EVALUATING THE RELATIONSHIP BETWEEN PRACTICE AND ACTUAL EXAMS IN AN INTRODUCTORY PSYCHOLOGY COURSE

Thomas Brothen[1], Jonathan Brown[1], Robert delMas[1]
[1]University of Minnesota (United States of America)

Abstract

The value of practice exams has been assessed for over three decades and researchers have made several important findings that show students learn more if they get feedback on their knowledge and how much additional study they need to do. Practice exams serve that purpose efficiently and effectively, with practice exam scores correlating highly with actual exam scores. This study provides information to instructors about the effectiveness of practice exams and how students can use them to improve their course performance. In this report, we focus on two questions. First, what practice exam taking strategies (e.g., how many or on what schedule) are associated with better student performance on actual exams? Second, are there interesting and useful to know interactions with measures of personality and ability that indicate specifically who will benefit more? We used a very large data set in this study to address these questions. In addition, we describe how our data suggests optimal test taking strategies for students. Finally, we describe the statistical procedures we used so as to assist others doing similar research.

Keywords: practice exams, testing, learning strategies

1 INTRODUCTION

A recent article in the Education section of the New York Times Magazine [1] reported on research and innovations in practice exams and student learning. The author described unpublished research by Elizabeth Ligon Bjork suggesting that giving on-paper practice exams to students before the semester began improved their exam scores at the end of the semester. The use of this specific technique may or may not be good advice to instructors but the zeitgeist is surely moving in the direction of providing practice tests to improve student learning—especially in online and hybrid courses. The category "practice test" includes both quizzes and exams but we focus on exams in this paper.

Much research has been done in various disciplines describing the uses and benefits of practice tests (e.g., [2], [3], [4], [5], [6], [7], [8], [9]). The value of practice tests administered in class and graded by instructors has been supported by research for over three decades and researchers have made several additional important findings during that period of time. Generally, practice exam scores predict exam scores quite well with correlations in the .70 range [4].

For example, in his overview of best practices for review sessions to enhance students' performance, Gurung [5] noted that instructors are reluctant to take class time away from lecture or discussion for study sessions. As an alternative, he reviewed relevant research and suggested that such exams are best "given in the same manner as the actual exam" (p. 2). This suggestion was supported by the study of Wenger, Hobbs, Williams, Hays and Ducatman [10], in which students were asked to rate relative time they used on various resources to prepare for exams. Their survey showed that students believed that practice exams that helped to reinforce course materials were useful for exam preparation. Gurung also cited Balch's [2] study of two groups of students who prepared for the final exams by either taking practice exams as if they were actual exams or only reviewing an answer keyed copy. Compared to the review group, the actual exam-taking group rated their activity more helpful and scored significantly higher on the final exam. Gurung noted that another advantage of practice exams is they can be used outside of class time. As another example, Oliver and Williams [11] cited research backing the proposition that practice exams have a positive effect on students' subsequent exam performance. In addition, their research showed that students who try to do well on the practice exams scored better on the subsequent exam than those students using the practice exam simply as a review.

The above research indicates that practice exams are most effective pedagogically when they give students a general sense of how prepared they are and also if they take them several times and use the feedback to guide their overall studying. Since 1998, the Internet and course management systems (CMS) have made this approach much more feasible. One would think students would be enthusiastic about using them to review the course material to get a deeper understanding of it and guidance for

further study. But any college instructor knows that just because this may well be the best course of action, it is unlikely to be universally adopted by students. In this study, we focus on two general questions. First, what practice exam taking behaviors (e.g., how many do they take or on what schedule) are associated with better student performance on actual exams? Second, are there meaningful interactions with measures of personality traits and academic ability that indicate specifically who will benefit more? For a single class, this may be difficult to determine if the class sizes of students possessing specific characteristics is small. The very large data set we used in this study provides an opportunity to address these questions and provide significant and reliable answers.

In this paper, we present and discuss results that bear on the above questions. In addition, we describe how our results suggest optimal test taking strategies. Finally, we describe our data management procedures to assist others doing similar research.

