ANALYSING DATA FROM A CLASS TAUGHT WITH CLICKERS

Enrique E. Alvarez
University of Connecticut, United States
ealvarez@stat.uconn.edu

During the fall 2005 the author taught a large introductory Statistics class at the University of Connecticut with the aid of a student personal response system (also known as "clickers"). Apart from its pedagogical value, the system has the secondary advantage of building vast day to day data on students' performance. This article presents an analysis of the data gathered from the course, together with conclusions and suggestions on how to collect and organize data for further educational studies.

INTRODUCTION

In the fall 2005 the author taught Introduction to Business Statistics. It is an introductory level course which typically has an enrolment of 800 students per term. The author taught one section of size 250. The class consisted of two components, lectures given by the instructor for the whole class and small discussion sections (of 25 students each) given by graduate student instructors. In order to promote participation and interaction in lecture sessions, it was decided to adopt a personal response system (also known as "clickers") in the classroom. A clicker is an electronic device of the size of a TV remote control capable of transmitting radio frequency signals. They are mostly used in education to give pop-up questions during large lectures; that is a means of assessing students understanding "on the spot."

It is not the purpose of this paper to asses how good the clickers are as a pedagogical tool. Instead, we focus on a second aspect of teaching with clickers. That is how to analyze the collection of a large set of day-by-day data on students' performance. We are specifically interested in two main questions: (*i*) How are the performances of students in the in-class clicker questions related to their overall success in the course; and (*ii*) Is it possible to predict students performance in the final exam at some point throughout the semester, based on the data available at each time, in order to approach them when they still have chances to improve or withdraw? This second question is very related to a procedure that teaching institutes are pushing in many U.S. universities. That is the so called "*early warning* system." The idea is that professors are expected to detect when a student is having problems with the course so as to alert them before they get a fail grade in the class.

CLICKERS MANAGEMENT AND DATA COLLECTION

We describe next how the clickers were distributed and administered, as well as were data was collected.

1. During the first two weeks of classes each student was given a clicker to take back home. In order to make the clickers operational each student needed to be registered on a website. Students were asked to register their clicker as soon as possible. As they did, a list of students was created with their clicker pad number and university ID.
2. As an incentive for students to register their clickers and bring to class, it was decided they were going to be used to record attendance starting by the fourth week; course attendance represented 5% of the overall course grade.
3. Starting the fourth week of class the instructor would go to the classroom with a laptop computer and a clicker receiver to plug into the USB port. At some point during the lecture the professor would switch from a *PowerPoint* presentation to the *Personal Response* software. An attendance screen would appear on the big wall-screen and each student would mark his/her attendance by pressing on button in their respective clicker.
4. Immediately after taking attendance the instructor would put a few multiple choice questions on the screen for students to answer with their clicker. In order to incentive

students to try their best those in-class questions were worth 5% of the overall course grade.

5. Right after each question, a bar chart would appear on the class screen with the distribution of answers. The correct answer would be highlighted. That graph was useful to the instructor as well as to the students. For the instructor, that distribution would tell whether the topic was understood by students or whether additional explanation was needed. For the students, they would have a day by day assessment of their understanding.
6. The software automatically kept records for each student's attendance as well as their answer to each question, for every single class. Hence at the end of the term we had a large dataset with information on students' performance.

At the end of the term all the students records where compiled into large dataset. We present next the variables that we constructed along with a basic summary

