Assessing the Impact of New Teaching Methods by Predicting Student Performance

Abstract

Many teachers try new things in the classroom with the intent of making learning more effective. In most cases, assessment of the impact is anecdotal; the teacher surveys the students about the new technique and draws conclusions based on their feedback. In order to more definitively prove the impact, better assessment tools are needed. In a recent study, the authors attempted to predict performance in a course and then measure the improvement due to a major change in the available resources for study outside the classroom in our fundamentals of engineering course. To measure the effectiveness, we used the GPA of the students at the start of the semester to predict their performance in the course. We then assessed the impact by comparing actual grades in the course to the predicted grades. Using historical data as a baseline, we thus conclude with some certainty the amount of impact our change made in academic performance. This paper focuses on the method of assessment and measurement rather than the classroom changes.

Introduction

Assessing the effect of various changes in the engineering classroom is a daunting task. Student reactions cannot be measured with strain gages or voltmeters. Instead, changes in things such as student attitudes, evolving love of knowledge and student achievement, all the desired outcomes, are difficult to measure with certainty and subject to the wide variations typical of studies involving human subjects (Borrego, 2007). The predictive model discussed in this paper is broadly applicable to the measurement of only student achievement, but is still a powerful tool for making a direct measure of the impact of a particular change on the achievement of a student group.

The best metric by which to predict undergraduate student performance in engineering, science, and technology-related courses is a topic open to further exploration. Researchers have found varying levels of success using diverse student characteristics as indicators. For example, Eskew and Faley (1988) proposed a complex model of factors affecting student performance in beginning accounting courses; among the factors found to be most significant were SAT scores, high school and college GPAs, student motivation, and earlier exposure to the subject material. Based in part on this earlier study, Kruck and Lending (2003) constructed a multi-variable model of student performance and found that student motivation and incoming college GPA did predict student success in introductory Information Systems courses, but that prior performance in related coursework did not. This study also found that SAT scores predicted student success well for male students but not for female students. Van Zwanenberg et al. (2000) found that learning style was not a reliable predictor of success in engineering and business classes and speculated that this may be caused partly by difficulties in reliably measuring learning styles and partly by the wide variety of available ways to accomplish learning objectives. Pritchard and Wilson (2003) found a strong relationship between various indicators of emotional and social health and overall college-level academic performance, though this study was not targeted to success in a specific course.

Felder, Felder, and Dietz (1998), among others, have used control and experimental groups to assess the effectiveness of novel teaching techniques. In this study, the decision was made early in the research period to make a new learning tool available to all of the course's students and allow them to use it if they chose to do so. While this choice maximized the potential benefit of a new learning tool, it removed the availability of a control group against which to test the tool's effectiveness. To provide this needed insight, a "virtual control group" was created using a simple performance prediction model based on the students' incoming cumulative GPAs. The choice to base the prediction solely on prior college performance was motivated in part by the likely availability of GPA information to other instructors hoping to use this technique, and partly by the prior research cited above, which indicates that college performance is generally a good indicator of performance in a specific course.

The broad research question the authors are seeking to answer is: How does the presence of on-demand supplemental resources (videos, instant messaging with instructors, etc) effect student learning? This work has been summarized in other papers (Bruhl, Klosky and Bristow 2008a, 2008b; Klosky and Ressler 2007; Klosky et al 2006). This particular paper seeks to describe only one piece of that work; how well do incoming Grade Point Averages (GPAs) correlate with student performance and can they be used to measure changes in student achievement?

The Prediction Models Examined

Our goal in developing our performance predictor was to quantify the likely student performance in a course which they were entering. This predicted value could then be compared to actual performance, allowing for a quantitative measure of the student's achievement versus rough potential. It should also be noted here that while we show in this paper using the predictor to predict overall performance in a course, the same method can be applied to specific course objectives if the student achievement on those objectives can be broken out in a tested event or problem set and recorded over time.

