How Early Is Early Enough: Correlating Student Performance With Final Grades

Jason M. Pittman
High Point University

Kimberly Titus
High Point University

Lloyd Williams
High Point University

Student retention is one of the greatest challenges facing computer science programs. Difficulties in an introductory programming class often snowball, resulting in poor student performance or dropping the major completely. In this paper, we present an analysis of 197 students over 6 semesters from 11 sections of an introductory programming class at a private four-year liberal arts university in the southeastern United States. The goal of this research was to find the earliest point in the assessment sequence which could predict final grade outcomes. Accordingly, we measured the degree of correlation between student performance on quizzes, labs, programs, and tests compared to final course grade. Overall, the results show a strong positive correlation for all four assessment modalities. These results hold significance for educators and researchers insofar as the body of computing education research is extended by evaluating the relative effectiveness of early semester subsets of each of the four categories of student data to model class outcomes.

Keywords: retention, educational data mining, learning analytics, grades, student performance

INTRODUCTION

Difficulties in an introductory programming class often quickly snowball, resulting in poor student performance and eventual failure in the class. Far too often, the quite legitimate challenges faced by predictably computationally illiterate students enrolled in a first-year programming class result in them dropping the major completely. The course serves as an initiation to the computer science major at a private four-year liberal arts university in the southeastern United States. In some semesters the department has observed up to 25-33% attrition in the introductory course, consistent with observations at other universities where 28% of CS majors change their major within the first 3 years (Leu, 2017). This remarkably high rate makes the class an ideal candidate for targeted early interventions to reduce the attrition rate.

One aspect of the underlying problem in these circumstances is interventions with struggling or failing students may only become practical after mid-semester grades are reported. This dramatically affects student success in these introductory programming courses. Moreover, research (Quille & Bergin, 2019; Salguero, McAuley, Simon, & Porter, 2020) has shown that the likelihood of dropping the major rises sharply as the semester goes on. Accordingly, the purpose of this research was to find the earliest point in the course assessment sequence it might be possible to predict final grade outcomes. If such points exist, targeted intervention may potentially lead to reduced course failure or abandonment rates as well as increased degree retention.

RELATED WORK

CS educators have investigated models to predict student performance for several decades. Early studies (Butcher & Muth, 1985) were reasonably successful in using high school data and ACT scores as predictors for both the introductory programming course and general college performance. Several studies since the early 2000s showed enormous promise in employing educational data mining and learning analytics (Ihantola et al., 2015) although replication of such studies have not always proven consistent (Bornat & Dehnadi, 2008). More recent studies have considered gender, years of prior programming experience, previous college course grades, cognitive development, cognitive style, or personality type as predictors (Werth, 1986; Guo, 2020). Other work has found improvement in retention and CS1 pass rates through media computation, paired programming, and peer instruction (Salguero et al., 2020). Studies focused on student self-assessment and self-efficacy determined that self-predictions were seldom accurate and discovered that students reporting more frequent negative self-assessment had lower self-efficacy (Gorson & O'Rourke, 2020; Sobral, 2020).

Quille and Bergin's (2019) predictive model developed over 13 years was found to be 71% effective in determining early success, can identify weak students at 80-89% accuracy leading to intervention that improved grades up to 9%. Our work focuses on finding a single assessment predictor in the first half of the semester that provides an intervention point.

Our introductory course has been taught in the C language using a structured programming approach. A student's final grade is calculated from 3 tests each worth 10%, 5 programming assignments each worth 6%, weekly labs collectively worth 10%, weekly quizzes collectively worth 10% and a final exam worth 20%. The final exam grade can replace a missed or lower test score at the end of the semester. Content of the first test, first 3 labs, first 3 quizzes and first program has typically included basic UNIX commands, variable declaration, assignment statements, arithmetic operators, debugging, code tracing and code writing. The second test, quizzes 4-6, labs 4-6, and program 2 cover for, while, and do…while loops, conditional statements, and operators, debugging, code tracing and code writing. Functions, arrays, structs and pointers are covered in the second half of the semester. Assessment items for these later topics were not considered in our analysis since an early predictor is desirable.

Methodology

The goal for this study was to find at what point, if any, in a freshmen introductory programming course it might be possible to predict final grade outcomes. The motivation for this goal was to identify the earliest opportunity for intervention targeted towards increasing computer science degree retention. To that end, we measured the degree of correlation between four categories of course assessments compared to final course grades.

Data

We used historical data from 11 course sections arrayed over 6 semesters ranging from 2016 until 2019. Data were organized according to four pre-existing course assessment categories: laboratory assignments (e.g., labs), quizzes, programming assignments, and tests. Within those assessment categories, we isolated data from the first half of the six semesters. Data associated with assessment occurring after the mid semester point might be actionable in various contexts but is outside the scope of any meaningful intervention in our specific operating environment.

