® ASSESSING ASSESSMENT: A FORMAL STUDY OF THE BENEFITS OF ASSESSMENT IN A FINAL YEAR UNDERGRADUATE STATISTICS COURSE

MURPHY, Patrick
University College
Ireland

Students learn by doing: unfortunately many only do as little as is required of them to meet formal assessment requirements. We examine how a radical change in assessment strategy was chosen as a method of improving the learning experience for a group of students engaged in a final year undergraduate course in Time Series. In 2006, following attempts to change student attitudes through the use of different teaching techniques, we decided instead to focus on altering our assessment strategies. This paper will show that assessment appears to be one of the most effective tools to improve learning. While we remain committed to investigating and implementing improved methods for course delivery, our recent experience indicates the need to complement those methods with innovative assessment to improve the whole learning experience.

INTRODUCTION

What do we mean by assessment? Is its purpose to measure how well our students have mastered the topics we chose to teach them? Many educators and most students consider its purpose as *assessment of learning*. Our aim in this paper is discuss the role of *assessment for learning*: to examine how assessment itself can be used as a tool for improving learning.

HISTORY AND CONTEXT

The last decade has brought increasing wealth in Ireland which, in 2007, places it in the top five wealthiest nations in the world as measured by GDP per capita. This new-found wealth has brought many benefits, but it has also brought distractions.

Students in the 1980s were forced to work hard at their studies in order to achieve any prospect of employment: at that time only the best graduates could be assured of attaining a position upon graduation. This can be contrasted with the situation today where students can expect to find full-time employment, almost irrespective of the grade of their degree. Indeed, most students actually hold part-time employment during their years at a university and see no conflict with this employment and their role as *full-time* university students. The purpose of the part-time employment, they indicate, is to enable them to have a higher quality social life while fully enjoying the *"college experience.."* In this context, education becomes an increasingly difficult challenge.

For several years I have devoted myself to improving the methods by which I instruct students in statistics. Autumns following ICOTS or JSM conferences have typically involved integration of new ideas into my courses; in addition I have tried to adopt new technologies early in order to aid exposition. My students receive instruction through three complementary methods: lectures, small tutorial classes, and computer based laboratory classes. *Activity-based learning* and *problem-based learning* have both been fully incorporated into all three of these environments.

I have placed considerable emphasis within my courses on deep rather than superficial learning; critical thinking philosophies have guided the development of all course material. My goal throughout is to facilitate the development of students into statistically literate and computer literate practitioners of statistics. In addition I continue to hope that some of my students will be transformed into *statistics zealots* ready to spread the word to a wider audience.

Despite considerable effort over the course of the last seven years I have found myself somewhat disappointed at the level of improvement in students' understanding of statistical concepts and in their ability to practice statistics. While new teaching methodologies have brought some improvement each year, it has been less than I had hoped. This led me to question whether there was a more effective tool available.

Education in Ireland is presently undergoing heavy restructuring. In particular, during the past two years, University College Dublin (UCD) has undergone major changes. While not all changes have been in the best interest of the students, one has been very positive and provided me with the new weapon that I sought for my arsenal in the battle to educate.

The Irish education system has traditionally always had more in common with the British system than the system in the United States. Consequently, students usually had one final exam at the completion of an academic year which accounted for one hundred percent of the marks for that course. It should be noted that under this traditional system a course conducted in the autumn might finish at Christmas but not be examined until the following summer. At University College Dublin, during the last two years, restrictions on the timing and the amount of continuous assessment have been eliminated.

RESEARCH QUESTION

The changes to the regulations at my university have provided me with an opportunity to formally examine the effect of different assessment techniques on the learning that occurs for students studying statistics in their final year of an undergraduate degree. I decided to investigate whether the introduction of regular, well designed, testing into a course can have direct benefits on students' learning. *To clarify, the aim is not just to use this testing as a measure of student performance but rather to examine how the testing itself can directly affect student learning.*

THEORETICAL BACKGROUND

"What you assess is what you get; if you don't test it you won't get it." These words from psychologist Lauren Resnick, first quoted in a seminal paper by Wiggins (1990), have appeared time and again in the education literature. Wiggins introduces the term "*authentic assessment*," urging teachers to place more emphasis on properly assessing student learning on "*worthy intellectual tasks.*" Inherent here is the idea that most students only learn what is directly required of them.

