Evaluating a predictive model of student performance in introductory calculus-based physics

M Greene and R Lopez
Physics Department, The University of Texas at Arlington, Arlington TX 76019 USA
michaelagreene@mavs.uta.edu

Abstract. Various factors are known to correlate with performance in physics and other STEM courses. In this study, we report on students’ performance in an introductory calculus-based physics course at the university level and correlate their performance (as measured by final grade) with a variety of measures of academic and cognitive ability. These measures include mathematical and English language ability (as measured by the SAT), incoming GPA, and spatial and scientific reasoning ability (as measured by the Mental Rotation Test and Classroom Test of Scientific Reasoning). We also consider the effect on student performance of factors such as employment status, and science and mathematics self-efficacy (as measured by a survey). In this study, we report the prediction efficiency of the ability of two multivariate linear models to predict student success in the course.

1. Introduction
Numerous studies have shown that poor performance in early STEM courses result in students dropping the STEM major altogether [1–3]. There has been some work in developing models to predict the performance of STEM students accurately, but these techniques so far have relied on data that are collected part-way into the course, sometimes even using midterm and quiz grades in the model [4, 5]. The purpose of this study is to develop, construct, and evaluate predictive models of student performance in physics using data that can be collected very early in the semester (or even before classes begin). If accurate and reliable predictive models could be constructed, researchers, instructors, and school administration staff could identify the students who may be at risk for failing the course. Additionally, these models could also be used to identify potentially high-performing students who may benefit from additional projects or material to enhance their understanding.

1.1. Student Population
The participants of this study are primarily STEM-major students attending The University of Texas at Arlington, an R1 Hispanic Serving Institution, enrolled in a calculus-based introductory physics course. Data was collected during four consecutive semesters from fall 2015 to spring 2017 (N=316) in accordance with Institutional Review Board policy. Taking this course is a requirement for their majors and many of the students must earn a final grade of at least 70% since students in the College of Engineering (the primary population served in the course) as well as various science majors must retake the course if they earn less than the required 70%.
2. Methodology

In this study, we develop two predictive models of student performance using multiple linear regression. In these models, we use factors that are well known to correlate or influence performance in physics as the input values, and the students’ final grade as the output value. The chosen factors, how they were measured, and how they were collected are shown in Table 1. The accompanying references include previous studies on the relationship between the factor and success in STEM courses. Data and final grades were collected from students over three semesters (fall 2015 through fall 2016), and used as the foundation of the empirical data set. From this set, two different multiple regression models were developed; one used all eight of the factors shown in table I, and another using three factors. During the spring 2017 semester, data were collected at the beginning of semester like the rest, but were used as input values, yielding a predicted final grade as the output value. In other words, data from three semesters were used to develop the models, and a fourth semester was used to test the predictive ability of the models.

To determine the prediction efficiency, a measurement of how well the model makes a reliable prediction, we compare the predicted final grade to the actual final grade earned in terms of simple YES or NO questions such as “Will the student get an A?” or “Will the student fail the course?” The meteorological community has developed an extensive set of tools for measuring the accuracy of such dichotomous predictions, and it begins with a contingency table, shown in Table 2, which accounts for the four possible combinations of YES/NO events for predictions and observations [6].

In Table 2, \( H \) is the number of hits, \( F \) is the number of false alarms, \( M \) is the number of misses, and \( N \) is the number of correct negatives. A hit represents a correct prediction (student X was predicted to fail and that occurred) a false alarm is an incorrect prediction (student X did not fail). A miss represents an event which did occur which was not predicted, while a correct negative represents a NO event occurring with a correct forecast.

### Table 1. Factors that influence performance in physics

| Factor                      | How it’s measured       | How it’s collected       |
|-----------------------------|-------------------------|--------------------------|
| Mathematical Ability [7]    | SAT Math Score          | Student Transcript       |
| Reading Ability [8]         | SAT Reading Score       | Student Transcript       |
| Previous Grades [9]         | Grade Point Average     | Student Transcript       |
| Spatial Ability [10,11]     | Mental Rotation Test [14]| In-class Assessment      |
| Scientific Reasoning Ability [12]| Classroom Test of Scientific Reasoning [15]| In-class Assessment |
| Employment Status [13]      | Number of Hours Worked  | In-class Survey          |
| Mathematics Self-Efficacy [13]| Likert-type Item       | In-class Survey          |
| Science Self-Efficacy [13]  | Likert-type Item        | In-class Survey          |