2 METHODOLOGY

Data for this study came from a large introductory psychology course taught in three formats each semester to approximately 1000 students via the Moodle CMS. The first format employed live lectures three days each week and the second utilized recorded online lectures. Both of these formats had live discussions once each week. Lectures and discussions were completely online in the third format. Students in all three formats completed practice exams and other activities online and took actual exams in a computerized testing center monitored by proctors. Practice exams were available any time on any Internet-connected computer before and during the actual exam period. Each practice exam session consisted of a new random selection of items from a large item pool. Students could do as many practice exam sessions as they liked prior to taking an actual exam. Their performance on the practice exams did not count toward the final course grade.

Analyses reported in this paper focused only on the final exam. Information on practice and actual final exams were recorded by Moodle. The recorded information was 1) the number of practice exams taken by a student prior to the actual final exam; 2) the date and time when each practice and actual final exam was taken; 3) the amount of time spent on each practice exam. Two measures of academic ability were obtained from university records; the composite score on the American College Testing Program (ACT) assessment and current cumulative college grade point average (GPA). Additionally, students' scores on the Big 5 Aspect Scales (BFAS) were collected from a class assignment on personality psychology. Data were aggregated over five years (10 semesters) of the course that was taught nearly the same each semester. The course enrollments of approximately 1,000 students a semester resulted in a database representing around 10,000 students.

The construction, format, and delivery for all exams were basically the same over the 10 semesters. First, all questions were multiple-choice with typically four response options and one correct answer. Second, for each exam, the sets of items that measured course content consisted of items of equivalent difficulty that had been established in the exam writing process and subsequent testing with students. Of the approximately eight items in each set, half were reserved for the practice exams and half for the actual exams. Moodle randomly selected items from the respective item pool for each exam. Practice and corresponding actual exams contained the same number of items. Third, students had unlimited attempts on the practice exams and one attempt on the actual exams. The practice exams provided feedback to students immediately after they submitted the exam.

As indicated, acquiring and processing the data for this analysis required combining information across multiple sources. Numerous data management and data quality issues were encountered during the data cleaning and combination process, which took a notable amount of time and effort to resolve. Broadly, these issues primarily involved understanding who was automatically included in the data tracked and provided by Moodle, defining which practice exam attempts were meaningful for this analysis, and the consistency of the data (missing data, consistency of variable measures) across variables and semesters.

In addition to student activity and scores, the behavior of any individual who was part of the course, regardless of role, was also tracked. Thus, the data for student practice exam attempts included some non-students, as well. To resolve this, a global list of students registered in the course was created from university data, outside of the Moodle data. While this eliminated non-students, there were rare cases where individuals present in the Moodle data appeared to be students, but for unknown reasons were not in the global list of students. Given the minimal instances of this occurrence relative to the final sample sizes for the analysis, any impact on results is presumably negligible.

Another challenging issue was determining which practice exam attempts were meaningful attempts. Distributions of the scores and time durations for practice exam attempts across semesters indicated a notable presence of extremely low scores and zero scores with durations on the order of seconds or a minute. Such low scores and times could represent numerous behaviors, such as viewing a single question out of curiosity without the intent to take a full practice exam or becoming distracted during an attempt and walking away.

To analyze practice exam behavior that involved earnest attempts to complete the whole practice exam, score and time criteria were applied. Nearly all questions had four response options, implying that randomly choosing answers would result in a score of 25%, on average. Thus, only attempts with scores above 25% were included. The minimum time taken to achieve a perfect score across all practice exams was three minutes. Therefore, only attempts with time durations of three or more minutes were included. Lastly, only practice exam attempts that preceded the corresponding actual exam were included. While it cannot be directly verified what was or was not an earnest attempt, the choice of these particular selection criteria were expected to minimize classification error. After applying these criteria, 15.5%, 13.4%, 11.5%, and 15.5% of attempts were excluded as “non-earnest” attempts for the first midterm, second midterm, third midterm, and final exam, respectively.

Other data consistency and missing data issues arose. For one, not all of those who took practice exams took the actual corresponding exam. Given the purpose of the analysis to explain variation in actual exam scores, this was resolved on a per-exam level by excluding any individuals who did not have a respective exam score. Thus, for a given semester students were only included in the data set for as many actual exam scores they had. The BFAS and other academic variables had varying levels of missing data resulting in varying sample sizes of complete data depending on which variables were considered in a given model. Finally, due to combining data across semesters and the opportunity for students to retake courses some students appeared in multiple semesters. Although this was not accounted for in the models, it represented a small number of instances relative to the sample sizes. Thus, any violation of independence from such repetition of students for a given model is presumably small.

