Differential Assessment, Differential Benefit: Four-year Problem Roulette Analysis of STEM Practice Study

Noah Weaverdyck
Department of Physics
University of Michigan, USA
nweaverd@umich.edu

Dhayaa Anbajagane
Department of Physics
University of Michigan, USA
dhayaa@umich.edu

August E. Evrard
Department of Physics
University of Michigan, USA
evrard@umich.edu

ABSTRACT

Using five million responses to thousands of practice examination questions on an optional study service known as Problem Roulette, we explore subject-specific differences in assessment style, grade benefit from usage of the service, and differential features in study behavior and grade outcome by birth sex. Our study includes more than 20,000 students enrolled in eight terms of introductory courses in general chemistry, physics and statistics. Student responses in the space of accuracy and response time reveal domain differences; by these measures, physics problems are typically both more difficult and more complex. Grouping students by term-length practice volume, we find significant positive grade benefits to higher volumes of study in chemistry and statistics. Across all subjects, we find that females gain more grade benefit from high study volume than males. Female students also outwork males during prime study hours yet, on average, earn 0.13 ± 0.03 lower grade points in chemistry than males with the same response accuracy in practice, with null results in statistics and physics.

Author Keywords

Computer applications: physical sciences and engineering; Instructional technologies; Educational equity and fairness

CCS Concepts

•Applied computing → Computer-assisted instruction;

INTRODUCTION: EQUAL ACCESS TO PRACTICE STUDY

In large introductory courses, timed examinations comprised of compact, focused questions developed by the instructor or instructional team remain an important component of summative assessment at American universities [6, 7]. For the sake of economy these questions are typically machine-gradable. The importance of this material to students means that authentic versions of past exams can become a form of campus currency. Stories are told of student organizations amassing copies of exams used in key courses to recruit and support their members, a practice that fosters inequity in student learning [5].

While students are free to employ open educational resources to support their study, the cost of discovering material that fits the substance and style particular to their locally-authored exam is typically high, especially for novice learners. Differing methods of expression, for example, conveying information via a graph instead of an equation, or resetting the problem context from gravity acting on an apple to gravity acting on a planet, may be small changes to an expert, but will require extra cognitive effort from a novice to recognize and adapt to the change. Minimizing this friction through self-regulated practice is a pathway to mastery [9].

At the University of Michigan, the Problem Roulette (PR) study service [3] provides equal access to authentic content collected from introductory course examinations. PR began in 2011 to support a single physics class, with questions published as Google docs and accessed via a simple, open server. By 2019, the service grew to support thirteen courses in science, engineering and public health.

PR is useful to students because it simulates the real test and democratizes access to quality content. The roulette name intentionally reflects the high risk environment of timed assessments. A typical exam contains perhaps a dozen or more questions, each of which requires correct conceptual interpretation and creative solution through application of skills and knowledge learned previously. Especially when locally-authored, the perspective and style of questions may differ from the assigned textbook and homework.

The order in which a student encounters these questions often feels random, and may indeed be set that way by the instructor to deter cheating. Given this setting, a practice service serving randomly chosen questions within a selected set of topics produces a cognitive load similar to that of the actual exam.

In a highly-cited survey of ten specific learning methods [2], *distributed practice* and *practice testing* were the only methods to be judged “*high utility*” as these approaches “*benefit learners of different ages and abilities and have been shown to boost students’ performance*”. PR supports distributed practice, and a 2017 redesign by the Center for Academic Innovation¹ supports practice testing by allowing students and instructors to generate a synthetic exam covering a set of topics, then administer the questions in a timed setting.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
L@S '20, August 12–14, 2020, Virtual Event, USA.
© 2020 Copyright is held by the author/owner(s).
ACM ISBN 978-1-4503-7951-9/20/08.
<http://dx.doi.org/10.1145/3386527.3406731>

¹<https://problemroulette.ai.umich.edu>

Item	CHEM	PHYS	STAT
# PR problems	1559	970	491
Total # students enrolled	7457	4745	13829
Majority birth sex (%)	F (52.5)	M (67.2)	F (56.0)
PR user fraction	97.6%	68.3%	80.4%
Problems served ($\times 10^6$)	3.687	0.346	1.766
Median response time	42 s	112 s	36 s
Median response accuracy	68.0%	59.3%	80.5%
Median skip rate	14.2%	37.2%	6.4%
Med. # responses/user/term	364	41	78
Mean # sessions/user/term	16.4	6.4	5.4

Table 1. Characteristics of PR data in the three courses of this study. Top rows summarize the courses involved while the middle rows characterize the sets of problems with at least ten responses. The bottom rows summarize statistics of users who attempted at least one problem.

