Predictors of Student Performance in Introductory Statistics Textbook Quizzes

Adani Abutto
aabutto@stanford.edu
Stanford University
Stanford, CA, USA

Abstract

Statistics literacy is key to comprehension of STEM subjects and other empirical disciplines. The present study examined U.S. college students' performance on quizzes embedded in an introductory statistics text book. More specifically, based on data sourced from five introductory statistics classes at one U.S. college, I investigated what predicts variability in individual student performance on assessments at the end of different statistics textbook chapters. Results suggested that student performance tended to decrease with the advancement of textbook content, revealing lower average performance on more advanced chapters. Moreover, performance on preceding chapters predicted performance on subsequent chapters, indicating that content comprehension compounded. Interestingly, results also revealed that students taking more attempts on textbook chapters tended to perform worse, suggesting that students who struggled with the textbook content tended to continue to struggle despite increased effort and engagement.

ACM Reference Format:
Adani Abutto. 2025. Predictors of Student Performance in Introductory Statistics Textbook Quizzes. In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Statistics literacy is key to comprehension of STEM subjects and other empirical disciplines. Recent developments in methods for teaching statistics as well as the application of routines and practices from open software development and improvement science have launched the "Better Book" approach as presented by Stigler et al. (2020). The Better Book approach enables (1) iterative curriculum design based on student feedback, and (2), large-scale data collection by use of course content and assessment quizzes embedded in an interactive online statistics textbook. The present study draws on data generated by this "Better Book" approach to evaluate student performance and student experience in introductory statistics classes at a U.S. college.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 Methods

The sample comprised N = 570 U.S. adult participants from five introductory statistics classes taught at one U.S. college. Not all students answered all questions in all chapters, so we report a range of N = 526 to 562 responses, depending on question.

2.1 Data and Measures

The data were obtained from students working through digital statistics books (two books, five versions). The books were highly similar in content, encompassing the same nine chapters. Students encountered a series of quiz questions while working through the books in question. While student performance was scored across chapters, students were not required to complete all nine chapters or the questions embedded in them but could freely skip select quiz items. Students' responses to individual questions, when available, were scored as correct (1) or incorrect (0). The dataset also contained additional information on students (student institution, student classroom), the textbook content (textbook title, textbook version, item IDs, page numbers), and question-related aspects (points possible, points earned, number of attempts, date and time of submission, whether all questions on a page were completed).

2.2 Inferential Statistical Analyses

My unit of analysis for this study was individual student performance (i.e., singular correct/incorrect responses) on individual quiz questions within each of the nine book chapters. My inferential statistical analyses assess the predictive power of three factors that may affect student performance at this level: Engagement (number of attempts for a given question in a given chapter), advancement of content (the location of the question's corresponding book chapter, whereby later chapters were more demanding), and student performance on the respective preceding chapter (proportion of correct responses within that chapter).

To examine how much these sources of variance were systematically related to variance in individual student performance at the question level, I ran mixed-effects logistic regression models. More specifically, I regressed performance (correct vs. incorrect response on a given question within a given chapter) on advancement of content (model 1), individual student engagement (number of attempts for a given question; model 2), individual student performance on the preceding chapter (model 3), and a combined model with all of the abovementioned predictors (model 4). In all models, I included the respective predictor plus book version as fixed effects, and random intercepts per classroom and individual student. I then conducted model comparisons to contrast how each model performed relative to the other models.

For each model, I report Chi-squared, degrees of freedom, unstandardized regression coefficients, standard error (SE), and p-values. Additionally, I report model metrics AIC, BIC, and unstandardized RMSE for comparability. All of the above was done using the R packages *lme4* (Bates et al., 2003), *ggeffects* (Lüdecke, 2020), *broom.mixed* (Bolker et al., 2019), and *emmeans* (Lenth, 2018).

3 Results

3.1 Descriptive Statistics

The average proportion of correct responses across all textbook chapters was 80.3%, with an average standard deviation (SD) of 13.1%. The table below shows mean % correct, SD % correct, min % correct, and maximum % correct per textbook chapter (collapsed across the two textbooks).