In order to develop an accurate predictor, we considered a number of factors for both the input and the prediction model. In choosing the input values we desired something that was easily obtained from the registrar rather than a value we would have to measure anew. Secondly, we considered the validity and timeliness of the predictor; for example, using SAT scores to predict performance in a freshman course may be reasonable but using it to predict performance in an upper-level course does not seem as sound as using recent academic performance. As Felder, Felder, and Dietz (1998) also chose, we decided to use the cumulative GPA of a student to predict their performance in a course. This numerical representation of past academic performance is a reasonable way to judge expected performance, but the authors did not compare this method to other possible predictors. Since our course is an engineering course, we considered using a GPA based only on their previous math, science, and engineering (MSE) courses to form a more accurate prediction. After pursuing this idea, we realized the difficulty in obtaining such a number. Each student has quite possibly taken a different combination and number of math and science courses and an "MSE GPA" is not something that our registrar maintains for each student. We would, therefore, have to obtain complete academic records for each student and manually compute this MSE GPA for each individual. Since we desired a

predictor that was readily available, we decided to use the simpler, easily obtained, cumulative GPA.

In addition to choosing the predictive characteristic of their GPA, we desired the most accurate prediction model. In other words, we set out to determine if simply converting GPA to a percentage was as accurate as other models that considered the distribution of grades. In order to investigate this question, we compared the accuracy of three models:

- Model 1 – "GPA to Percent". In this model, we converted each individual's GPA at the close of the previous semester to a percentage and used this percentage to predict performance.

- Model 2 – "GPA z-score with Historical Stats". This model makes use of each individual's z-score of their GPA (that is, how each student has historically performed with respect to his peers). This score is used with the historical course mean and standard deviation for the course to predict each student's score.

- Model 3 – "Gamma-Distribution Percentile". Knowing that the distribution of grades does not historically follow a normal distribution (unless curved – which is not our policy), we used the gamma-distribution in this model since it is a skewed distribution. We computed each individuals percentile with respect to their peers also entering the course. This percentile is based on a gamma-distribution using the GPAs of the population. Each student's percentile is then fed into a gamma-distribution based on historical course statistics to compute a predicted score.

Our goal was to develop a reasonably accurate predictive model which could be used to assess the effectiveness of new teaching methods or resources. In order to evaluate the effectiveness of the model, we examined its accuracy in predicting grades of the 240 students enrolled in CE300 (Fundamentals of Engineering Mechanics and Design) during the Fall 2007 semester. This course had been taught in its current form for 5 semesters prior to Fall 2007 and there were no significant, course-wide changes in methods or resources introduced for this semester. It was thus thought likely that the distribution of incoming and outgoing grades and trends in the variation between them would be similar in Fall 2007 as compared to past semesters. Following determination of the accuracy of the model with this population, we will then use it to assess effectiveness when a new resource is introduced.

Model 1 – "GPA to Percent"

The most straightforward method is Model 1. Using a linear fit based on the standard grade distribution scale used at the Academy (see Table 1), we computed an expected percentage in our course. We used Equation 1 to predict individual performance in our course based on the linear fit of GPA to percent, where $P_{CE300,i}$ is the predicted grade in the course for an individual and GPA_i is that individual's GPA at the start of the semester.

$$P_{CE300,i} = GPA_i \cdot 0.0971 + 0.5432 \quad (1)$$

Table 1 Letter Grade, Grade Points, Percentage Relationship

Letter	Points	Percent	Letter	Points	Percent
A+	4.33	96.67%	C+	2.33	76.67%
A	4.00	93.33%	C	2.00	73.33%
A-	3.67	90.00%	C-	1.67	70.00%
B+	3.33	86.67%	D	1.00	65.00%
B	3.00	83.33%	F	0.00	<65%
B-	2.67	80.00%			

This prediction model yielded reasonable results. The actual performance tended to be a bit above this prediction but was generally close, as seen in Figure 1. The variance for this prediction on the CE300 population from the Fall 2007 semester was 0.30 (sum of squares = 71.9, n = 240). Note that if the actual performance was exactly as predicted, the data points would fall on the diagonal line in Figure 1.

Model 1 tends to underpredict performance; that is, most students tend to perform slightly better in CE300 than they did in previous courses. This is noted in Figure 1 since the majority of the data points lie above the diagonal line. This is also seen statistically: the average difference between actual and predicted performance is 2.0%. This indicates that the average student performs 2% better than predicted and the distribution of this difference is negatively skewed (-0.64) indicating the bulk of the students outperform the prediction.

Also notice in Figure 1 that no students were predicted to finish with above a 93%. This is likely due to the fact that few students earn all A's and A+'s in previous coursework, so very few have a GPA higher than 4.0. This prediction, therefore, fails to predict that students will score an A+ in a future course.