1. `Sex:` Female (86), Male (101), or unknown (6).
2. `Level:` Freshman (103) , Sophomore (62), Junior (18), Senior (7), or Other (2).
3. `School:` ACES (75) Business (35) Liberal Sciences and Arts (LSA) (71), Other (12)
4. `TA` Graduate Student Instructor that the student had for the discussion section. There were three instructors that we renamed for confidentiality A, B and C
5. `Like` The first day of clickers we gave 3 questions in class to test the, the first was "How much do you like statistics?" and the possible answers were:
   A "I like it very much, I am in fact thinking of majoring in Statistics" (3%)
   B "I sort of like it, but not enough to major in it" (11%)
   C "I am neutral. Maybe I take another Statistics course down the road" (25%)
   D "I don't particularly like it, but I don't mind knowing a little more about it" (37%)
   E "I hate it. I will never ever take another Stats class again" (24%)
6. `Reason` "Why are you taking Statistics 100?"
   A "Because it is required for my major/minor" (66 %)
   B "Because my friends and taking it" (6%)
   C "Because I found no other course that fits my schedule" (10%)
   D "Other reasons" (12%)
7. `GrExp` "Where do you expect your grade in the first exam to be?
   A "In the top 10%" (33%)
   B "In the top 25%" (29%)
   C "In the middle 50%" (23%)
   D "In the bottom 25%" (4%)
   E "In the bottom 10%" (5%)
8. `PMI` Practice Midterm I (0 to 100). The class before the first midterm, we did a practice exam in class. The answers were recorded with the clickers.
9. `MI` Midterm Exam I (0 to 100).
10. `AtMII` Percentage of attendance until the second midterm exam.
11. `CPMII` Percentage of correct in-class clicker questions until the second midterm exam.
12. `PMII` Practice Midterm II (0 to 100)
13. `MII` Midterm Exam II (0 to 100)
14. `AtF` Percentage of class attendance throughout the course
15. `CPFI` Percentage of correct in-class clicker questions throughout the course
16. `PFI`Practice Final (0 to 100)
17. `Final` Final Exam (0 to 100)
18. `HW` Homework (0 to 100)

ANALYSIS

In this section we analyze some aspects of the clicker dataset.

Demographics and Teaching Assistant Effect

We explore now whether the Final grade is related to any of the demographic variables (i.e., sex, level, and school) and whether there is a Graduate Student Instructor effect. The respective boxplots are shown in Figure 1. They show that the variability *within* groups is very large relative to the variability *between* for all four grouping factors. We cannot assert thus that there exist any of these four categorical variables has a clear effect in the final grades.

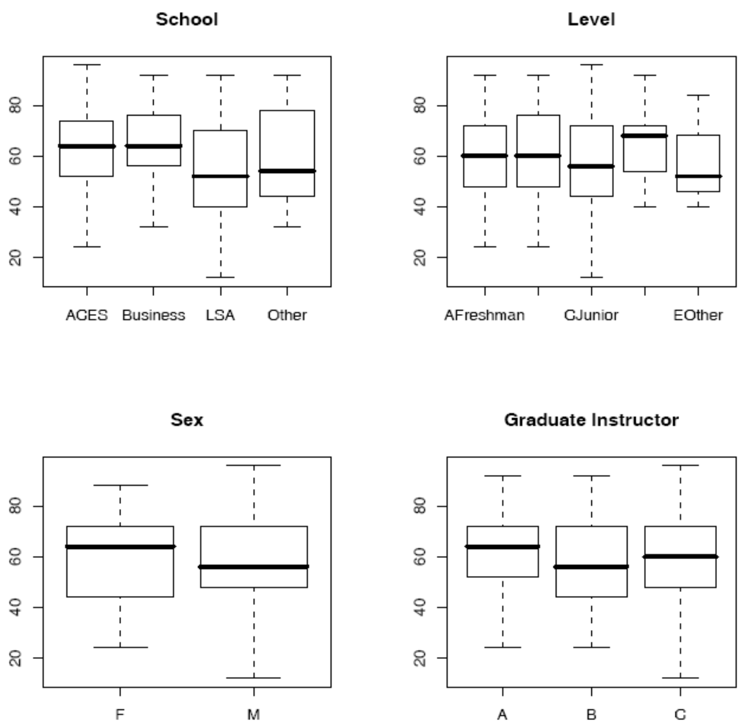

Figure 1: Final grade boxplots I

Attitude and Motivation

In order to asses whether the variables `Like`, `Reason`, or `GrExp` had an effect on final grades, we present Figure 2. Again, the variability *within* groups outweighs the variability *between* groups, to the effect that we do not find any significance.