Table 1: Means, SDs, response counts, and SEs for student performance (% correct) per book chapter

Chapter	Mean	SD	*n*	SE
1	0.978	0.051	562	0.002
2	0.854	0.106	556	0.004
3	0.841	0.118	540	0.005
4	0.792	0.130	532	0.006
5	0.800	0.135	537	0.006
6	0.759	0.149	536	0.006
7	0.733	0.151	536	0.007
8	0.744	0.166	535	0.007
9	0.714	0.170	526	0.007

Figure 1 gives a visual overview of student performance per chapter and book. Individual students' performance was averaged across questions within a given chapter to yield a single % correct estimate per chapter, and individual values were plotted as individual dots for each chapter. Individual data points are also colored by book to visualize any possible differences in student performance across book versions. Overall, we see that the proportion of correct responses per chapter ranged from near ceiling (98%) in chapter 1 to about 70% on the last chapter. Thus, the visual trend suggests that student performance tended to decline with chapter number (i.e., more advanced chapters were harder for students). Notably, the spread in individual variability also increased with chapter number (i.e., we observed more variability in later, more advanced chapters).

I also computed 3 engagement metrics: (1) what proportion of questions a given student completed in full on a given chapter page, (2) how many attempts the student took for a given question in a given chapter, and (3), how long a given student took to answer a given question within a given chapter. Figure 2 visualizes the variability on these different measures. The proportion of fully completed question sets was low, ranging from 12.6 to 27.6%. In other words, most students did not complete all the questions on a given question page of a given chapter. The average amount of time students spent on a given question within a given chapter ranged from 9 to 14.2 minutes. The standard deviation on this measure was quite large, ranging from 12 to 26 minutes depending on book chapter. The mean number of attempts individual students made on individual question was consistently close to 1 (with the mean ranging from 1.08 to 1.23 depending on book chapter, and SDs ranging from .08 to .34).

3.2 Inferential Statistical Analyses

3.2.1 Model 1: Book Chapter Progress Predicting Student Performance. To examine whether performance indeed decreased with advancement of content (i.e., lower % correct on later chapters) as suggested by Figure 1, I examined the student data from both books (193,312 responses from N = 570 students). I predicted that, on average, progression in book chapters would negatively predict proportion correct, because later book chapters were more advanced (i.e., harder). The results of the mixed-effects logistic regression supported this, with the predicted % correct decreasing across book chapters: χ^2 = 4976.56, df = 1, unstandardized β = -.169, z = -70.55, SE = .002, p < .001; RMSE = .0.382. This effect showed no difference by book version.

3.2.2 Model 2: Student Engagement Predicting Student Performance. Out of the three student engagement metrics I computed (see above), I focused on regressing student performance on the number of attempts made on a given individual question. Number of attempts is a continuous variable with a straightforward interpretation: The more attempts made for a given question, the greater a given student's effective engagement with the respective question. I predicted that with greater engagement (i.e., more attempts made), student performance would increase. However, the results of the mixed-effects logistic regression contradicted this, whereby % correct tended to decrease with higher average number of attempts made: χ^2 = 5014.04, df = 1, unstandardized β = -.041, z = -6.95, SE = .006, p < .001; RMSE = .0.3818. However, the size of this effect was small, and it again showed no difference by book version. Figure 3 visualizes average model predictions for 1, 2, and 5 model attempts (the average number of attempts was close to 1).

3.2.3 Model 3: Preceding Chapter Performance Predicting Concurrent Performance. Lastly, I regressed students' performance on a given chapter on their performance in the preceding chapter. I predicted that greater performance on a preceding chapter would relate to greater performance on the concurrent chapter. The results of the mixed-effects logistic regression supported this, whereby % correct tended to decrease with higher average number of attempts made: χ^2 = 740.59, df = 1, unstandardized β = 1.01, z = 27.21, SE = .037, p < .001; RMSE = .0.3816. This effect also showed no difference by book version.