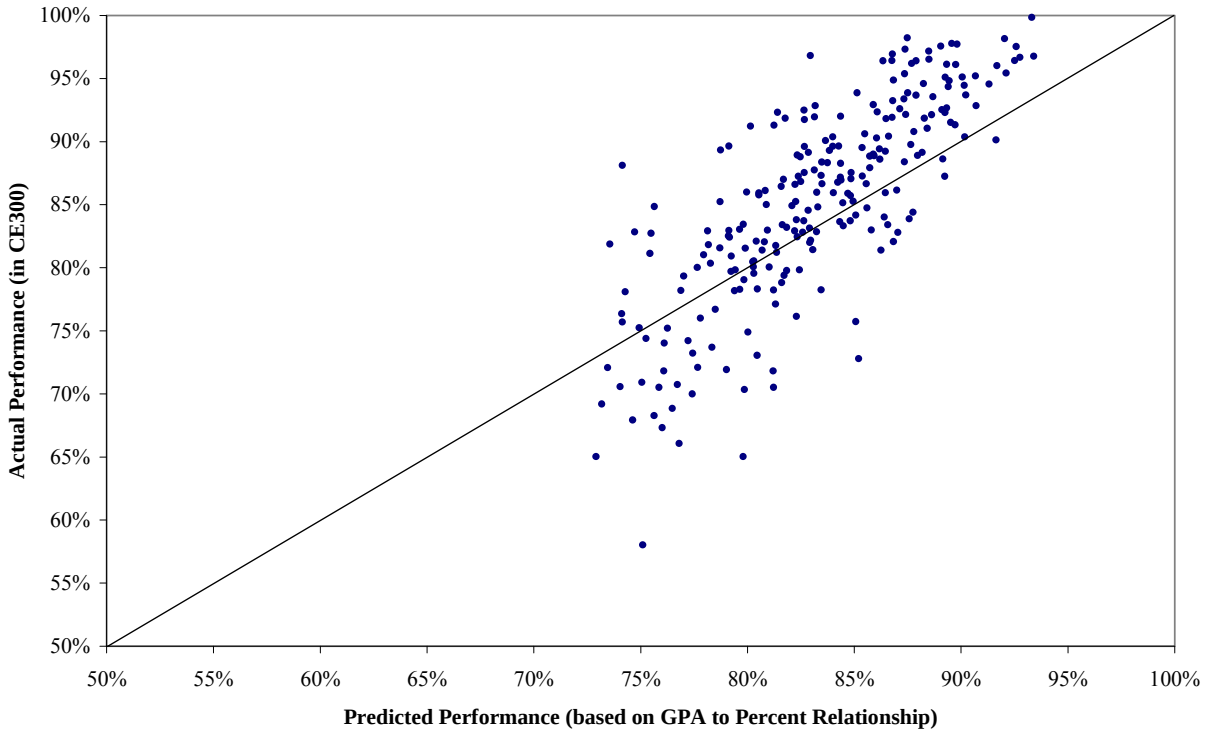

Figure 1 Model 1 Actual versus GPA-Predicted Performance (CE300, Fall 2007)

Model 2 – "GPA z-score with Historical Stats"

For Model 2, we calculated the z-score for each individual (Equation 2) and used it to predict their performance in our course (Equation 3), where $x_{GPA,i}$ is the individual's incoming GPA, x_{GPA} and σ_{GPA} are the incoming group's GPA mean and standard deviation. The predicted score is then computed by adding the historic mean for all past semesters with the individual's z-score multiplied by the historic standard deviation for the course.

$$z_{GPA,i} = \frac{x_{GPA,i} - \bar{x}_{GPA}}{\sigma_{GPA}} \tag{2}$$

$$P_{CE300,i} = \bar{x}_{CE300} + z_{GPA,i} \cdot \sigma_{CE300} \tag{3}$$

This predictive model yielded a slightly more accurate prediction. The reduced spread of data can be seen graphically in Figure 2. The variance of this prediction for the same population as before was 0.26 (sum of squares = 63.4, n = 240) – slightly better than Model 1.

