# *Increasing Available Attempts: Changes in Student Correctness on Formative Introductory Physics Problems*

**Citation:** MacDonagh, Aidan. 2023. "Increasing Available Attempts: Changes in Student Correctness on Formative Introductory Physics Problems."

**As Published:** https://doi.org/10.1145/3573051.3596202

**Publisher:** ACM|Proceedings of the Tenth ACM Conference on Learning @ Scale

**Persistent URL:** https://hdl.handle.net/1721.1/152141

**Version:** Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

# Increasing Available Attempts: Changes in Student Correctness on Formative Introductory Physics Problems

Aidan MacDonagh
aamacdon@mit.edu
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA

## ABSTRACT

Student submission data from formative graded physics problems were analyzed to detect effects from increasing the number of available attempts. The problems were used in two subsequent runs of MIT's general requirement introductory electricity and magnetism course, 8.02, with 190 students in the 2021 course and 203 in the 2022 course. Students completed the problems asynchronously in interactive online lessons as part of a blended learning design. The problems awarded full credit on any attempt and gave correctness as submission feedback. A number of available attempts was set for each problem by the course designers, varying across problems. Between 2021 and 2022, this number was increased for some problems but not others, creating a natural experiment. Qualitative effects were evaluated relative to a control group of unmodified problems and were found to differ between constructed response and selected response problems. Constructed response problems saw a significant decrease in first-attempt success rate and increase in rate of abandoning with attempts remaining, despite an insignificant increase in overall success rate. Selected response problems saw a significant increase in overall success but negligible changes in first-attempt success and abandoning. For both types, there was a significant increase in the overall number of attempts used. These results suggest that increasing attempts for constructed and selected response problems may encourage undesirable approaches to answering the problems, even in a formative context.

## CCS CONCEPTS

• **Applied computing → Interactive learning environments**.

## KEYWORDS

introductory physics, online problems, multiple attempts, formative assessment, blended learning, constructed and selected response

**ACM Reference Format:**
Aidan MacDonagh. 2023. Increasing Available Attempts: Changes in Student Correctness on Formative Introductory Physics Problems. In *Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S '23), July 20–22, 2023, Copenhagen, Denmark.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3573051.3596202

*L@S '23, July 20–22, 2023, Copenhagen, Denmark*
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0025-5/23/07.
https://doi.org/10.1145/3573051.3596202

## 1 INTRODUCTION

The prevalence of blended and online learning has made online interactive problems a common instruction and assessment tool [9]. They are particularly suitable as online formative assessment [2], and they can provide timely feedback to aid in self-regulated learning [10]. Various work has demonstrated positive and negative effects from implementing such problems with multiple attempts.

On the positive side, multiple attempts seem to benefit learners in ways that learners themselves may perceive. Problem scores have been found to increase when students are merely given follow-up attempts, even without feedback hints and even on constructed response items [3], and high final scores are observed when many attempts are granted [8]. Student perception of overall effort spent may be lower when given multiple attempts, regardless of cumulative time [12]. Multiple attempts may also mitigate student anxiety and improve score reliability in an assessment context [4].

Possible negative effects also extend to student behavior and quality of assessment. Multiple-attempt graded problems may overprioritize the correct answer instead of encouraging good problem solving processes [5]. In that vein, "gaming the system" strategies like random guessing have been linked to higher success rates and lower consistency on selected response problems [3, 12]. Other work has shown that high maximum attempt numbers are associated with lower success on early attempts, inferring that students' tries are generally less effortful [8]. The role of resilience in student success on multiple-attempt problems has been modeled to explore how it may be conflated with knowledge [13].

This work seeks to explore similar questions in the context of multiple-attempt, graded online problems in the formative assessment materials of an introductory physics course. It leverages a blended learning system with a large set of student engagement data as well as modifications between course runs that allowed for a natural experiment.

## 2 BACKGROUND

### 2.1 The course

The course in question is a one-semester general-requirement undergraduate physics course at MIT, "8.02 Physics II: Electricity and Magnetism," taught in the fall semester in the Technology Enabled Active Learning format [6, 7]. The course covers first-year electricity and magnetism and incorporates a variety of graded components, including in-class participation (Clicker Questions, Group Problems, and Experiments), asynchronous assignments (Online Lessons and Problem Sets), and summative assessments (Exams). Active learning, group problem solving, and a flipped classroom are all aspects of the course design. This version of the course is

