as long as the process exists. The “HOW” of testing students is only limited to the imagination of those who design the testing procedures themselves.

REFERENCES

Clabaugh, Gary K., and Edward G. Rozycki, comps. Cheating Trends. Vers. 2nd Edition. 25 Sept. 2005. 14 Feb. 2008 <http://www.newfoundations.com/PREVPLAGWEB/CheatingTrends1.html>.

Colwell, J.L., and C.F. Jenks. "Student Ethics in Online Courses." Frontiers in Education, Proceedings 35th Annual Conference FIE (2005): t2d17-t2d19.

Parsons, Robert. Statistical Analysis: A Decision-Making Approach. New York: Harper & Row, 1974. 3-6,141-152.

Peterson, Iver. "Cheating Prompts Air Force to Halt Honor Boards." New York Times 20 Sept. 1984, Late City Final Edition ed., sec. A: 1-9. New York Times Online. 14 Feb. 2008 <http://select.nytimes.com/gst/abstract.html?res=F30E13FB3C5F0C738EDDA00894DC484D81>.