Model 2 tends to overpredict performance; fewer data points in Figure 2 lie above the diagonal line than below. The average difference between actual and predicted performance is -1.4%; the average student performs 1.4% worse than predicted. The distribution of the difference is slightly negatively skewed (-0.07).

Unlike Model 1, this model also provides predictions that some students will score an A+ in the course. In fact, the model must be capped at 100%; if Equation 3 is used without such a cap, some of the top incoming students are predicted to score better than 100% in the course.

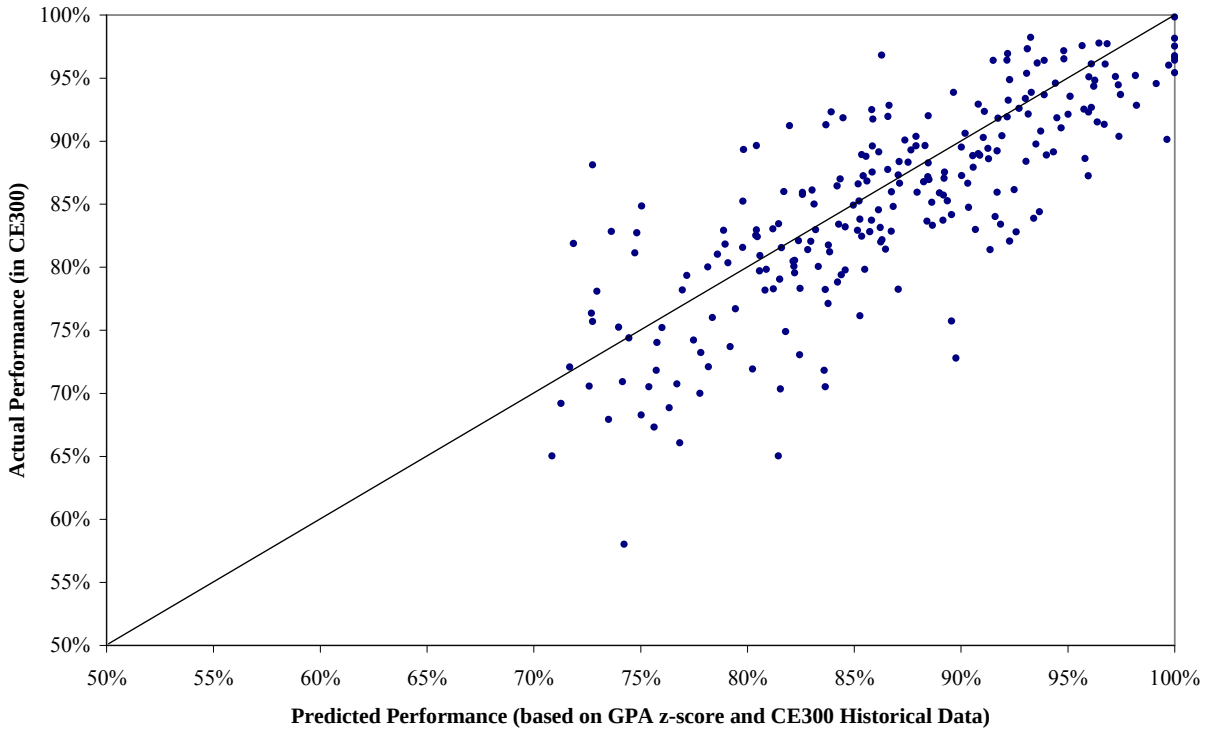

Figure 2 Model 2 Actual versus GPA z-score Predicted Performance (CE300, Fall 2007)

Model 3 – "Gamma-Distribution Percentile"

For Model 3, we computed each student's percentile rank among all students entering the course. This was accomplished using the skewed gamma-distribution with the scale (α) and shape (β) factors computed from the mean and variance of the incoming GPAs as inputs into Equations 4 and 5. We used Microsoft Excel's GAMMADIST function to compute this percentile for each student.

$$\alpha = \frac{\bar{x}}{\beta} \tag{4}$$

$$\beta = \frac{\sigma^2}{\bar{x}} \tag{5}$$

The computed percentile is then used along with scale and shape factors computed from the mean and variance of the historical records for the course to compute a predicted grade in the course. Microsoft Excel's GAMMAINV function was used to compute this predicted grade.