### Table 2. Standard contingency table for dichotomous predictions.

| Observe YES | Observe NO | Total |
|-------------|------------|-------|
| Predict YES| Hit \((H)\) | False Alarm \((F)\) |
| Predict NO  | Miss \((M)\) | Correct Negative \((N)\) |
| Total       | \(OY = H + M\) | \(ON = F + N\) |

\[ PY = H + F \]

\[ PN = M + N \]

\[ T = PY + PN = OY + ON \]
For our purposes we asked the model, "Will the student fail the course?" In this course, a failing grade is defined as less than 70 out of 100 points. A failing grade is a relatively rare event, and the best metrics to evaluate the prediction efficiency of rare events are the probability of detection (POD) and the false alarm ratio (FAR) [16]. POD is defined as the ratio of “hits” to the total “yes” observations, and produces a number between zero and one. A zero means the model does not detect any hits, and a one indicates perfect predictions. FAR is defined as the ratio of false alarms to the total predicted “yes” answers, and produced a number between zero and one. A zero indicates no false positives, and a one indicates that every event predicted was a false positive.

\[
POD = \frac{H}{H + M} \quad (1)
\]
\[
FAR = \frac{F}{H + F} \quad (2)
\]

3. Results

3.1. Eight-factor model
Aggregating the three semesters of data for each of the eight factors listed in Table 1, we had a sample of 33 students from fall 2016 to fall 2017 with which to construct a multiple regression model with those eight factors. We were able to collect data on all eight factors from Table 2 from 16 students in the spring 2017 semester. The POD and FAR for the model are shown in Table 3, as well as the \( R^2 \) value between predicted grade and earned grade.

Table 3. Eight-factor model summary

| Model     | POD | FAR  | \( R^2 \) | N  |
|-----------|-----|------|-----------|----|
| Fnl_Grd = 26 + 0.0592*SAT_M - 0.0142*SAT_R + 8.36*GPA + 0.306*MRT + 0.594*CTSR - 0.736*HrsWrk -1.06*Conf_S - 0.45*Conf_M | 0   | Undefined | 0.7314 | 16 |

3.2. Three-factor model
Using the top three factors with the highest individual linear correlation coefficients (GPA, SAT math score, and scientific reasoning ability), we had sample of 45 students with which to construct a multiple regression model with the three factors. The POD and FAR for the 31 students in the spring 2017 semester for whom we could collect data for all three factors are shown in Table 4.

Table 4. Three-factor model summary

| Model     | POD  | FAR  | \( R^2 \) | N  |
|-----------|------|------|-----------|----|
| Fnl_Grd = 26.4 + 0.0342*SAT_M + 9.36*GPA + 0.579*CTSR | 0.75 | 0     | 0.5503   | 31 |

4. Discussion
The biggest problem of the eight-factor model is that it doesn’t detect either of the two students in the spring (for whom all eight factors could be measured) who actually ended up failing the course – it incorrectly predicted that those two students would earn a passing grade. Another problem is that these eight factors may be too constraining. The number of students upon which the 8-factor model was constructed from is small. The low number is due to the fact that not every student will have exactly all eight factors measured, and the transcripts used either provides SAT or ACT scores, not both, so there is a problem with missing SAT scores. Additionally, GPA is recorded only from UTA courses taken, so students who are in their first semester, whether as freshman or transfers, will not have a GPA to report. In the future, a larger dataset may eliminate some of these constraints and provide better prediction
outcomes. In the meantime, the three-factor model is based on factors that can be easily measured and collected early in the semester, which helps increase the number of students that can be included in both the creation and implementation of the predictive model. Now out of the 31 (instead of 16) students for whom we could predict from the spring 2017 semester, we were able to identify three out of the four students who actually did fail using this three-factor model with no false alarms. Upon investigating the fourth student who was not correctly predicted to fail, we found that this individual failed other courses that semester as well, even though the prior academic history of the student was strong. We speculate that something happened in this individual’s personal life that could not have been predicted by any academic model. In future studies, reflecting on the false alarms—those who were predicted to fail but passed—may provide insight into additional factors that are influential to academic performance.

