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 student performance for 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, 5 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 both iterative curriculum design based on student feedback and 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.

2 Methods

Our sample comprises N = 570 U.S. adult participants from five introductory statistics classes taught at one U.S. college. For cross-validation purposes, the participant data was split into two samples: Training (80%) and test (20%).

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 then 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 on a question in any given chapter), advancement of content (the location of the question's corresponding book chapter, whereby I assumed later chapters were more demanding), and student performance on the respective preceding chapter (i.e., across all questions within that chapter). To examine how much these sources of variance were systematically related to variance in question-level student performance, I ran mixed-effects logistic regression models. More specifically, I regressed performance (correct vs. incorrect responses on questions within a given chapter) on advancement of content (model 1), student engagement (average number of attempts across chapter questions; model 2), performance on the preceding chapter (model 3), and a combined model (model 4). For 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.

Conference'17, Washington, DC, USA

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

Additionally, I report model metrics AIC, BIC, and unstandardized RMSE as model metrics 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 value, and individual values were then plotted per chapter. Individual data points are also colored by book to highlight 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) how many chapter pages were fully completed in a given chapter, (2) how many attempts student took, on average, across questions within a chapter, and (3), how long students took, on average, to answer questions within a chapter. Figure 2 visualizes the variability on these different measures. The proportion of full pages completed was low, ranging from 12.6 to 27.6%. In other words, most students did not complete all questions on a given 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 students made on a given 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), I examined the student data from both books (193,312 responses from N = 570 students). In line with what Figure 1 indicated, I predicted that, on average, progression in book chapters would negatively predict proportion correct. 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 (proportion of book chapter pages fully completed, average number of attempts taken across questions within chapter, and average time spent on questions within chapter), I focused on regressing student performance on average attempts made across questions. Number of attempts is a continuous variable with a straightforward interpretation: The more attempts made on various questions within a given chapter, the greater a given student's effective engagement with the respective question(s). 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 = .0.382. Notably, however, the regression coefficient effect size was small. This effect showed no difference by book version. Figure 3 visualizes average model predictions for 1, 2, and 5 model attempts:

3.2.3 Model 3: Preceding Chapter Performance Predicting 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 = -.041, z = -6.95, SE = .006, p < .001; RMSE = .0.382. Notably, however, the regression coefficient effect size was small. This effect also showed no difference by book version.

3.2.4 Model 4: Combining Predictors of Student Performance. Given the predictive power of all of the above predictors, in my final model, I combined all three regressors (on top to random effects; see Methods) in a single model. 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. RMSE = 0.381. These effects again showed no difference by book version. The table below shows model metrics compared across all four models,