This predictive model is the most accurate of the three investigated. The reduced spread of data can be seen graphically in Figure 3. The variance of this prediction for the same population as before was 0.24 (sum of squares = 57.9, n = 240) – just barely better than Model 2.

Model 3 tends to neither to over- or underpredict performance; notice that roughly the same number of data points in Figure 3 lie above the diagonal line than below. The average difference

between actual and predicted performance is 0.1%; the average student performs almost exactly as predicted. The distribution of the difference is slightly negatively skewed (-0.05).
Like Model 2, this model provides predictions that some students will score an A+ in the course. Unlike Model 2, however, there appears to be no need to cap the model at 100% - the highest predicted grade was 99.8% for the 240 students in this study.

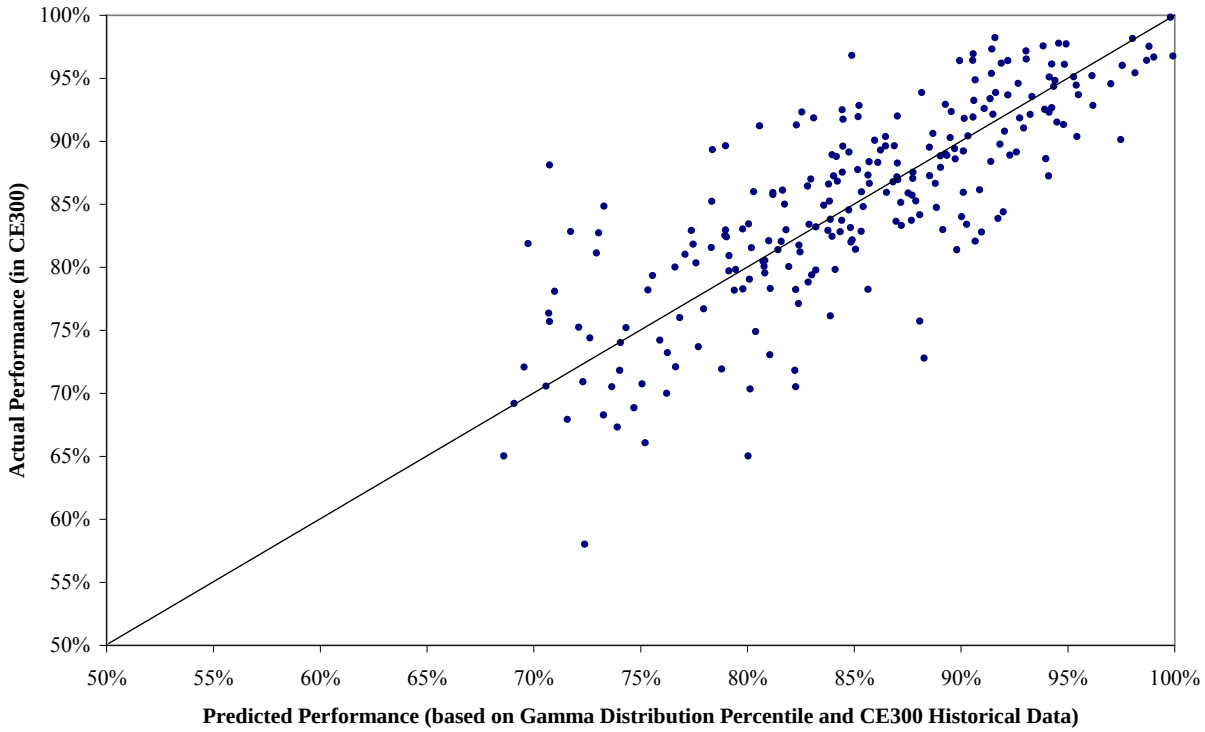

Figure 3 Model 3 Actual versus Gamma-Distribution Percentile Predicted Performance (CE300, Fall 2007)

The Most Accurate Prediction Model

Model 3 – that is, using each student's percentile rank (based on a gamma-distribution of their GPA entering the course) along with the historical mean and standard deviation for the course – is the most accurate predictor of actual performance of the three models considered herein. Examining various statistics, Model 3 most closely mirrors the actual distribution of grades (see Table 2). The mean and median scores along with the standard deviation from this prediction are reasonably close to those from the actual grades achieved in the course and the skewness is closer than the other models. Additionally, the variance is smaller for Model 3 than the other two prediction models investigated.