**Table 1: Course component grades (student average) and weights compared between the 2021 and 2022 course runs. In cases where grade differences are statistically significant, the higher of the two average grades is boxed.**

| | Grades (%) | | Weights (%) | |
|---|---|---|---|---|
| Component | 2021 | 2022 | 2021 | 2022 |
| Group Problems | $\boxed{90}$ | 85 | 6 | 5 |
| Experiments | 93 | 90 | 6 | 7 |
| Clicker Questions | $\boxed{90}$ | 83 | 10 | 7 |
| Online Lessons | 88 | 89 | 14 | 13 |
| Problem Sets | 93 | 92 | 14 | 13 |
| Exams | 75 | $\boxed{79}$ | 50 | 55 |
| TOTAL | 83 | 84 | 100 | 100 |

run once each year with a fairly stable syllabus and typical enrollment of about 200 students. Two consecutive runs of the course are considered here: fall 2021 with 190 students, and fall 2022 with 203 students. Before analyzing the graded problems themselves, it is important to consider differences between the 2021 and 2022 course runs.

Both 2021 and 2022 course runs featured the same types of graded components, but there were some adjustments to the component weights between the two runs, and there were differences in average student grades. These values are summarized in Table 1. The asynchronous Online Lessons and Problem Sets showed little change, but both in-class components and summative assessments showed statistically significant differences, which illustrates that the courses and students are indeed different and that student performance on any assessments may naturally differ.

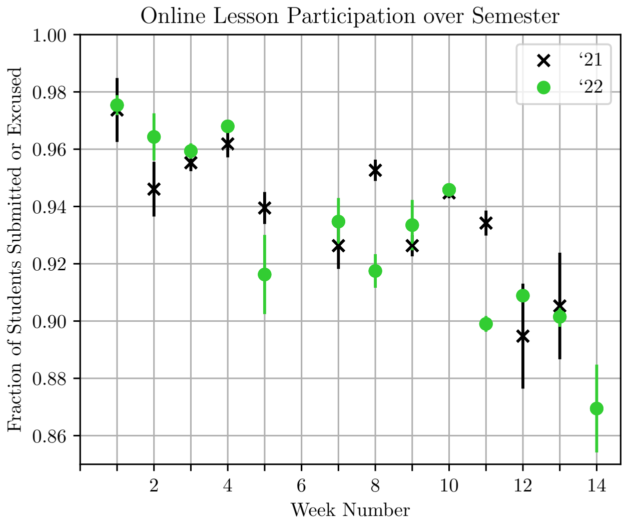

**Figure 1: Fraction of Online Lessons grades that were non-zero, averaged per week, across the 2021 (black, 'x') and 2022 (green, circle) courses. Error bars indicate standard errors.**

## 2.2 The problems

The graded problems analyzed in this work are housed in only one of the course components, the Online Lessons, which are part of a flipped classroom model and are hosted on the course learning management system (LMS). The Lessons are sequences of instructional videos and readings interwoven with ungraded and graded practice problems. The graded problems are designed to be low weight with multiple but limited attempts, feedback of correctness, and full credit on any attempt. They are intended as low-stakes formative assessment that encourages effortful practice in service of self-regulated learning.

The organization and content of the Online Lessons were not modified significantly between the 2021 and 2022 runs of the course, and student engagement was comparable across both runs. Figure 1 shows that the number of students participating in Lessons each week was similar in both course runs throughout the semester. The total number of Lessons increased from 36 in 2021 to 42 in 2022 (largely because of an extra set in week 14), but the average number of problems per Lesson was similar at 4.6 and 4.9 respectively. Taken together with similar component weights and average grades in Table 1, this suggests that student engagement in the Online Lessons was comparable across 2021 and 2022.

Finally considering the graded problems themselves, there were no fundamental changes to how they function within an Online Lesson. For a given graded problem, a student is granted a maximum number of attempts with which to submit a correct answer. Upon submission, the LMS provides the feedback of whether the submitted answer is correct. Full credit is awarded for a correct submission on any attempt, and zero credit is awarded for an incorrect submission. Each graded problem has an associated number of points, displayed by the LMS, that constitutes part of a Lesson total. Each individual Lesson is a fraction of the total Lesson component weight. The average across the problems considered here was approximately 1 point per problem, or roughly 0.07% of the total course grade.

## 3 METHODS

After the 2021 course run, many of the graded problems were given an increased number of available attempts due to student feedback that attempts were too limited. This change provided an opportunity for a natural experiment of sorts, in which the 2021 students encountered all problems in their unmodified state, and the 2022 students encountered some problems unmodified (the **control** problems) and some with increased available attempts (the **modified** problems). The principal research question of this work is thus,

- **RQ1** What effect does increasing the number of available attempts for a graded problem in an Online Lesson have on student performance on that problem?