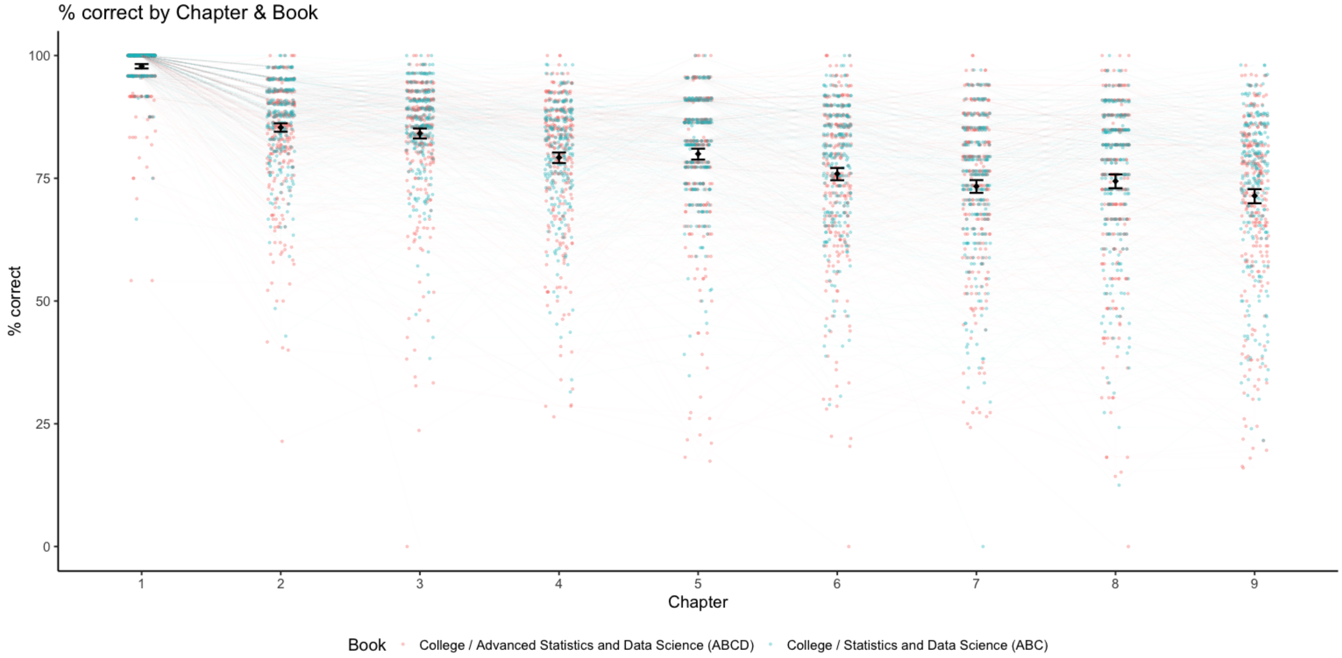

Figure 1: Average Student Performance by Book Chapter

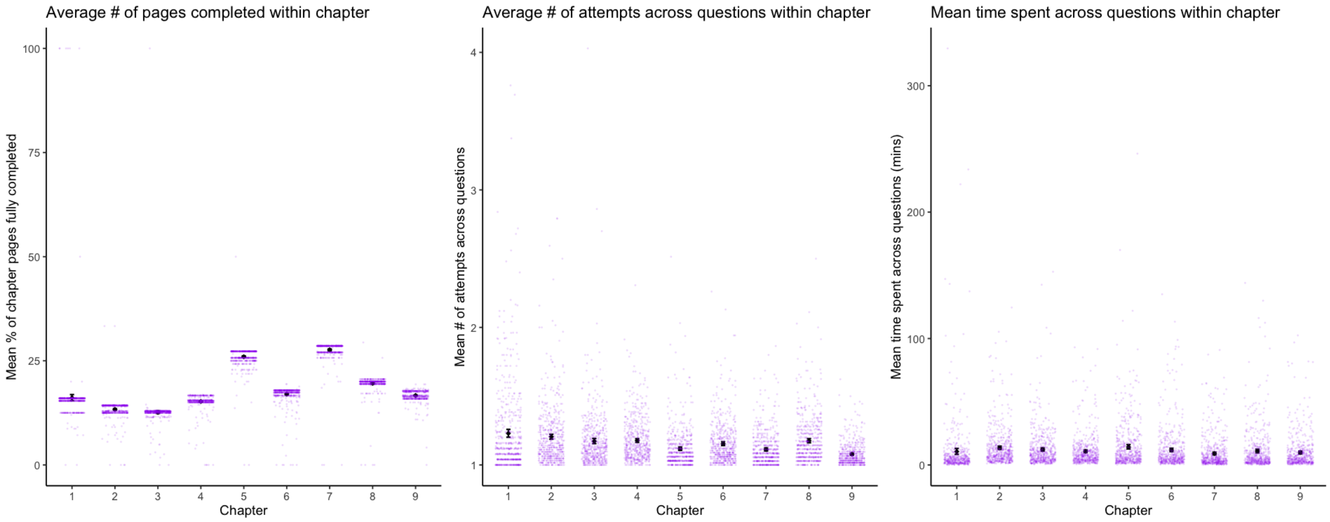

Figure 2: Student Engagement Metrics by Book Chapter

showing that while RMSE is lowest for the full model, AIC and BIC (standardized) rank the simpler models more favorably.

4 Discussion

Assessments of data visualization literacy often present the test-taker with a series of items entailing a question and various graphs to answer said question (Brockbank et al., 2024). Here, I examined if five widely-used data visualization literacy assessments pass a "sanity check" on two fronts: 1) Are the test items the designers intended to be more difficult indeed more difficult as per the test-takers' performance? And 2), are the tests (reasonably) safe from low-literacy test-takers scoring highly through mere chance/guessing? The answer to both questions was yes: Two common-sense predictors of item difficulty—question length and item nature—predicted item-level performance. Additionally, my model simulating a low-literacy