To characterize the relationship between practice exam behavior and exam scores, bivariate correlations, linear models, and linear contrasts were explored. Analyses reported in this paper focused only on the final exam. The primary response variable was final exam score. Explanatory variables included best practice final exam score, first practice final exam score, last practice final exam score, number of earnest practice final exam attempts, GPA, ACT composite score, and the five BFAS variables. All analyses were executed in R version 3.4.4 [12] and all figures were produced with the ggplot2 package [13].

3 RESULTS

Initial review of bivariate correlations informed the model building process. Across the 10 semesters, 9,433 students had a final exam score. Due to missing data, analyses based on different combinations of variables resulted in sample sizes smaller than n = 9,433. Among all simple correlations with pairwise complete observations, best practice exam score resulted in the highest correlation with final exam score (r = .69, n = 6,967), while GPA closely followed (r = .68, n = 8,877). First practice exam score and last practice exam score were both correlated with final exam score at r = .6 (n = 6,967); however, they were also highly correlated with best practice exam score (r = .76 and .87, respectively; n = 6,967 for both). Thus, first and last practice exam scores were not considered for further use, in favor of choosing best practice exam score as an explanatory variable. ACT and number of attempts were modestly correlated with exam score (r = .46 and .42, respectively; n = 8,197 and 9,433, respectively). Lastly, all five BFAS variables were weakly correlated with exam score, ranging in magnitude from r = .02 to r = .13 (n = 6,227 for all BFAS variables).

Considering personality variables further, scatterplots between final exam score and each individual BFAS variable showed no discernable pattern. A linear model that regressed final exam score onto all five BFAS variables produced an R^2 of only 0.04. While the partial regression coefficients for three BFAS variables had *p*-values < 0.001 and the other two variables indicated marginal statistical significance (p = 0.04, 0.051, respectively), all coefficients were extremely small in the context of the scale. Consequently, BFAS variables were not included in subsequent models.

Two models were built to explain final exam score. Both models included the same set of explanatory variables except that one model also included best practice exam score whereas the other model also included number of attempts. Separate models were built because all students with zero attempts would

be excluded from any model with best practice exam score given that a minimum of one attempt is needed to have a best practice exam score. Because including students who never attempted a practice exam might account for variance in final exam scores, number of attempts was included in a separate model.

To identify a model with the appropriate form to represent the relationship between final exam score and each of best practice exam score and number of attempts, respectively, scatterplots were explored (see Figure 1 and Figure 2). A general additive linear model ('gam') smoothing function was added to each scatterplot to estimate the localized trend.

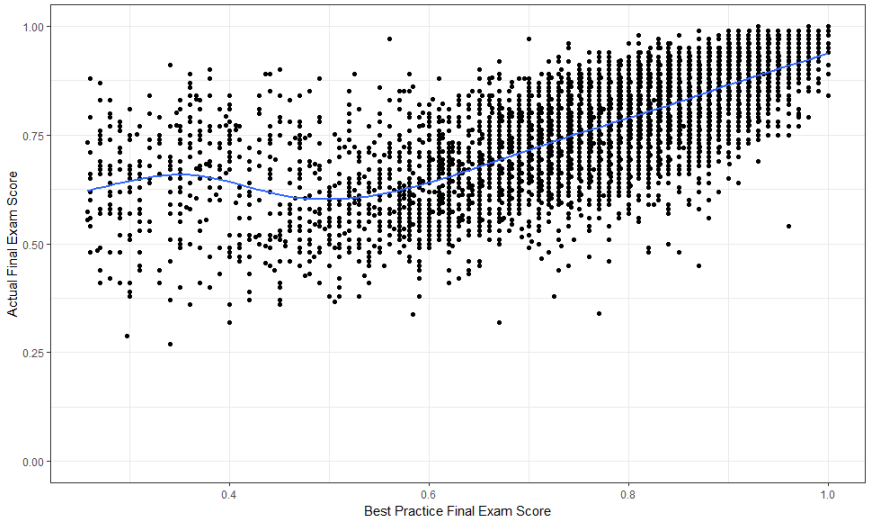

Figure 1. Scatterplot of actual by best practice exam score (proportion correct) for final exams across all semesters (n = 6,967). A generalized additive linear model ('gam') smoothing function is shown.