While the so-called "interested student" does actually exist, in most institutions educators realize that these students are a rare and endangered species. Much more common is the student who will only complete the minimum workload required of him or her. In the affluent Ireland of 2007, this student is unfortunately very prevalent. How can this attitudinal shift in students be combated?

Garfield (1994) discusses two fundamental principles that should underpin assessment in statistics courses:

- *"The Content Principle: Assessment should reflect the statistical content that is most important for students to learn.*
- *The Learning Principle*: *Assessment should enhance learning of statistics and support good instructional practice."*

While end of semester or end of year exams can validly achieve the goals of the Content Principle, it is very questionable as to whether they can *"enhance the learning of statistics."* Indeed Hubbard (1997) emphasizes that *"it is important to recognize that assessment determines not only what students learn but how they go about learning it."* She continues: *"Assessment drives the whole learning process. As a simple example, if the only assessment is at the end of the course, then students tend to defer the learning process until examination time draws near. By doing this, they may waste many hours listening to lectures and class discussions on topics about which they know very little."*

Clearly the literature indicates that the traditional end of year assessment procedure in Ireland could be improved upon.

AN EXPERIMENT TO ASSESS ASSESMENT

This section is divided into three subsections: the first will delineate the new assessment strategies that I chose to replace the existing single end of year examination. The second subsection will briefly describe the actual assessments used in my course on Time Series Analysis. Finally, I outline how I conducted a comparison between these new procedures and the previous single end of year examination approach.

Authentic Assessment

The motivation behind the decision to introduce new assessment strategies has been described above. What I should like to do here is describe the types of assessment that were chosen and why they were chosen. Many assessment choices are available to instructors, some requiring significant amounts of the instructor's time and some being very easy to deliver. My priority was that each component of my assessment should serve three purposes, namely:

- to genuinely assess students understanding of the concepts of Time Series Analysis,
- to clearly establish the ability of students to put these concepts into practice in analyzing real Time Series data,
- and finally, that the assessments should motivate students to increase the quantity and enhance the quality of their learning.

Garfield (1994) and Chance (1997) discuss the merits of several different assessment methods and recommend that new methods should be introduced in a gradual manner. Because I believed a radical shift in assessment strategies was required to change the level of motivation among my students, I decided to adopt a *big bang* approach and move from an end of year final exam to a multifaceted testing regime in one step.

The obvious and perhaps easiest approach to moving away from an end of year assessment would be to intersperse one or more multiple choice tests through the semester itself. My view on this approach is in agreement with Cobb (1993) who considers un-enhanced multiple choice tests to be a very unsatisfactory tool. Cobb instead recommends the use of projects which the students must complete over the course of a semester.

Clearly the workload involved for the instructor in designing and assessing projects is higher than that in a multiple choice test. Indeed the extra burden imposed upon the instructor by more creative and more authentic types of assessment conflicts directly with the aim of the instructor to use only the best assessment methodologies. This workload issue is one that demands serious consideration in advance of designing a new assessment regime. My conclusion was that instead of grading all of a students work we should instead grade a representative sample of their work. In addition to reducing the grading required, this approach brings other considerable benefits.

Consider the issue of homework assignments. I wished to allocate 15% of the total course mark to these assignments. However, one significant problem with homework assignments is the inverse relationship between the number of assignments in a semester and students' perceptions of the relative importance of each assignment. Students may be tempted to choose to skip an assignment on the basis that each one counts for such a small percentage of their final grade. Our solution to this problem consists of assigning 10 problem sets, but not grading all assignments. Instead at the conclusion of the semester a random selection of two assignments is chosen and graded. If students do not know a-priori which assessment will be graded, then they quickly realize that a potential 7.5% of the final course grade rests with each assignment, significantly increasing their level of engagement with each and every assignment.

Another approach to assessment that is commonplace in the USA, but which is not widely practiced in Ireland, is to grade in-class participation. In Ireland, students are used to coming to classes unprepared and expecting that it is their right to be spoon-fed by the instructor. This is not an educationally beneficial approach for students to adopt; more of an onus should be placed on students to take responsibility for their own learning. I hoped that by indicating that in-class participation would be graded, students' learning would be transformed from being a passive process to an active one.