Table 2 Comparison of Predictor Distributions to Actual Grades (Fall 2007 term)

	Actual Grade	Model 1	Model 2	Model 3
Mean	85.2%	83.2%	86.7%	85.1%
Median	86.0%	83.1%	86.6%	85.2%
Standard Deviation	7.9%	4.7%	7.2%	7.2%
Skewness	-0.55	-0.11	-0.15	-0.22
Variance from Actual		0.30%	0.26%	0.24%

Another way to investigate the accuracy of these models is to examine the "normalized gain" of each student when compared to the prediction. "Normalized gain", N, is defined as the percent improvement of the total improvement possible. This enables a more accurate comparison of improvement between students of differing skill levels. We used Equation 5 to compute the normalized gain (A is actual performance, and P is the predicted performance).

$$N_i = \frac{A_i - P_i}{100\% - P_i} \tag{5}$$

Figure 4 shows histograms of the normalized gains for each of the three prediction models. Note that all three models result in normalized gains that are lumped around 0% gain; most students perform close to predicted.

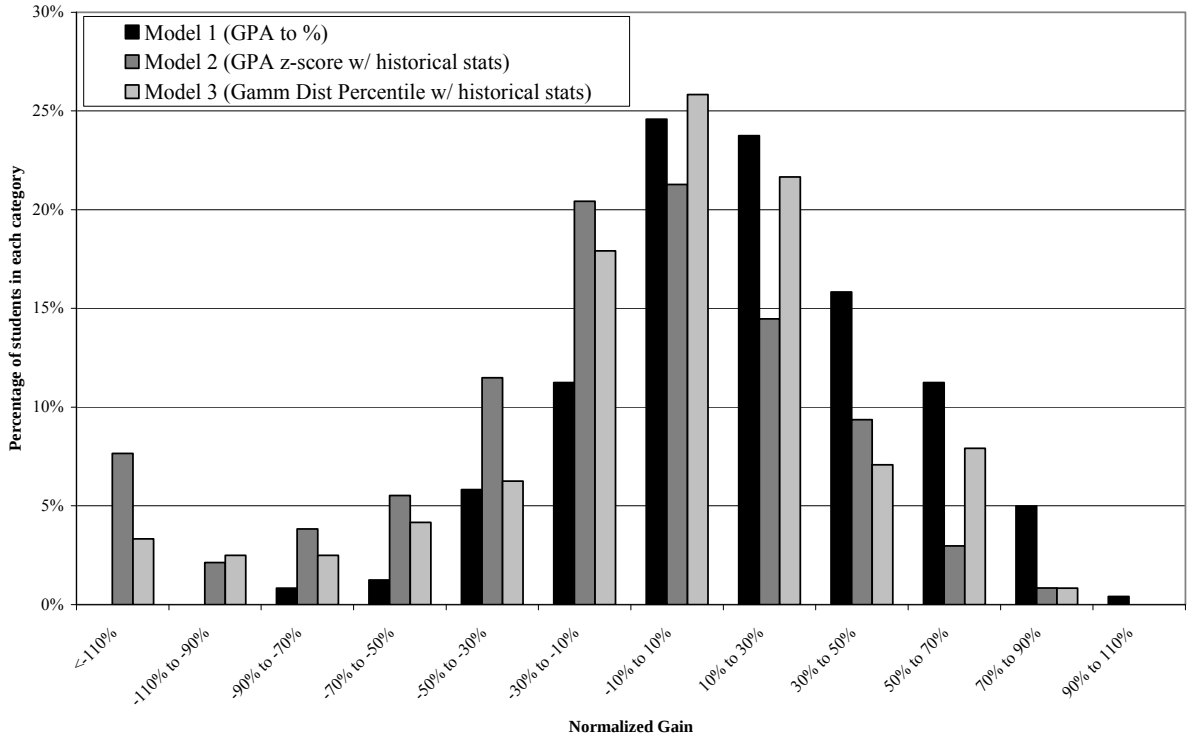

Figure 4 Histogram of Normalized Gains for Various Prediction Models

These histograms corroborate what is seen graphically in Figure 1, Figure 2, and Figure 3 about each model's tendency for over- or underpredicting performance. Model 3 is the most evenly distributed – 38% of the students fall to the left of the (-10% to 10%) category and 37% fall to the right. This distribution is significantly skewed for Model 1 (19% to the left, 56% to the right) and Model 2 (51% to the left, 28% to the right).