Figure 1 indicates there is a clear transition in the relationship around a best practice exam score of .55, suggesting two different linear relationships. For scores below .55, Best Score seems to have no relationship or even a weakly negative relationship. Above .55, the relationship appears sharply positive. Since this transition appears abrupt, a piecewise or spline modeling approach was employed, where two different slopes were applied to the best practice exam score variable depending on its value.

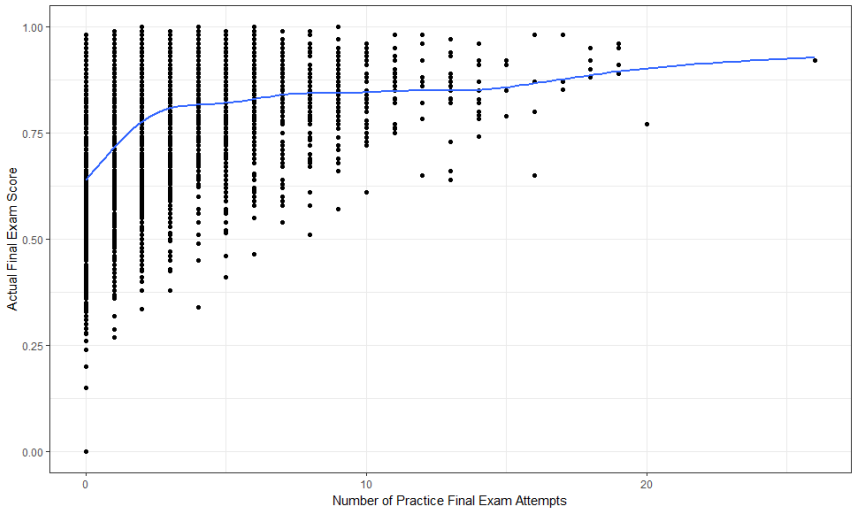

Figure 2. Scatterplot of actual exam score (proportion correct) by. number of practice exam attempts for final exams across all semesters (n = 9,433). A generalized additive linear model ('gam') smoothing function is shown.

Similar to Figure 1, the pattern in Figure 2 appears to abruptly shift at a specific value. From 0 to 3 attempts, a strong, positive relationship is apparent. However, for more than 3 attempts, the effect of

number of attempts seems to diminish. A spline model was built with number of attempts as the focal explanatory variable using a slope transition point at 3 attempts.

Two primary linear regression models employing splines were tested, with each predicting final exam score. For the first model, best practice exam score was the focal variable, which included a transition slope at .55, as well as the academic performance variables of GPA and ACT as covariates. The second model tested number of attempts as the focal variable, with a slope transition at 3 attempts, and also GPA and ACT as covariates. Results are displayed in Table 1 and Table 2, respectively.

Table 1. Linear model spline results predicting final exam scores for all semesters, with a knot at a best practice exam score of .55 (n = 5,762).

	Coefficient (SE)	t-score	p-value
Intercept	0.265 (0.015)	17.6	< 0.001
Best Practice Score < .55	-0.128 (0.023)	-5.5	< 0.001
Best Practice Score Adjustment > .55	0.623 (0.029)	21.6	< 0.001
GPA	0.073 (0.002)	33.6	< 0.001
ACT	0.008 (0.0003)	28.0	< 0.001
R^2 = .6654 F (4, 5757) = 2865 p < 0.001			

For the first model (Table 1), all explanatory variables were statistically significant and jointly explained 66% of the variation in final exam scores. Notably, the slope transition for the effect is of high practical significance. For best practice exam scores below 0.55, each additional 0.1 increase in score is associated with an average final exam score decrease of 0.13, controlling for GPA and ACT. However, above 0.55, there is a slope gain of 0.623, which results in a final slope of 0.495. Thus, a 0.1 increase in best practice exam score is associated with an average final exam score increase of approximately 0.05, controlling for GPA and ACT.