Here we report preliminary findings from student engagement with PR over a four-year period, 2013 to 2017, in three foundational² courses in the subjects of physics, chemistry, and statistics. In this working paper, we first report on subject-level differences in assessment before turning to grade benefits of practice and student behaviors and outcomes by birth sex.

DATA AND SCOPE OF STUDY

After a 2011 pilot in first-semester physics, other courses adopted PR by 2013, including general chemistry and introductory statistics. At that time, the U-M Library deployed the service with an authentication layer based on university credentials. This layer allows us to map practice study to grade outcomes. We examine data across eight academic semesters, Fall 2013 to Winter 2017 terms, in these three courses. The data include over 20,000 students who produced 5.8 million responses to a combined ensemble of 3000 questions. Each question maps to one or more topics covered by the class.

The first five rows of Table 1 summarize the scale of the data. Average class sizes in general chemistry (CHEM), physics (PHYS), and statistics (STAT) were 1746, 593, and 921, respectively. Over the eight-terms, CHEM and STAT were each led by a single senior lecturer, whereas PHYS employed instructional teams, comprised of lecturers and tenure-track faculty, of varied composition. PR questions were extracted from past examinations authored by the instructors, with numbers ranging from just under 500 in STATS to over 1500 in CHEM.

Student use of the optional PR service varies, with the fraction of students who have attempted at least one problem on PR being lowest in PHYS (68%) and highest in CHEM (98%). The enrollments, engaged user fraction, and student study behaviors factor together to give total instances of problems served to students ranging from approximately 0.35 million in PHYS to over 3.6 million in CHEM, with 1.8 million in STATS. While the student cohort in physics is over two-thirds male, the other courses are nearly equal in terms of birth sex. Female user fractions are one to six percentage points higher than that of males across the three courses.

²See <http://crlt.umich.edu/fci>.

Figure 1. Distribution of problems in the space of response accuracy and response time, smoothed via Scott's method. We only include problems with at least 10 responses. Stars indicate the median values in each dimension (see also Table 1).

User Interaction and Final Grades

Once authenticated to a PR course site, a student selects one or more topics to study, and the set of problems linked to the choice is constructed from those never-attempted or incorrectly answered by the student. One is randomly served as the first question in that study session, and the user can either skip or submit an answer. In the latter case, the user is informed as to response correctness, along with other feedback [3], before the next question is served. We define a study session as a chain of problems served to a user with less than 20 minutes elapsing between adjacent links in the chain. Over the four-year study period, students engaged in nearly 215,000 study sessions.

We obtained final letter grade information for the students in these classes via the LARC data release [8] from the Office of Enrollment Management at the University of Michigan.

SUBJECT-LEVEL DIFFERENCES IN ASSESSMENT

While all three courses employ timed examinations, the character of the questions that students are asked to address differs by subject. For every problem with at least 10 responses, we compute its mean *response accuracy*, the fraction of correct responses, as well as its *response time*, defined by the median time interval between serving and receiving an answer. Note that skips do not contribute to these “response” statistics.

Mean response accuracy can be considered a proxy for the level of *difficulty* of a problem, while the median response time reflects its *complexity*, or the *level of effort* required by a student to produce a solution. Figure 1 uses kernel density estimates to map how the set of problems in each subject populate the space of difficulty and complexity. Note that the response time axis is logarithmic.

Whereas questions in physics extend down to a mean response accuracy of 20% (corresponding to simple guessing, with one among five options), questions in chemistry and statistics are generally less difficult. The median response accuracy for statistics problems (80.5%) is significantly higher than those of chemistry (68.0%) and physics (59.3%).

Physics questions are also more complex. The median time for solving a physics problem, 112 seconds, is nearly three times longer than the values for statistics (36 s) or chemistry (42 s). As illustrated in Figure 1, the range of response times is broad in all subjects, with a standard deviation of roughly a factor of

Subject	Grade Anomaly			
	All	High PR	Low PR	High-Low
CHEM	-0.473	-0.378	-0.545	0.167 ± 0.014
PHYS	-0.425	-0.400	-0.435	0.035 ± 0.019
STAT	-0.251	-0.161	-0.309	0.149 ± 0.011
CHEM-F	-0.615	-0.477	-0.737	0.260 ± 0.020
CHEM-M	-0.321	-0.251	-0.366	0.116 ± 0.019
PHYS-F	-0.518	-0.461	-0.543	0.083 ± 0.034
PHYS-M	-0.382	-0.369	-0.387	0.018 ± 0.023
STAT-F	-0.292	-0.194	-0.363	0.170 ± 0.015
STAT-M	-0.202	-0.136	-0.261	0.137 ± 0.016

Table 2. Mean grade anomaly, $\langle GPE - GPAO \rangle$, for all students and complementary halves of high- and low-volume PR users. The last column is the difference in mean grade earned between the high and low practice subsets along with bootstrap-estimated 1σ statistical error.

two. Difficulty and complexity/effort are somewhat linked, as easier questions tend to be quicker to solve. This tendency is most pronounced in STAT, where a 10% increase in response accuracy is associated with a 15% shorter response time.