The following table outlines the complete set of assessment tools as used in the 2006 delivery of the Time Series Analysis course at UCD.

Table1
Assessment Types Used in 2006 Time Series Course

Assessment Type	*Percentage of Final Grade*
In-Class Participation	10%
In-Class Tests	10%
Assigned Homework	15%
Project	15%
End of Semester Exam	50%

Description of Assessment Components

Students were assigned numbers at the beginning of the first class and were told to print these numbers in 400pt times-roman font on a sheet of A4 paper. Students brought these pages to class each day and when asking or answering a question they raised the page which allowed a note to be taken of their number. Clearly this low-technology approach could be significantly improved through the use of a proper electronic student response system or *"clicker."*

Three in-class tests were delivered during the course of the semester; two of these were of 20 minutes duration with the other one lasting for an hour. Each of these tests used an open rather than multiple choice format for questions. Two such questions were:

- Express a centered 4-MA smoother as a weighted MA smoother,
- Compute the first three auto-correlations for the series defined as

$$X_t = \varphi_1 X_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}.$$

Eight homework problem sets were assigned over the course of the semester and two were randomly chosen at the end of the semester for grading. Problem sets generally consisted of four or five problems and it was estimated that each problem set should require between one and two hours work by a student.

The final continuous assessment component involved a project being assigned to students in week six of the semester for completion by week twelve. Students were instructed to each find a time series containing between 100 and 200 observations. Students could gather their own data or could locate data on the internet; the only stipulation was that no two students were allowed to use the same data set. Three hours of laboratory classes were specifically devoted to guiding the students on how to complete this project. Outside of these classes, students were not allowed to seek assistance from myself or from post-graduate students and were obliged to sign a declaration indicating that all work was solely their own. Students were asked to describe their data and their sources in detail. They then were required to set aside the final two observations in their data sets and to use any techniques they deemed appropriate to determine the *"best"* model for their data. Following this, they should produce one and two step ahead forecasts using their data which would be compared with the actual observations.

One week after the end of the semester students took a written final examination of two hours' duration. There was a temptation to make this examination shorter and less detailed than it had been when it was the sole form of assessment. However, I decided to persist with the precise format of earlier years. The examination consisted of six open format questions. Each question consisted of several parts and measured knowledge under one or more of the following categories:

- questions to measure students' understanding of definitions such as difference stationarity,
- questions to measure students' ability to manipulate equations such as to derive expressions for a partial autocorrelation for a given ARIMA process or to show how a 5x3 double MA smoother could be written as a weighted moving average smoother,
- questions to measure students' ability to seasonally decompose an actual time series or to compute sample statistics for a real time series,
- questions which test students' higher level understanding and their ability to combine different techniques from the course to address a real life situation.

Experimental Design

The Time Series Analysis course was delivered in 2005 to a class of 86 students and to 88 students in 2006. There were five separate groups of students that took the Time Series course together. The first group was studying for degrees in Statistics andwere in their third year of taking courses in Statistics. The second group was studying for degrees in Actuarial and Financial Studies; these students had also taken three years of statistics, but not to the extent of the first group. However, this degree program has one of the most competitive entry requirements of any in Ireland, and so the students usually possess above average academic ability. The third group consisted of graduates from other disciplines (such as engineering or business) who were studying for a one year diploma in Actuarial Studies. The Fourth group included students who were taking a BSc in Mathematical Science; these students all studied honors level Mathematics and Mathematical Physics, as well as Statistics. Finally, there was a group of students studying Economics and Finance and took this course as an option in their degree program. This group generally had only one year of introductory level statistics.

The composition of this Time Series class did not change between 2005 and 2006, and the description above is appropriate for both years. No changes were made to teaching methods between the two years other than the changes in assessment described here. Consequently we may regard the 2005 class as the control group and the 2006 class as the treatment group for measuring the effect of the new assessment procedures.

RESULTS

We will describe the results of our experiment under two headings: students' perceptions and students' performance. As we shall see there were quite interesting results under both measures.