The Predictor to Assess New Methods

Accurately predicting performance in a class is only part of the purpose of this paper. Perhaps the more important aspect is using such a prediction to assess the impact of changes to teaching

methods or resources. This, indeed, is what many teachers hope: that what they are doing makes a difference – that it improves learning. Having a reasonably accurate prediction of how students should perform in a course provides the teacher with a baseline with which to compare actual results and then draw conclusions.

The authors used the prediction for just such a purpose during the Fall semester of 2008 when they introduced a new resource to assist students in the understanding of material, the completion of assignments, and the preparation for exams. This resource consisted of short, instructor-created, tutorial videos on various topics within the course (Bruhl, Klosky, and Bristow 2008a, 2008b). We called this resource "Video AI" ("AI" stands for Additional Instruction) and made it available to all 181 students enrolled in the course.

In order to assess the impact of this new resource, we used Model 3 to predict performance in the course. 112 students chose to watch at least one video during the course of the semester. Therefore, we had two self-selected groups: those who watched Video AI, and those who did not. Comparing the statistics of the groups in Table 3 it is clear that there is some academic benefit to using this resource as a supplement to the other, more commonly used, resources such as textbooks, notes, and example problems worked in class.

Table 3 Comparative Statistics (Predicted vs. Actual Grades) by Population

	Statistic	Predicted Grade	Actual Grade
Entire Population (n=181)	Mean	86.2%	86.4%
	Median	86.7%	87.4%
	Standard Deviation	7.6%	7.6%
	Skewness	-0.17	-0.62
Watched Video AI (n=112)	Mean	85.4%	86.2%
	Median	86.1%	87.7%
	Standard Deviation	7.8%	7.7%
	Skewness	-0.20	-0.56
Did not Watch Video AI (n=69)	Mean	87.3%	86.9%
	Median	88.2%	87.3%
	Standard Deviation	7.1%	7.5%
	Skewness	0.02	-0.73

As noted in Table 3, the prediction is very close to the actual grades for the entire population, corroborating the findings in the previous section. When examining the sub-populations, it is noted that those who watched Video AI slightly outperformed the prediction while those who did not slightly underperformed. While further analysis does not show this difference in performance to be statistically significant, these videos clearly have some (albeit, small) impact on academic performance.

Another way to examine the difference between the populations is through a histogram of the normalized gains. A glance at Figure 5 shows that those who saw the largest gains were from the group of students who made use of the videos while studying. Examining Figure 5 makes clear that this resource is not a sure-fire way to succeed in the course. More study of the impact

of these videos is necessary before drawing any substantive conclusions. However, without a reasonably accurate predictive model, it would be difficult to include such a quantitative discussion of the academic impact of this new teaching resource and we would be forced to rely more significantly on anecdotal data.