Attempting to answer the question could provide insight into a more difficult question that would inform future course design:

- **RQ2** How many attempts should be granted for a given graded problem in an Online Lesson?

To contextualize the results and discussion that follow, it will be helpful first to look in more detail at features of the graded problems. Each graded problem has a number of available attempts ($M$) set by the course designers. Both constructed response (**CR**) and selected response (**SR**) problems are present. **SR** problems offer a limited set of choices with only one correct answer, and so the space of possible submissions is exactly that set of choices; $M$ for **SR** problems is always less than the number of choices, ensuring that random guessing is not guaranteed to earn credit. In contrast, the space of possible submissions for **CR** problems is not constrained to a set of choices, and so $M$ tends to be somewhat higher for **CR** as a result. These differences warrant **CR** and **SR** being analyzed separately in this work. Attempt number distributions for both problem types are shown in Figure 2.

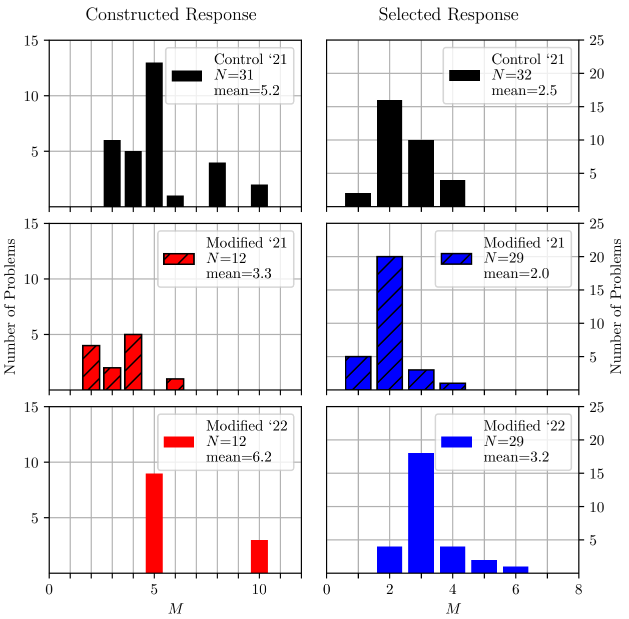

**Figure 2: The distributions of problems in the control group (top row) and modified group (bottom two rows) sorted by number of available attempts, $M$, for both constructed response CR (left column) and selected response SR (right column). The top two rows show control and modified problems from the 2021 course run before any changes to the modified group. The bottom row shows the increased attempts for the modified problems in the 2022 course run. In that order, the mean $M$ values for CR are approximately 5, 3, and 6, and for SR they are approximately 2.5, 2, and 3.**

To quantify student performance on these problems, the following outcome quantities are considered for each problem: first-attempt success rate ($s_1$), calculated as the fraction of students who succeed (answer correctly) after one attempt; cumulative success rate ($S$), calculated as the fraction who succeed after all attempts; abandon rate ($A$), calculated as the fraction who do not succeed but also do not exhaust the available attempts; average number of attempts used ($\overline{m}$); average fraction of available attempts used for success ($\overline{m}_S/M$); and average fraction of available attempts used before abandoning ($\overline{m}_A/M$). For each problem, these averages are evaluated over students who attempted the problem.

**Table 2: Results of two-sample $t$-tests, modified minus control, for constructed and selected response problems. The mean change from 2021 to 2022 in each outcome quantity is compared between modified and control groups. The qualitative effects visualized by Figure 3 are quantified here with corresponding $t$-statistics and $p$-values from the $t$-tests. The $t$-statistics are boxed if the $p < 0.05$ criterion is met.**

| | Constructed | | Selected | |
|---|---|---|---|---|
| Outcome | $t$ | $p$ | $t$ | $p$ |
| $s_1$ | $\boxed{-2.3}$ | 4e-2 | +0.2 | 8e-1 |
| $S$ | +1.2 | 2e-1 | $\boxed{+6.0}$ | 1e-6 |
| $A$ | $\boxed{+2.2}$ | 5e-2 | +0.1 | 9e-1 |
| $\overline{m}$ | $\boxed{+2.9}$ | 1e-2 | $\boxed{+4.2}$ | 2e-4 |
| $\overline{m}_S/M$ | $\boxed{-5.7}$ | 1e-4 | $\boxed{-9.2}$ | 3e-10 |
| $\overline{m}_A/M$ | +0.5 | 6e-1 | -1.0 | 3e-1 |

## 4 RESULTS

The observational nature of this work means that direct comparison between 2021 and 2022 performance is difficult. The students in the 2021 and 2022 course runs had outcomes that were statistically significantly different at the level of course grades, let alone at an individual problem level. In order to account for the effects of course run on the outcomes, the difference between the changes of the control and modified groups—rather than those changes themselves—can be evaluated. For example, the first-attempt success rate $s_1$ may decrease from 2021 to 2022 for both control and modified problems; but if the decrease for modified problems is significantly larger or smaller than that for control problems, this suggests an effect from increased available attempts. This analysis can be likened to characterizing an interaction effect in a multiple regression model [1].