The character of the problems that a student encounters is likely to affect their desire to practice and, ultimately, their attitudes toward the subject. Not only is the level of student engagement lowest in PHYS, the rate of problems being skipped is highest in that subject as well. While many factors are at play, including instructor variability that leads to shifting topical emphasis, the enhanced difficulty of physics problems (which results in less reward) may discourage student engagement with PR when combined with the extra effort typically required to solve a physics problem. In addition, PHYS problems are often cumulative and map to multiple topics, so students tend to skip problems involving subjects not yet covered in class.

GRADE BENEFIT

Grade point average (GPA), determined on the standard 4.0 academic scale, is commonly used to characterize student aptitude at college. Since grading behaviors traditionally differ across academic divisions [4], we follow earlier studies [6, 7] and employ a measure known as the *grade anomaly*, the difference between the grade point earned by a student in a particular class (GPE), their end-of-term GPA computed from all other classes (GPAO).

The first column of Table 2 shows the mean grade anomaly, $\langle GPE - GPAO \rangle$, for all students who took these classes during the eight terms of our study (first three rows), with lower rows disaggregating by birth sex. On average, students in CHEM and PHYS earn grades at a level 0.4 – 0.5 below their GPAO, and 0.25 below in STAT. Consistent with previous studies [6, 7], we find that on average females receive lower average grades than males in all three subjects, with CHEM showing the largest difference of nearly 0.3 in grade point.

Given that benefits to memory of distributed practice are well established [1], one might expect improved grade outcomes for students who study more. In the remaining columns of

Figure 2. Average study rate, in 15-minute intervals, of females (F) is higher than that of males (M) in all subjects.

Table 2, we examine the benefits of PR practice by dividing the student population into high- and low-volume users based on the median number of responses per term among users (Table 1). In CHEM, the median number is largest (364), followed by STAT (78) and PHYS (41) for the full, eight-term collection of responses. We find the results insensitive to minor changes in these thresholds.

The remaining columns of Table 2 show mean grade anomalies for high- and low-volume subsets. We bootstrap within each component to compute the mean offset, and its 68% statistical uncertainty, presented in the final column. For CHEM and STAT, students in the high-volume subset earn an average grade benefit of ~ 0.15 , with very high statistical significance. A smaller grade boost is found in PHYS, 0.035 ± 0.019 .

Previous analysis of PR use in a single 2012 term of a physics class found a grade benefit of 0.22 ± 0.04 [3], consistent with our findings for CHEM and STAT. That class (531 students) had a very high PR participation rate of 86%, with nearly half of enrolled students completing 50 or more problems. We speculate that both content misalignment — problems from older examinations not reflecting currently authored problems — and diffused instructor engagement are potential factors muting the grade benefits to PHYS students.

DIFFERENCES BY BIRTH SEX

The gender gap in STEM fields is an active area of research. While gone are the days of active exclusion of women from such fields, barriers to gender parity persist, requiring active engagement on the part of institutions to achieve equity.

Figure 2 shows the aggregate rate at which students of each birth sex study, measured in 15-minute intervals and averaged across all eight terms. Overall, practice activity peaks in the late afternoon, drops sharply when the dining halls open at 5pm, and then peaks again near 10pm. Females tend to study more than males in all three subjects, viewing 21 – 26% more problems per student, on average.

In the last six rows of Table 2, we see that higher practice volumes seem to boost female grades more than males in all three subjects. The boost is especially significant in CHEM, with mean differential grade point shifts of 0.26 ± 0.02 for females and 0.12 ± 0.02 for males.

A students' response accuracy on PR provides a subject-specific measure of their understanding of course material,

Figure 3. Grade point differential between female and male students in ten bins ordered by PR response accuracy. Points are plotted at the mean accuracy and grade differential, with the horizontal bar indicating the full width of the decile bin and vertical error bars giving central 95% confidence interval on the mean from 10,000 bootstrap resamplings.

and so provides an alternative and more precise control for subject-level aptitude when assessing differential outcomes. In Figure 3 we examine grade biases by birth sex conditioned on student response accuracy. Points show the mean difference (F–M), and 95% uncertainty, in mean grade point earned within each decile bin of practice accuracy. In CHEM, females earn -0.13 ± 0.03 (95% confidence error) grade points lower than males with the same response accuracy. We find no significant difference in PHYS or STAT, with 95% confidence ranges of $[-0.098, 0.023]$ (PHYS) and $[-0.007, 0.048]$ (STAT).