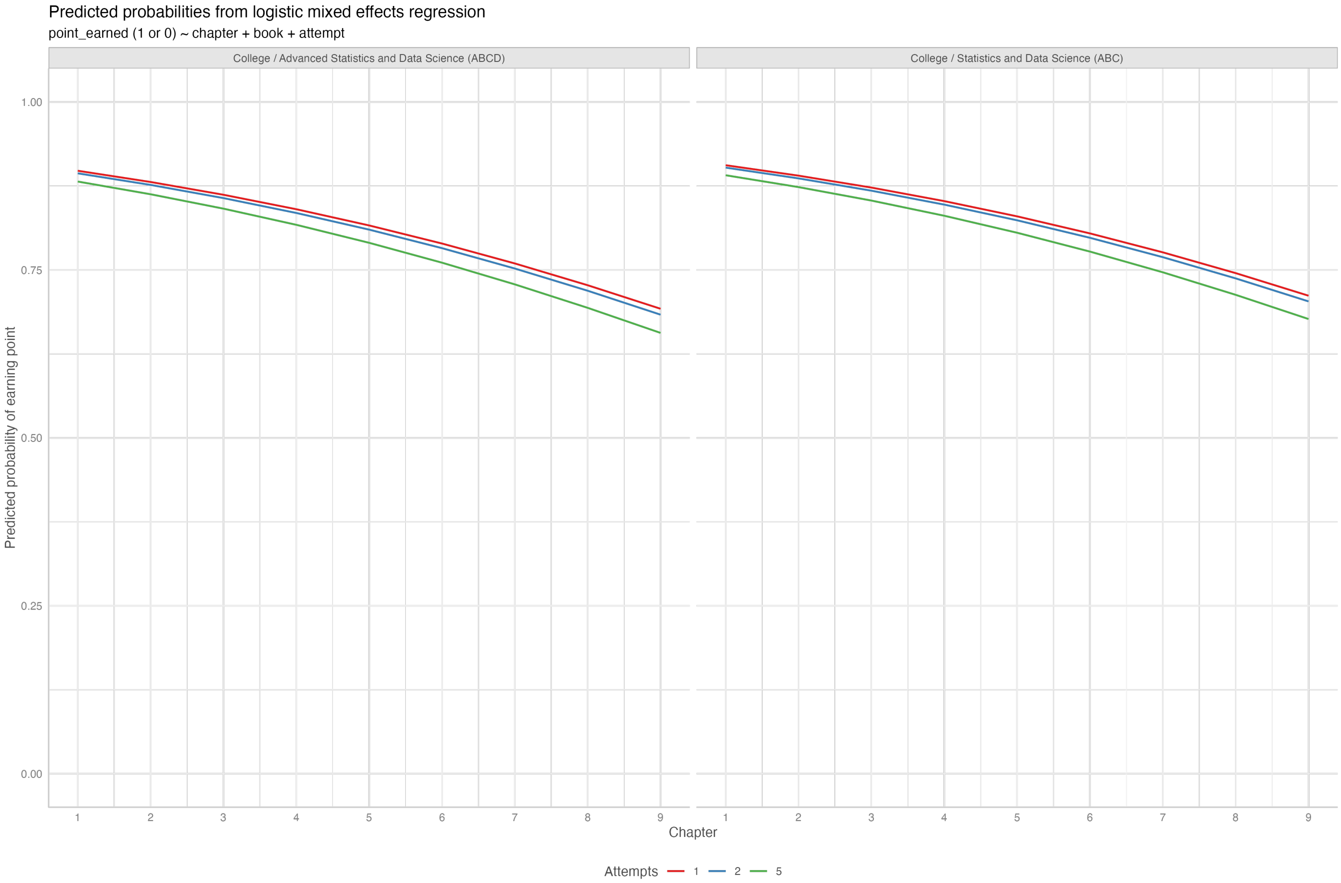

Figure 3: Model 3 Predictions

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

subject who followed a exclusion-and-guessing strategy did not align well with the observed data.

4.1 Limitations and Future Directions

The present approach was limited in various ways. At the analysis and modeling level, I did not perform nested model comparisons or mixed-effects multiple linear regressions to assess the predictive power of my ’common-sense’ predictors while holding the other predictor(s) constant, including predictors found to be significant by previous work (e.g., graph type and task type of the item; cf. Brockbank et al., 2024). Moreover, my ’exclusion-and-guessing’ model had hard-coded parameters that were not empirically derived but rather a best guess for what a true test-taker with such a strategy could plausibly be characterized as. Future research should look into an empirical approach to setting these parameters to increase the robustness of this ’sanity check.’

Secondly, for the IRT analyses, I exclusively fit a 2PL model. Various literature discusses the advantages and drawbacks of such a model compared to a more traditional 1PL (Rasch) model and multi-parameter models of other kinds. Future research should fit a variety of models and conduct nested model comparisons to assess which one best captures the structure underlying the data on the data visualization literacy assessments used 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/datasci294l_mini-project-2/tree/main/code.

References

[1] Margaret Wu Alexander Robitzsch, Thomas Kiefer. 2013. TAM: Test Analysis Modules. doi:10.32614/CRAN.package.TAM Institution: Comprehensive R Archive Network Pages: 4.2-21.

[2] 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

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

[4] Jeremy Boy, Ronald A. Rensink, Enrico Bertini, and Jean-Daniel Fekete. 2014. A Principled Way of Assessing Visualization Literacy. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (Dec. 2014), 1963–1972. doi:10.1109/TVCG.2014.2346984

[5] Mirta Galesic and Rocio Garcia-Retamero. 2011. Graph Literacy: A Cross-Cultural Comparison. *Medical Decision Making* 31, 3 (May 2011), 444–457. https://doi.org/10.1177/0272989X10373805

[6] Lily W. Ge, Yuan Cui, and Matthew Kay. 2023. CALVI: Critical Thinking Assessment for Literacy in Visualizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–18. doi:10.1145/3544548.3581406

[7] Sukwon Lee, Sung-Hee Kim, and Bum Chul Kwon. 2017. VLAT: Development of a Visualization Literacy Assessment Test. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (Jan. 2017), 551–560. doi:10.1109/TVCG.2016.2598920

[8] Dimitris Rizopoulos. 2005. ltm: Latent Trait Models under IRT. doi:10.32614/CRAN.package.ltm Institution: Comprehensive R Archive Network Pages: 1.2-0.