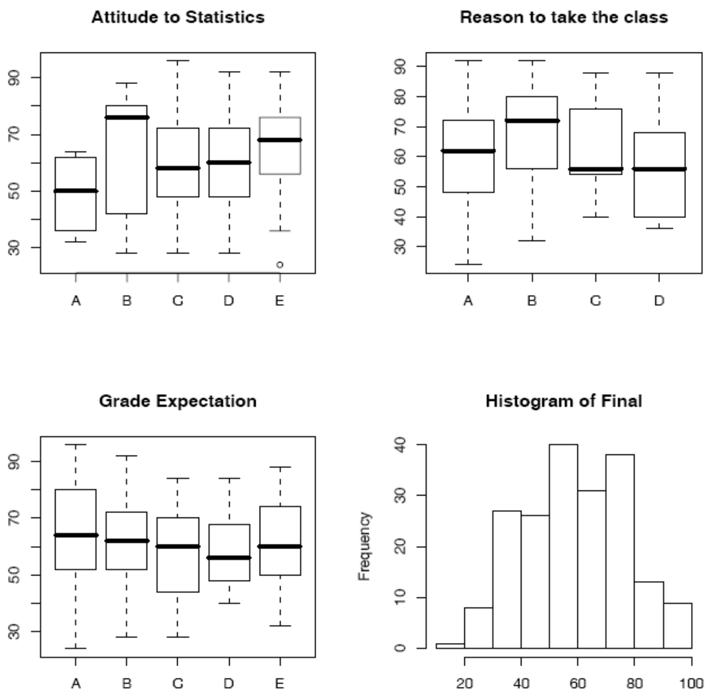

Figure 2: Final grade boxplots II

It is interesting to observe, however, the boxplot of "`Like`," as students who answered "A" were the worst. However, because the percentage of those students is so small (3%), they are not significantly different from the rest. The analysis of the variable `GrExp` about grade expectations deserves special comments. Recall the answers and its distribution {A = "top 10%" (33%), B= "top 25%" (29%), C "middle 50%" (23%), D "bottom 25%" (4%), and E "bottom 10%" (5%)}. Students' grades expectations are thus mathematically impossible. It is also remarkable that they seem to show over optimism on students' part. Now we explore how good students were in predicting their own grade.

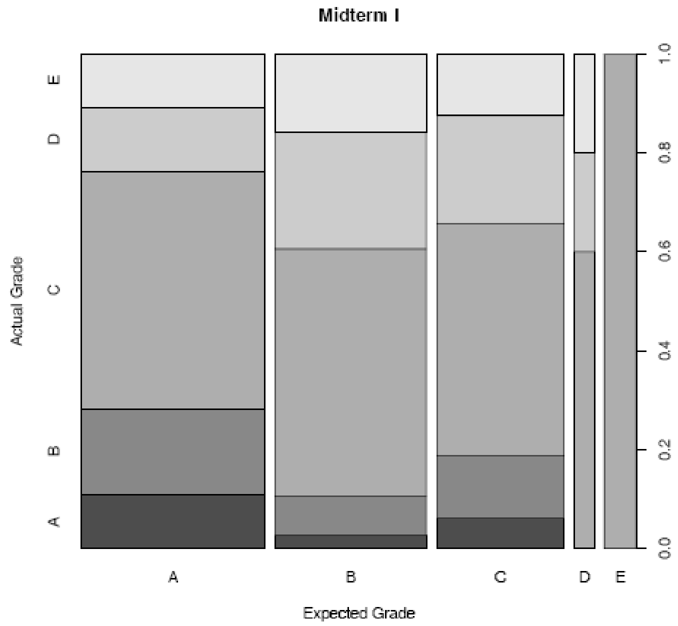

Figure 3: Expected grade versus reality

From Figure 3 we can conclude that the only students who somehow succeeded in predicting their grade were those who expected to be in the bottom 25% or 10%. As for the other students, there is no apparent pattern.

How good are practice-exams?

Data on students performance in practice versus real exams is presented in Figure 4. The location of the distributions for the practice exams is consistently but we see a very weak linear relation.