3.2.4 Model 4: Combining Predictors of Student Performance. Given the statistical significance of the tested predictors, in my final model, I combined all three regressors in a single model (on top to random effects; see Methods). This served to examine whether each predictor of interest explained variance in student performance when holding all else constant. Indeed, the results of the mixed-effects logistic regression confirmed this, with all predictors remaining significant: Book chapter progress: unstandardized β = -.103, z = -31.15, SE = .003, p < .001; number of attempts: unstandardized β = -.042, z = -7.10, SE = .006, p < .001; Performance on previous chapter: unstandardized β = .931, z = 13.85, SE = .067, p < .001; Model RMSE = 0.381. These effects again showed no difference by book

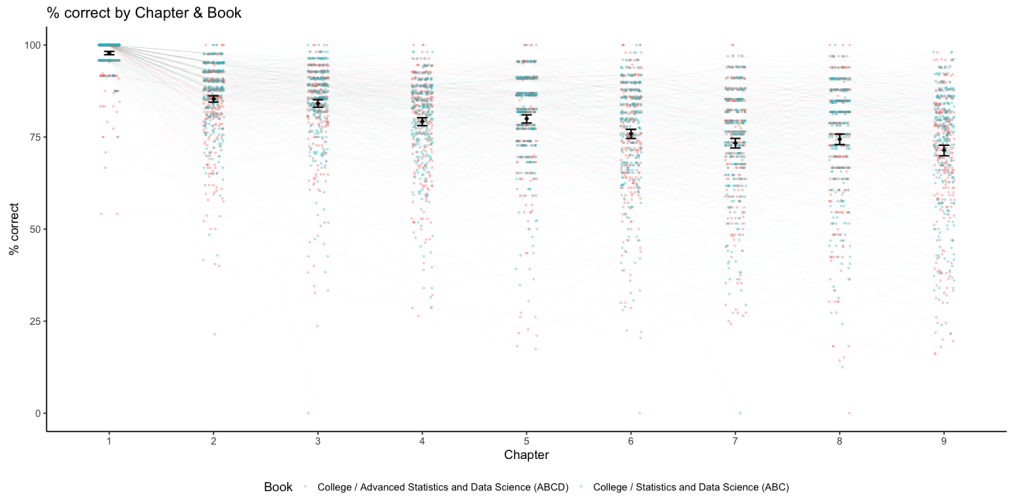

Figure 1: Average Student Performance by Book Chapter

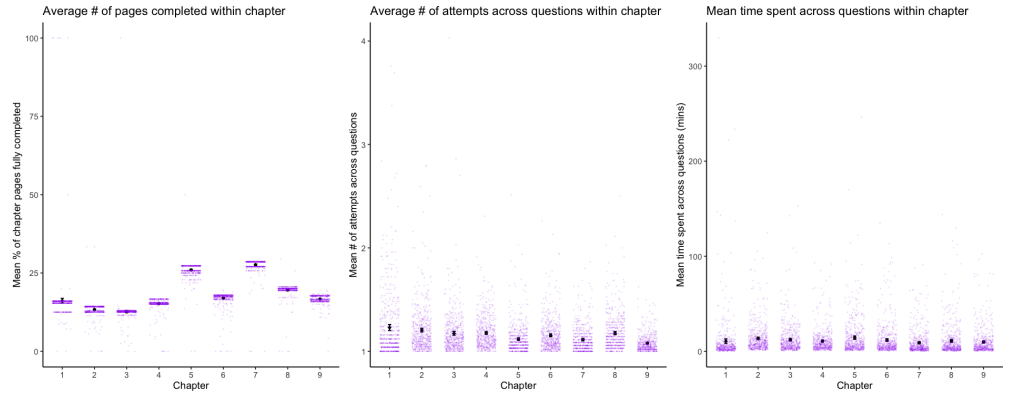

Figure 2: Student Engagement Metrics by Book Chapter

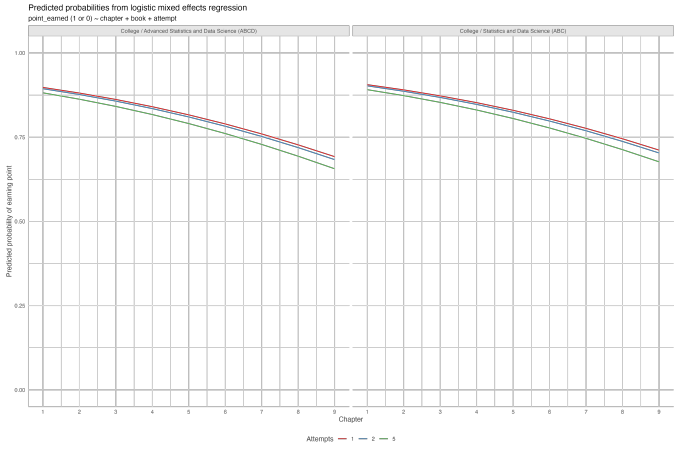

Figure 3: Model 3 Predictions

version. The table below shows model metrics compared across all four models, showing that while RMSE is minimally lower for the full model, AIC and BIC rank the simpler models more favorably (standardized; 0 is best, 1 is worst).

Table 2: Model-Fit Indices by Model

Model	AIC (std.)	BIC (std.)	RMSE
Model 1	0	0	0.38190
Model 2	0	0	0.38186
Model 3	0	0	0.38158
Model 4 (Full)	1	1	0.38054

4 Discussion

Statistics are hard to teach and learn. Researchers and practicioners are testing new approaches to helping students build statistics literacy and assessing their progress. Here, I evaluated what predicts student performance on quiz questions embedded in a digital introductory statistics textbook. Results showed that more advanced textbook content in later chapters tended to be harder for students on average. Moreover, students who struggled on earlier chapters also tended to continue struggling on subsequent chapters; conversely, students who did well on earlier chapters tended to continue doing well on subsequent chapters. Interestingly (and perhaps sadly), the number of attempts students took at a given question did not help their chances of getting that question correct.

4.1 Limitations and Future Directions