Collectively, the data span 48 weeks of instruction and student work. Within that time frame, labs, and quizzes both had five assessment measurements per half-semester while tests had two per half-semester. The data represented 197 students yielding 2,559 unique assessment observations.

Procedure

We analyzed the data using a three phased approach. First, a scatterplot with linear trend line revealed the nature of correlation between assessment category and final grade. Then, we ran a total of 14 linear regressions (five for labs, five for quizzes, two for programs, and two for tests) against individual assignments and final grades. Finally, we ran a series of progressive multivariate linear regressions within each category to supply a cumulative correlation up to and including the entire category (which aligned with the scatterplot r values).

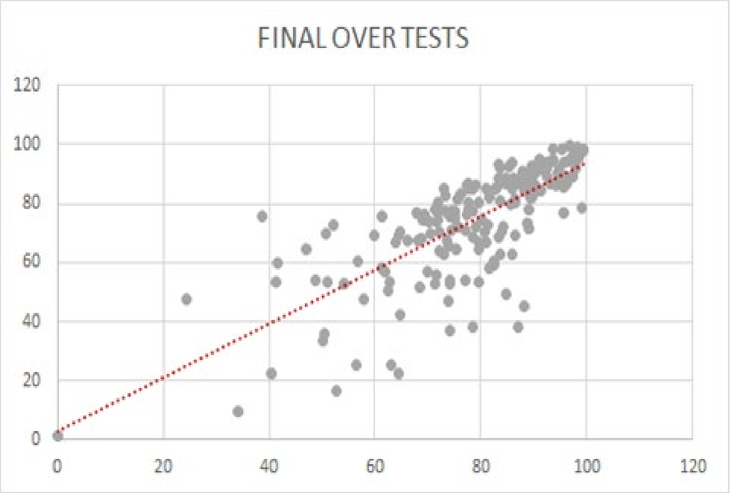
Limitations

This work has three limitations. First, data were historical and may not have bearing on current educational contexts. Second, the data potentially included students not submitting work and thus earning zero points. Lastly, the work is limited in statistical power by a small and tightly bounded student sample.

RESULTS

Broadly, the results show strong positive correlations between all four assessment categories and final grades. Interestingly, the data also revealed specific opportunities for intervention prior to what the institution has traditionally considered the midway point of a semester. Furthermore, every individual assessment (X variables) showed a statistically significant result and showed a test statistic supporting rejection of null hypotheses.

FIGURE 1
CORRELATION BETWEEN TWO TESTS AND FINAL COURSE GRADE

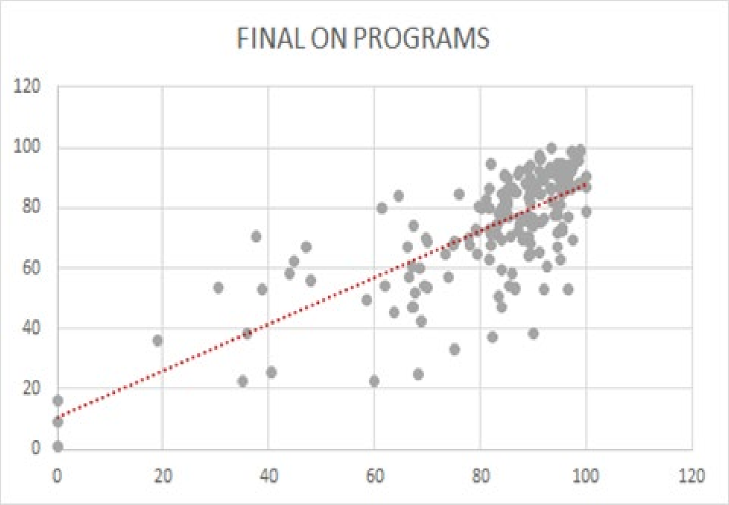


At face value, such results make sense because all assessments contribute to the final grade to varying degrees. However, the question we sought to answer was not what assessments contribute to final grades, but which assessments may serve as early predictors of final grade. With all of that in mind, we present the results in order of which assessment categories may best fit such use.

Tests

Figure 1 shows a strong positive correlation between tests and final grades. This was the strongest correlation in the study with a r value of 0.752. There were two assessments in this category. The first test displayed a strong positive relationship with 55% of data conforming to the model. In contrast, the second test revealed a moderate positive relationship with 47% of data fitting the model. Moreover, the degree of relationship between both tests and final grades was similar to the first test alone but differed significantly from Test 2 alone.