Regarding model quality, the residuals in both models are sufficiently normally distributed, but appear to suffer from some heteroskedasticity. In both cases, the residual variance decreases with increasing predicted values of final exam score, though this effect is stronger for the model with best practice exam score as the focal variable. This may potentially reduce the validity of inference from each model.

Table 2. Linear model spline results predicting final exam scores for all semesters, with a knot at 3 practice exam attempts (n = 7,719).

	Coefficient (SE)	t-score	p-value
Intercept	-0.001 (0.009)	-0.076	.939
Practice Attempts < 3	0.031 (0.001)	30.9	< 0.001
Practice Attempts Adjustment > 3	-0.028 (0.001)	60.0	< 0.001
GPA	0.118 (0.002)	36.8	< 0.001
ACT	0.011 (0.0003)	-19.8	< 0.001
R^2 = .6038 F (4, 7714) = 2941 p < 0.001			

For the second model (Table 2), all explanatory variables were statistically significant and jointly explained 60% of the variation in Exam Scores. The slope transition on number of attempts indicates a contextually meaningful change. For 0 to 3 attempts, each additional attempt is associated with an average exam score increase of .03, controlling for GPA and ACT. This effect dissipates after 3 attempts, where the slope decreases by .028 to .003. Thus, for each additional attempt after 3, there is only an average exam score increase of .002, controlling for GPA and ACT.

Based on the apparently meaningful relationship between final exam score and number of attempts, a grouping variable named “attempt group” was created based on the number of attempts: groups of 0 attempts, 1 to 3 attempts, and 4 or more attempts. While the spline model captures the transition from 0 to 3 to above 3, it was of interest to determine if those who did not attempt any practice exams differed systematically relative to those who did. A possible profile of these groups was explored using a set of linear contrasts. Attempt group was used as the explanatory variable for each of the following response variables; final exam score, GPA, ACT, and the five BFAS variables. Pairwise mean differences in each dependent variable were evaluated via three linear contrasts for each response variable (0 vs. 1-3; 0 vs. 4 or More; and 1-3 vs. 4 or more). Select results are presented in *Table 3.*

Results from modified Levene’s tests were examined to assess the assumption of homogeneity of variance. For the five BFAS variables, there was no evidence for violation of homoscedasticity. On the other hand, Exam Score, GPA, and ACT all resulted in statistically significant p-values, suggesting some level of heteroscedasticity. However, the largest to smallest variance ratios for these three academic variables are all below 3, which suggests limited concern for heteroscedasticity.

Table 3. Descriptive and linear contrast mean comparisons of academic and BFAS variables across practice exam attempt groups. Sample sizes for each set of comparison vary. Within a column, means with different letters are statistically significant ($p < 0.05$), whereas means with the same letter are not. The Benjamini-Hochberg method was used to adjust p-values within each column. (OI = Openness/Intellect, CO = Conscientiousness, EX = Extraversion, AG = Agreeableness, NE = Neuroticism.)

Attempt Group	Max Group Size	Exam Score	GPA	ACT	OI	CO	EX	AG	NE
4 or more	1872	.829[a]	3.49[a]	28.0[a]	3.50[a]	3.53[a]	3.56[a]	3.82[a]	2.82[a]
1 to 3	5095	.757[b]	3.29[b]	27.4[b]	3.57[b]	3.42[b]	3.51[b]	3.86[b]	2.88[b]
0	2466	.633[c]	2.86[c]	26.9[c]	3.62[c]	3.31[c]	3.47[c]	3.83[a]	2.89[b]

For all three academic variables, all groups significantly differed from each other ($p < 0.001$). From a group standpoint, those with higher numbers of attempts exhibit higher averages of final exam score, GPA, and ACT. For BFAS variables, statistical significance varied. Raw values of mean differences are again extremely small, indicating small practical differences in personality traits between groups. While small in raw value, there are nonetheless statistically significant differences in mean Conscientiousness and Extraversion values favoring the higher attempt groups, and statistically significant differences in mean Openness/Intellect values favoring the 0 attempts group.