SUMMARY

Students in large-enrollment introductory courses often encounter timed examinations for summative assessment. The Problem Roulette study service reduces institutional inequities by providing uniform access to authentic practice material. From a corpus of over five million responses to questions over a four-year period, we identify disciplinary differences regarding the degrees of difficulty and solution effort that students encounter. What constitutes a “good” problem for an exam is a matter of both domain tradition and, for locally authored material, the attitudes of the particular instructor or instructional team. There is every reason to suspect that the latter can often be more influential than the former, so we caution against interpreting the differences we report here as broadly applicable to other campuses.

Student engagement with the optional service varies by subject, with physics having lowest levels and chemistry highest. Thresholding, based on the median number of practice problems per term, the student population into low- and high-practice volume subsets, we find grade benefits to the high-volume groups in all subjects, with higher significance in the subjects with high student engagement, chemistry and statistics. A novel finding of our study is that the grade benefit of high-volume study is larger to females than it is to males, particularly in chemistry.

Outfitting multiple campuses with PR instances, or similar offerings, that serve local content is a potential pathway to reduce inequities in access to quality practice as well as to enable federated research into the issues raised by initial study.

ACKNOWLEDGMENTS

We thank many early PR team members, especially Brenda Gunderson, Dotie Sipowska, Dave Winn, Tim McKay, Noah Botimer, and John Leasia, as well as Efrain Segarra, James Kakos, and Denny Walsh for early analysis. We thank Kyle Schulz and Cait Holman for LARC assistance, and acknowledge funding from the Center for Research on Learning and Teaching, the Instructional Support Services team in the College of Literature, Science, and Arts, and the Center for Academic Innovation, all at the University of Michigan.

REFERENCES

- [1] Aaron S. Benjamin and Jonathan Tullis. 2010. What makes distributed practice effective? *Cognitive Psychology* 61, 3 (2010), 228 – 247. DOI: <http://dx.doi.org/https://doi.org/10.1016/j.cogpsych.2010.05.004>
- [2] John Dunlosky, Katherine A. Rawson, Elizabeth J. Marsh, Mitchell J. Nathan, and Daniel T. Willingham. 2013. Improving Students’ Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. *Psychological Science in the Public Interest* 14, 1 (2013), 4–58. DOI: <http://dx.doi.org/10.1177/1529100612453266> PMID: 26173288.
- [3] August E. Evrard, Michael Mills, David Winn, Kathryn Jones, Jared Tritz, and Timothy A. McKay. 2015. Problem roulette: Studying introductory physics in the cloud. *American Journal of Physics* 83, 1 (2015), 76–84. DOI: <http://dx.doi.org/10.1119/1.4894061>
- [4] Joseph C Hermanowicz and David W Woodring. 2019. The Distribution of College Grades across Fields in the Contemporary University. *Innovative Higher Education* 44, 6 (2019), 497–510.
- [5] Amy Homkes-Hayes. 2018. Origin Story Podcast: Problem Roulette. Audio. (8 February 2018). Accessed May 13, 2020 at <https://ai.umich.edu/blog/origin-stories-problem-roulette/>.
- [6] Benjamin P. Koester, Galina Grom, and Timothy A. McKay. 2016. Patterns of Gendered Performance Difference in Introductory STEM Courses. (2016).
- [7] Rebecca L. Matz, Benjamin P. Koester, Stefano Fiorini, Galina Grom, Linda Shepard, Charles G. Stangor, Brad Weiner, and Timothy A. McKay. 2017. Patterns of Gendered Performance Differences in Large Introductory Courses at Five Research Universities. *AERA Open* 3, 4 (2017), 2332858417743754. DOI: <http://dx.doi.org/10.1177/2332858417743754>
- [8] U-M Office of Enrollment Management. 2020. Learning Analytics Data Architecture. (2020). <https://enrollment.umich.edu/data-research/learning-analytics-data-architecture-larc>.
- [9] Barry J. Zimmerman. 1990. Self-Regulated Learning and Academic Achievement: An Overview. *Educational Psychologist* 25, 1 (1990), 3–17. DOI: http://dx.doi.org/10.1207/s15326985ep2501_2