FIGURE 2
CORRELATION BETWEEN SCORES ON TWO PROGRAMS AND FINAL COURSE GRADES



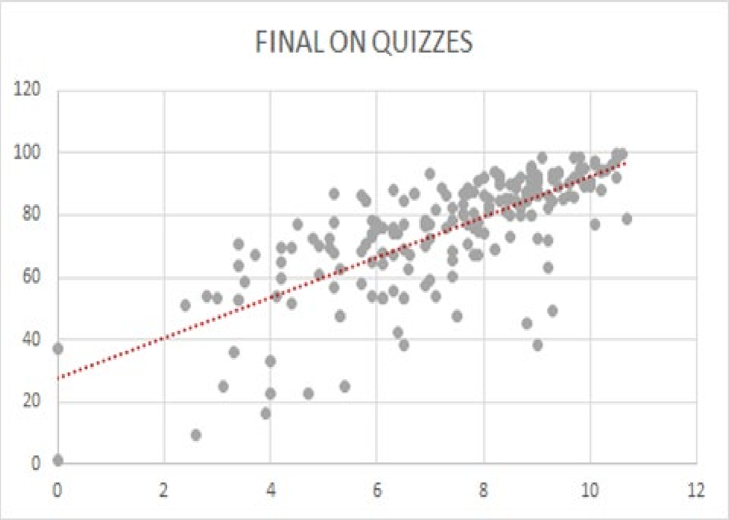
Programs

Programming assignments showed the second strongest correlation (Figure 2) with a r value of 0.745. As with tests, there were two assessments in this category. However, both assessments demonstrated moderate positive relationships. Meanwhile, the second programming assignment had a better model fit (45%) than the first (36%). Taken together, the programming assignments showed a 55% fit with a strong positive relationship.

Quizzes

Quizzes, as shown in Figure 3 were a close third compared to tests and programs with a r value of 0.729. Despite differing overall from Tests and Programs by small margin, individual quiz data were much less clustered around high grade scores. Individual linear regressions reflected such variance insofar as some quizzes (i.e., Quiz 2 and Quiz 4) had weak-moderate relationships with final course grades while other quizzes (i.e., Quiz 5) had moderate-strong relationships. Furthermore, the strength of the correlation increased dramatically as quizzes were analyzed in collections (i.e., Quiz 2, 3, 4, and 5) up to showing a strong relationship with 50% data fitting the model.

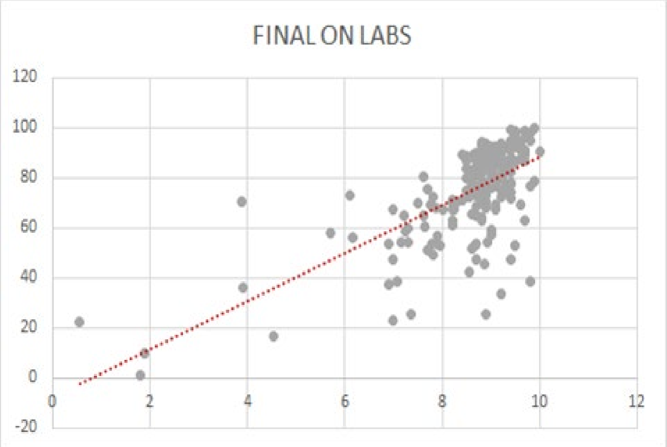
FIGURE 3
CORRELATION BETWEEN SCORES ON QUIZZES AND FINAL COURSE GRADES



Labs

Labs (Figure 4) diverged from other assessment categories in both form and the degree of clustering but still showed a positive correlation with an r value of 0.682. Interestingly, labs demonstrated a moderate positive relationship across all individual assignments but varied within the category quite a bit. For example, Lab 2 revealed a borderline weak relationship to final course grade and had just 18% fit. Yet, the next lab (Lab 3) showed a more robust moderate relationship with 31% data fit to the model. Further, several individual labs such as Lab 6 showed close to identical measures as a lab group consisting of labs two through four with moderate positive relationships and 34% data to model conformance.

FIGURE 4
CORRELATION BETWEEN SCORES ON LABS AND FINAL COURSE GRADES



CONCLUSIONS

Difficulties in introductory programming classes (i.e., CS1) are known to affect both immediate student performance in the course as well as long term degree retention. There has been a plethora of investigations into root causes for this phenomenon (Ihantola et al., 2015; Quille & Bergin, 2019) as well as explorations of potential interventions (Sobral, 2020) and alternative pedagogies which might address the situation (Salguero et al., 2020). Given that the problem has been studied since the mid-1980s (Butcher & Muth, 1985) and still consistently plagues undergraduate computer science education (Bornat & Dehnadi, 2008; Gorson & O'Rourke, 2020), there is merit in continuing to study the problem.