Students' Perceptions

In 2005, the students who took the course considered it to be difficult and challenging. They rarely interacted in class and, despite enormous efforts to stimulate their interest through the use of simulations and real data in class, from week four onwards most students had become bored and lost interest in the class. The disaffection that I experienced in this class was the prime motivator behind the following year's radical restructuring. It was clear to me that the main reason students were so bored was because they did not understand what was being taught in class. It was also clear that this was because they were not studying the course material and, as the weeks went by, they became more and more lost.

In 2006 students were forced to keep up with the course material or lose significant course marks because of the new assessment regime. Initial student perceptions were uniformly negative. They formed committees and made representations first to myself and then to course administrators, student advisors and finally to my Head of School declaiming what they perceived as being an entirely unfair workload. As the new UCD regulations permitted this new form of assessment, the students' protests made no progress.

Interaction in class, however, was the best that I have seen in over 15 years of teaching. And while, at first, this was most definitely influenced by students just wanting to gain marks, later in the term students were asking questions and not bothering to raise their numbers in search of credit. By the final four weeks of the semester the conclusion of each class would involve me spending half an hour fielding questions on course material.

The following quote from a student sums up the reversal of fortune that occurred: *"at the beginning of term I wanted to drop this course but now I am so glad I didn't, I learned more than I ever thought possible."*

Student's Performance

Table2
Students' Average Final Grades on 2005 and 2006 Courses

Group	2005 - Control Group Average Final Grade	2006 - Treatment Group Average Final Grade
Bachelor of Actuarial and Financial Studies	60%	75%
BSc Statistics	50%	62%
H Dip Actuarial Science	46%	53%
BSc Mathematical Science	66%	70%
BSc Economics and Finance	49%	68%

As can be seen from the above table, the introduction of new assessment strategies led to a measurable improvement in students' mastery of the course material. Comment appears to be superfluous; I will just add that an independent samples test of the difference between population means produced a P-Value of less than 0.001.

CONCLUSION

This paper describes the results of a study to examine the effect of assessment on students' learning in a Time Series course. Students' attitudes to the introduction of increased testing were initially extremely negative. However, testing forced students to change their study habits and they no longer postponed their learning until the end of the semester.

The type of testing used is also of the utmost importance. It must authentically simulate real world problems to be of any lasting benefit. O'Connell (2002) discusses students' perceptions of different assessment strategies in a multivariate statistics course. She found that students preferred assignments which were very structured over ones that required the students themselves to make decisions on the types of analyses to be performed. But, of course, it is this latter type of assessment which is most realistic.

My experience of using this type of assessment is that it causes frustration at first among students but, in overcoming this frustration and adequately completing such assessments, students eventually achieve a deeper mastery of Statistics. The revisions that were made to assessment in this course have delivered highly significant results, and we would urge others to consider seeing assessment as an important tool for improving student learning.

REFERENCES

Chance, B. (1997). Experiences with Authentic Assessment Techniques in an Introductory Statistics Course. *Journal of Statistics Education, 5(3).* Retrieved January 28, 2007, from http://www.amstat.org/publications/jse/v5n3/chance.html

Cobb, G. (1993). Reconsidering Statistics Education: A National Science Foundation Conference. *Journal of Statistics Education, 1(1).* Retrieved January 28, 2007, from http://www.amstat.org/publications/jse/v1n1/cobb.html

Garfield, J. (1994). Beyond Testing and Grading: Using Assessment to Improve Student Learning. *Journal of Statistics Education, 2(1).* Retrieved January 28, 2007, from http://www.amstat.org/publications/jse/v2n1/garfield.html

Hubbard, R. (1997). Assessment and the Process of Learning Statistics. *Journal of Statistics Education, 5(1).* Retrieved January 28, 2007, from http://www.amstat.org/publications/jse/v5n1/hubbard.html

O'Connell, A. (2002). Student Perceptions of Assessment Strategies in a Multivariate Statistics Course. *Journal of Statistics Education, 10(1).* Retrieved January 28, 2007, from http://www.amstat.org/publications/jse/v10n1/oconnell.html

Wiggins, G. (1990). The Case for Authentic Assessment. *Practical Assessment, Research and Evaluation, 2(2).* Retrieved January 28, 2007, from http://pareonline.net/getvn.asp?v=2&n=2