5. Conclusion
We have constructed a preliminary predictive model of student performance in a calculus-based introductory university mechanics course. Using an empirical data set aggregated from three semesters, we constructed two multiple regression models to predict student performance in the course of an upcoming semester. One model uses eight factors known to correlate or influence performance in physics (see Table 1), and the other uses only the top three strongest-correlating factors (SAT math, GPA, and scientific reasoning ability). We used model prediction metrics developed in meteorology to evaluate the prediction efficiency of these models. The three-factor model gave a high probability of detection (POD) with a low false alarm ratio (FAR), which is an ideal result. Further studies are being conducted to test the reliability and generalizability of the models.

Acknowledgements
We acknowledge support from Department of Education grants P200A130247 and P200A090284.

References
[1] Crisp G, Nora A, and Taggart A 2009 Student characteristics, pre-college, college, and environmental factors as predictors of majoring in and earning a STEM degree: An analysis of students attending a hispanic serving institution Am. Educ Res. J. 46 924–42. DOI: 10.3102/0002831209349460
[2] Graham M J, Frederick J, Byars-Winston A, Hunter A, and Handelsman J 2013 Increasing persistence of college students in STEM Am. Ass. Adv. Sci. 341 1455–6. DOI: 10.1126/science.1240487
[3] Maltese A V and Tai R H 2011 Pipeline persistence: Examining the association of educational experiences with earned degrees in STEM among US students Sci. Educ. 95 877–907. DOI: 10.1002/sce.20441
[4] Jishan S T, Rashu R I, Haque N, and Rahman R M 2015 Improving accuracy of students’ final grade prediction model using optimal equal width binning and synthetic minority oversampling technique Decis. Anal. 2(1). DOI: 10.1186/s40165-014-0010-2
[5] Meier Y, Xu J, Atan O, and van der Schaar M 2016 Predicting grades IEEE T. Signal Proces. 64 959–72 doi: 10.1109/TSP.2015.2496278
[6] Stanski H R, Wilson L J, and Burrows W R 1989 Survey of common verification methods in meteorology WMO/ TD 358. World Meteorological Organisation, Geneva, Switzerland.
[7] Meltzer D E 2002 The relationship between mathematics preparation and conceptual learning gains in physics: A possible “hidden variable” in diagnostic pretest scores Am. J. Phys. 70 1259–68. DOI: 10.1119/1.1514215.
[8] Koch A, Eckstein S G 1995 Skills needed for reading comprehension of physics texts and their relation to problem-solving ability J. Res. Sci. Teach. 32 613–28. DOI: 10.1002/tea.3660320607.
[9] Allen J, Robbins S B, Casillas A and Oh I 2008 Third-year college retention and transfer: Effects of academic performance, motivation, and social connectedness Res. High. Educ 49 647–64.
DOI: 10.1007/s11162-008-9098-3

[10] Wai J, Lubinski D, and Benbow C P 2009 Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *J. Educ. Psychol.* **101** 817–835 DOI: 10.1037/a0016127

[11] Wai J, Lubinski D, Benbow C P, and Steiger J H 2010 Accomplishment in science, technology, engineering, and mathematics (STEM) and its relation to STEM educational dose: A 25-year longitudinal study. *J. Educ. Psychol.* **102** 860–71 DOI: 10.1037/a0019454

[12] Coletta V P, Phillips J A 2005 Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability *Am. J. Phys.* **73** 1172–82. DOI: 10.1119/1.2117109.

[13] Pike G R, Kuh G D, and Massa-McKinley R C 2008 First-year students’ employment, engagement, and academic achievement: Untangling the relationship between work and grades *NASPA J.* **45** 560–82. DOI: 10.2202/1949-6605.2011.

[14] Vandenberg S G, Kuse A R 1978 Mental rotations, a group test of three-dimensional spatial visualization *Percept. Mot. Skills* **47** 599-604. DOI: 10.2466/pms.1978.47.2.599

[15] Lawson A E 1978 The development and validation of a classroom test of formal reasoning *J. Res. Sci. Teach.* **15** 11–24. DOI: 10.1002/tea.3660150103

[16] Lopez R E, Hernandez S, Wilberger M, Huang C, Kepko E L, Spence H, Goodrich C C, and Lyon J G 2007 Predicting magnetopause crossings at geosynchronous orbit during the halloween storms *Adv. Space Res.* **5** S0105 DOI: 10.1029/2006SW000222