The present approach was limited in various ways. For one, I assessed only one simplified metric of student engagement (number of attempts). Other possibilities would be to have a more granular look at how long students actively view and engage with a page (e.g., by obtaining click-level data). I also did not perform an outlier analysis; some students performed a very large number of repeated attempts which may reflect a technical issue rather than true attempts, and these data points likely skewed the model estimates presented here. Another avenue for taking a closer look at the merits (or downfalls) of repeated attempts is to analyze nonlinear relationships. For instance, it is possible that repeated attempts are indeed helpful up until a certain point (e.g., trying another 2 or 3 times), whereas repeated attempts past that point may be accompanied by significant demotivation or disengagement or other student-related issues that were not captured here.

Secondly, at the analysis and modeling level, I did not perform growth modeling. Growth curves, for example as used by Koedinger et al. (2023), would likely more accurately capture how performance on preceding chapters predicts performance on later chapters. Future research should compare model fit of such an approach to the model fit of the approach employed here.

Finally, these results are limited in generalizability. The students in the dataset used here were recruited from one specific U.S. college. With this, all estimates rest on a convenience sample. It is unclear whether and how the present findings apply to other student groups, cultures, and languages. Future work should expand its scope accordingly.

5 Code and Data Availability

All analysis scripts are publicly available in the following GitHub repository: https://github.com/adaniabutto/mini-project-2.

References

[1] Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker. 2003. lme4: Linear Mixed-Effects Models using 'Eigen' and S4. doi:10.32614/CRAN.package.lme4

[2] Ben Bolker, David Robinson, Dieter Menne, Jonah Gabry, and Paul Buerkner. 2019. Package "broom. mixed". (2019).

[3] Kenneth R. Koedinger, Paulo F. Carvalho, Ran Liu, and Elizabeth A. McLaughlin. 2023. An astonishing regularity in student learning rate. *Proceedings of the National Academy of Sciences* 120, 13 (March 2023). https://www.pnas.org/doi/abs/10.1073/pnas.2221311120 Publisher: Proceedings of the National Academy of Sciences.