4 CONCLUSIONS

In this study, we explored how students’ practice exam taking behaviors are associated with their performance on actual exams and also whether there are meaningful relationships between personality traits and academic ability. Our goal was to suggest to instructors and students how best to use practice exams. Consistent with previous research, we found a strong relationship between practice exam taking and actual exam scores and this basic relationship held when we controlled for academic ability and personality differences.

We found little effect for personality differences, which suggests that all students will benefit from practice exams. Obviously, students differ in many other ways but our results indicate that personality differences do not have a strong effect on the practice/actual exam relationship. We did find differences related to academic ability however. Table 3 shows that students taking no practice exams had lower final exam scores and were notably lower in their GPAs—which might not be surprising if they typically do not make use of good study methods. Although their slightly lower ACTs can explain some of that, our data suggests that their lower achievement is at least partly due to less than optimal study behaviors.

In addition to a conclusion that practice exams are good learning tools for instructors to assign and students to use, our data from nearly 10,000 students over several semesters suggested some important ways in which students should use them. First, as they make serious attempts at doing them, students should persist until they reach a reasonable level of achievement. Our data found an

inflection point at 55% correct—when students reach that level, more work on the practice exams is associated with increasing actual exam scores. Second, students need to do them more than once. Our analysis showed that three serious attempts was associated with higher actual exam scores. Taking practice exams more than that helped, but with diminishing returns.

There are more relationships for us to explore in our very large data set and we intend to pursue them to draw a clearer picture of how students can make best use of practice exams. For now, we can state with some confidence that students should use them and if they want to be as efficient as possible, they should take them until their scores are reasonably good and should take at least three of them before taking their actual exams.

ACKNOWLEDGEMENTS

This paper was supported in part by National Science Foundation Grant NSF/IIS-1447788.

REFERENCES

[1] B. Cary, "Why flunking exams is actually a good thing," *New York Times Magazine,* 2014. Retrieved from http://www.nytimes.com/2014/09/07/magazine/why-flunking-exams-is-actually-a-good-thing.html

[2] W.R. Balch, "Practice versus review exams and final exam performance," *Teaching of Psychology,* vol. 25*,* pp. 181-185, 1998.

[3] T. Brothen, "Comparison of non-performers and high performers in a computer-assisted mastery learning course for developmental students," *Research & Teaching in Developmental Education,* vol. 13, pp. 69-73, 1996.

[4] T. Brothen, Z. Lv, and H. Bai, "Can We Develop Best Practice Guidelines for Online Practice Exams?" Paper presented at the Lilly Conference On Evidenced-Based Teaching and Learning - Newport Beach, California on February 19-22, 2015.

[5] R. A. Gurung, "Enhancing learning and exam preparation," *Observer,* vol. 21. Retrieved from https://www.psychologicalscience.org/index.php/publications/observer/2008/january-08/enhancing-learning-and-exam-preparation.html

[6] K.J. Knaus, K.L. Murphy, and T.A. Holme, "Designing chemistry practice exams for enhanced benefits: An instrument for comparing performance and mental effort measures," *Journal of Chemical Education,* vol. 86*,* pp. 827, 2009.

[7] J.A. Kulik, C.C. Kulik, and R.L. Bangert-Drowns, "Effects of practice on aptitude and achievement test scores," *American Education Research Journal*, vol. 21, pp. 435-447, 1984.

[8] W. H. Lee-Sammons, and K. A. Wollen, "Computerized practice tests and effects on in-class exams," *Behavior Research Methods, Instruments, & Computers,* vol. 21, no. 2, pp. 189-194, 1989.

[9] R. H. Maki, and M. Serra, "Role of practice tests in the accuracy of test predictions on text material," *Journal of Educational Psychology, 84*, no. 2, pp. 200, 1992.

[10] S. L. Wenger, G. R. Hobbs, H. J. Williams, M. Hays, and B. Ducatman, "Medical student study habits: practice questions help exam scores," *Journal of International Association of Medical Science Educators*, vol. 19, no. 4, pp. 170-172, 2009.

[11] R. Oliver, and R. L. Williams, "Direct and indirect effects of completion versus accuracy contingencies on practice-exam and actual-exam performance," *Journal of Behavioral Education*, vol. 14, no. 2, pp. 141-152, 2005.

[12] R Core Team, *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria, 2018. URL https://www.R-project.org/

[13] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, 2009.