Accordingly, the goal of this research was to find the earliest point in the course assessment sequence it might be possible to predict final grade outcomes. If such points exist, targeted intervention may potentially lead to reduced course failure or abandonment rates as well as increased degree retention. To that end, we measured the degree of correlation between student performance on quizzes, labs, programs, and tests compared to final course grade. This work presented an analysis of 197 students over 6 semesters from 11 sections of an introductory freshman-level programming class.

The analysis revealed positive correlations across four assessment categories (tests, programs, quizzes, and labs). Further, we discovered that each assessment category contained an instrument which is administrated before the normal mid-semester intervention point which significantly correlated with final course grades. Overall, the results show that as the degree of relationship between both tests and Final Grade was very similar to considering Test 1 alone, we conclude that Test 1 may be the best and earliest tool to predict final grade. As Test 1 typically occurs early in week 5 of a given semester, using this indicator allows for sufficiently early intervention and adequate time for a course correction on the part of the student. The next intervention opportunity based on the results would be after Program 1 and Program 2 are completed which corresponds with the middle of the semester.

Overall, we speculate that potential early interventions might include an improvement in study strategy or time management skills or may utilize tutors or other resources, an investigation left for future work. With appropriate intervention at this early point, a student can improve their performance on subsequent assessments. As one test can generally be replaced by a student's final exam grade if they earn a higher score, their score on the first test might not weigh into their final grade at all. Interestingly, only 9% of the students in the study scored lowest on the first test, replacing that test with their final exam grade. The average Test 1 dropped grade across the 11 sections was 77%, as compared to an average Test 3 dropped grade of 60%. The implications of this are left for future investigations. Improvement on subsequent assessments builds towards their success in the course and in the program, thereby improving retention.

Lastly, while it might be possible to dismiss these findings as an epiphenomenon of the grading structure, more interestingly, a deeper analysis of the data reveals the legitimate potential to identify specific opportunities for early intervention solely based on data prior to the standard midterm grading period. These results hold significance for educators and researchers by extending the body of computing education research through an evaluation of the relative effectiveness of employing early semester subsets of each of the four categories of student data to model class final grade outcomes. There is a compelling need for these types of tools, especially at larger institutions, where extremely large class sizes in introductory programming classes often mean that a professor has very little personal contact with many students enrolled in the class.

REFERENCES

Bornat, R., & Dehnadi, S. (2008). Mental models, consistency and programming aptitude. In *Proceedings of the tenth conference on Australasian computing education*, *78*, 53–61.

Butcher, D.F., & Muth, W.A. (1985). Predicting performance in an introductory computer science course. *Communications of the ACM*, *28*(3), 263–268. https://doi.org/10.1145/3166.3167

Gorson, J., & O'Rourke, E. (2020). Why do CS1 Students Think They're Bad at Programming?: Investigating Self-efficacy and Self-assessments at Three Universities. *Proceedings of the 2020 ACM Conference on International Computing Education Research*, pp. 170–181. https://doi.org/10.1145/3372782.3406273

Guo, A. (2020). *Analysis of Factors and Interventions Relating to Student Performance in CS1 and CS2*. Retrieved from https://www2.eecs.berkeley.edu/Pubs/TechRpts/2020/EECS-2020-22.pdf

Ihantola, P., Rivers, K., Rubio, M.Á., Sheard, J., Skupas, B., Spacco, J., . . . Petersen, A. (2015). Educational Data Mining and Learning Analytics in Programming: Literature Review and Case Studies. *Proceedings of the 2015 ITiCSE on Working Group Reports - ITICSE-WGR '15*, pp. 41–63. https://doi.org/10.1145/2858796.2858798<

Leu, K. (2017). *Beginning College Students Who Change Their Majors within 3 Years of Enrollment*. Data Point. NCES 2018-434. National Center for Education Statistics.

Quille, K., & Bergin, S. (2019). CS1: How will they do? How can we help? A decade of research and practice. *Computer Science Education*, *29*(2–3), 254–282. https://doi.org/10.1080/08993408.2019.1612679

Salguero, A., McAuley, J., Simon, B., & Porter, L. (2020). A Longitudinal Evaluation of a Best Practices CS1. *Proceedings of the 2020 ACM Conference on International Computing Education Research*, pp. 182–193. https://doi.org/10.1145/3372782.3406274

Sobral, S.R. (2020, July 17–19). CS1 student grade prediction: unconscious optimism vs insecurity? In *4th International Conference on Education and Distance Learning Conference (ICEDL2020)*. Roma, Italia. Retrieved from http://hdl.handle.net/11328/3147

Werth, L.H. (1986). Predicting student performance in a beginning computer science class. *ACM SIGCSE Bulletin*, *18*(1), 138–143.