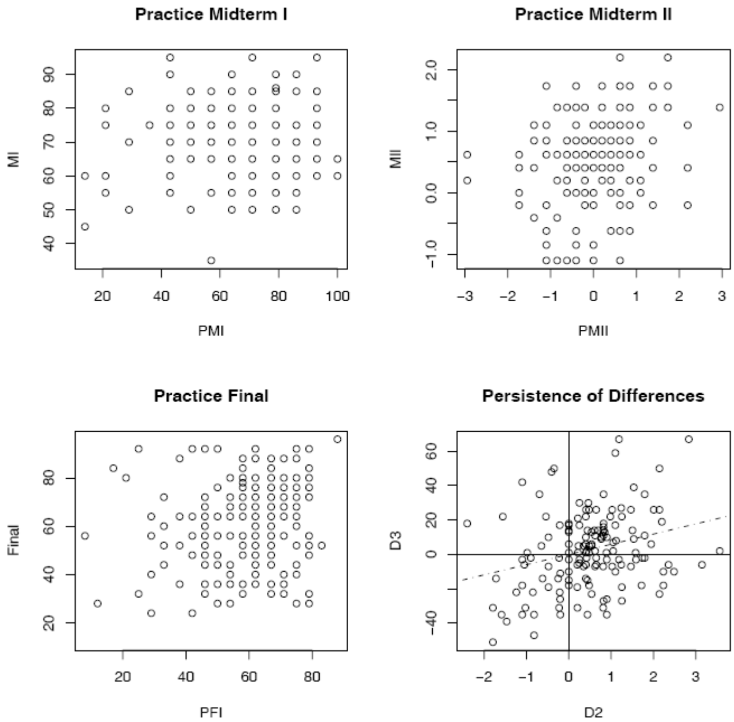

Figure 4: Practice vs. real

Finally, we find a feature in the data that is worth special comment. We constructed two new random variables, `D2` and `D3`, defined as the differences between real and practice exams for `MII` and `Final` respectively. In the last plot of Figure 4 (bottom right) we exhibit the corresponding scatterplot, along with a fitted linear regression line. These data show a somewhat stronger positive relationship. This indicates that, while performance in practice and real exams are virtually unrelated, there is a tendency for students to consistently do better or worse in real exams as opposed to practice.

Are Early Warnings Possible?

We now investigate whether it is possible, during the course of the semester, to "warn" students they are on a path to fail the class. To this extent we fit two linear regression models with the information available at two points in time: (*i*) right after the first midterm, and (*ii*) right after the second midterm exam. For the regression the response `Final` was regressed on `PMI` and `MI`. Both variables were found significant but the R^2 coefficient was only 0.25. Secondly, by the time of the second exam we regressed `Final` on `PMI, MI, AtMII, PMII, CPMII, MII,` and `HW;` the R^2 increased to 0.43. The estimates are very stable as there are very low pairwise correlations between predictors; this is shown in Table 1. We convene this may still not be high enough to predict the final grade with confidence. However, what is relevant is not to predict the final grade, but the overall letter grade for the course. The latter is usually calculated from a weighted average of the two midterms, the final exam and the homework. By the time after the second midterm two of those formula components (`MI` and `MII`) are known exactly. The homework component is known almost exactly, as the final `HW` score is the average of 10 problem sets and 8 of them were turned in and grading by the time of the second midterm. Only the last component, the final grade, is thus the only one that has to be estimated. We conclude that by the end of the second midterm there is enough information to warn a student that he/she is on the road to failing. The only problem is that this warning is no longer *early*, as by the time the second midterm was given and graded only 2 weeks of classes remained!

Model Selection

Finally, we perform model selection for regression models with the `Final` grade as response and the available covariates as predictors. Our aim is to identify with are the factors that best predict success in the class. The three most important variables are Midterm II, Midterm I, and lecture Attendance, in that order. After inclusion of those 3 variables the goodness of fit almost does not improve at all with the inclusion of other factors. This finding is not surprising as they coincide with what most statistics instructors would say. If we asked experienced Statistics lecturers what are the factors that best predict success in their final, most of them would say "midterms and attendance." It is surprising however that clickers do not appear to help at all to predict students success at least for the present dataset. Before dismissing clickers as a predictive tool, however, we would like to repeat the experience with more courses and more extensive clicker questions in the future.

CONCLUSIONS

Teaching with clickers brings the advantage to collect vast day to day data on students' performance. That data has important potential uses: (*i*) fist it can be used *off-the-line* to conduct studies in education aiming at improving the learning process; and also (*ii*) it could be used while the courses are ongoing to reach out to weak students in order to offer help. We carried out an extensive data analysis of an introductory level Statistics class taught with clickers. These are our conclusions:

1. Who is most capable of providing an *early warning* is the student (him)herself. This is because for the instructor to provide an accurate warning, the amount of data necessary requires a wait till almost the end of the course. On the other hand, we saw that all students who predicted their grade to be in the bottom of the class indeed attained low grades.
2. Practice exams are not reliable to predict students' performance in real exams. However there are students who consistently perform better in reality than in practice, and vice versa. We presume it would be promising to collect more data in future educational experiments aimed at characterizing those over and under-performers.
3. The most important variables in predicting students' success in an introductory statistical course are the performance in exams and attendance.

We invite the statistical education community to repeat and revisit our findings with their own data.

ACKNOWLEDGEMENT

This support for the educational experiment described in this article was generously provided by Pearson Education Inc. and Prentice Hall Publishing Co. The author wishes to thank the sponsors for making this rewarding teaching experience possible.