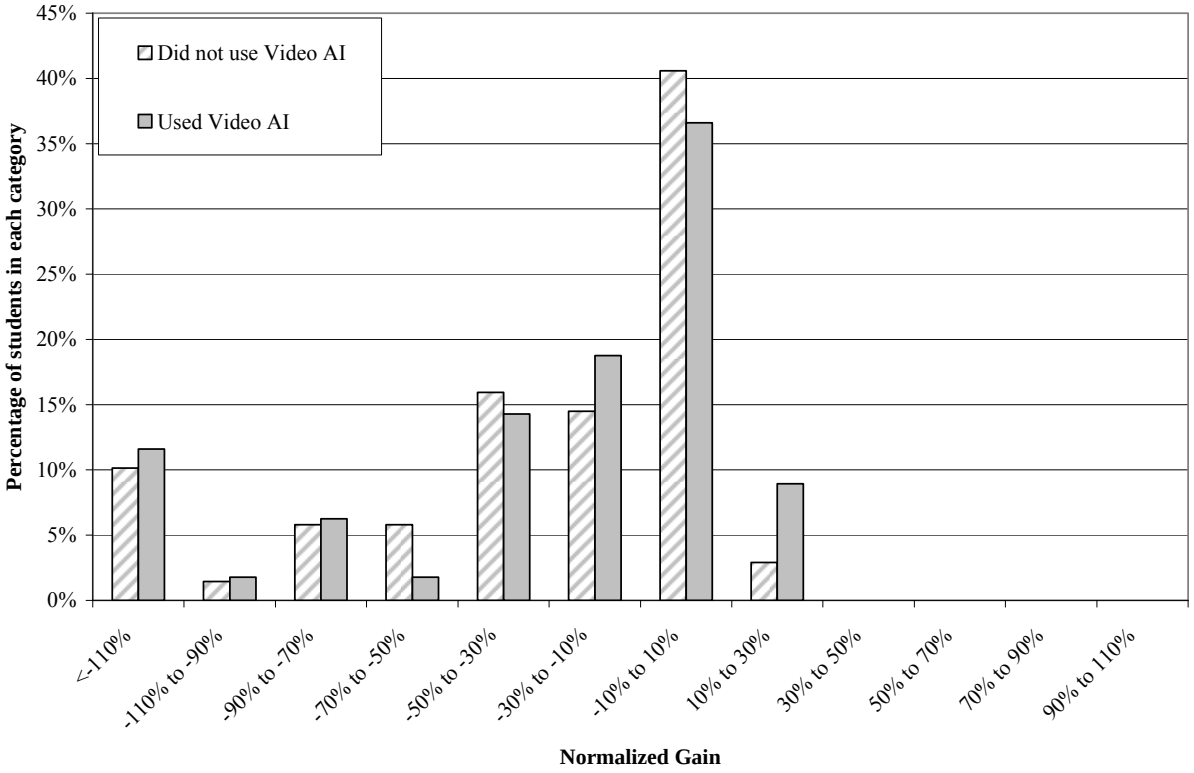

Figure 5 Histogram of Normalized Gains (by Population, CE300, Fall 2008)

Conclusion

Assessing the impact of new teaching methods or resources is important to improving the learning that occurs in our courses. In order to prepare an assessment based on both quantitative and qualitative data, a reasonably accurate prediction of performance in a course is necessary.

Using the gamma-distribution model of incoming GPAs to compute a percentile for each student coupled with historical statistics of the course they are entering results in a reasonably accurate grade prediction. This prediction can then be compared to actual performance in a course in which a teacher is trying something new to assess its impact.

Bibliography

Borrego, Maura (2007). "Conceptual Difficulties Experienced by Trained Engineers Learning Educational Research Methods". Journal of Engineering Education, Vol 96, No 2, pp 91-102.

Bruhl, Klosky and Bristow, 2008a. "Watching Videos Improves Learning?" American Society for Engineering Education, 2008 National Conference.

Bruhl, Klosky, and Bristow, 2008b. "On Demand Learning – Augmenting the Traditional Classroom" American Society for Engineering Education, 2008 National Conference.

Klosky and Ressler. 2007. "Asynchronous delivery of engineering courses to a widely dispersed student body". American Society for Engineering Education, 2007 National Conference, Honolulu, HI.

Klosky, Hains, Ressler, Evers and Erickson. 2006. "AIM for Better Student Learning: Best Practices for Using Instant Messaging and Live Video to Facilitate Instructor-Student Communication." American Society for Engineering Education, 2006 National Conference, Chicago, IL.

Eskew, R.K. and Faley, R.H. (1988). "Some Determinants of Student Performance in the First College-Level Financial Accounting Course." *The Accounting Review*, 63(1), pp. 137-147.

Felder, R.M., Felder, G.N., and Dietz, E.J. (1998). "A Longitudinal Study of Engineering Student Performance and Rentention: v. Comparisons with Traditionally-Taught Students." *Journal of Engineering Education*, 87(4), pp. 469-480.

Kruck, S.E., and Lending, D. (2003). "Predicting Academic Performance in an Introductory College-Level IS Course." *Information Technology, Learning, and Performance Journal*, 21(2), pp. 9-15.

Pritchard, M.E. and Wilson, G.S. (2003). "Using Emotional and Social Factors to Predict Student Success." *Journal of College Student Development*, 44(1), pp. 18-28.

Van Zwanenberg, N., Wilkinson, L.J., and Anderson, A. (2000). "Felder and Silverman's Index of Learning Styles and Honey and Mumford's Learning Styles Questionnaire: How do they compare and do they predict academic performance?" *Educational Psychology*, 20(3), pp. 365 – 380.