Results are shown qualitatively in Figure 3 for both **CR** and **SR** problems. In each plot, the course run is on the horizontal axis, and the outcome is on the vertical. Thus each single line indicates the change from 2021 to 2022 for an outcome averaged over a single problem group, either the control problems (black, dashed) or the modified problems (red or blue, solid). The effect of increased attempts may be inferred from the difference between these two lines' slopes [1]. For example, in the $s_1$(CR) plot, the modified line has a more negative slope than the control line, suggesting that increasing the available attempt number had the effect of lowering success rate on first-attempt, relative to the control. Error bars (proportional to $1/\sqrt{N}$) are presented for each point to visualize how the outcome quantities compare to each other across groups and runs. For a more quantitative analysis, a $t$-test for two independent samples with unequal variance was conducted for the difference in slopes for each pair of lines [11]. This difference in slopes is here equivalent to the difference in changes, $(\Delta y)_{\text{modified}}$ and $(\Delta y)_{\text{control}}$, where $\Delta y = \text{mean}(y_{2022} - y_{2021})$ for an outcome quantity $y$. The null hypothesis is that the mean change from 2021 to 2022 for an outcome is the same in the modified and control populations. Results of these $t$-tests are compiled in Table 2.

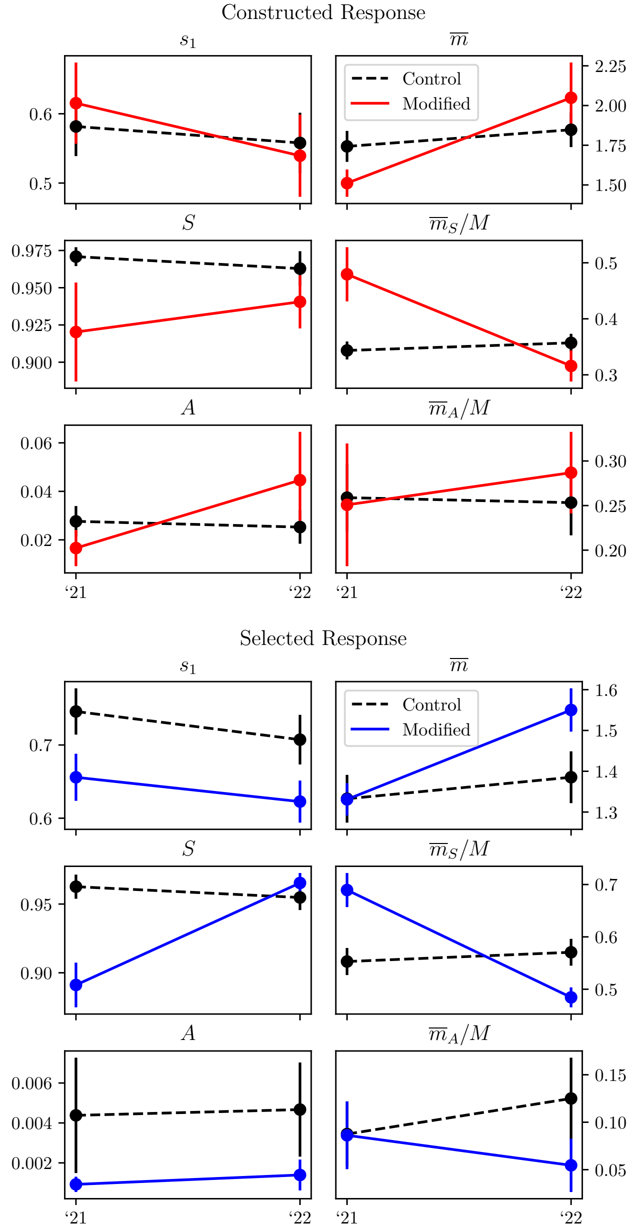

**Figure 3: Changes in outcomes from 2021 to 2022 visualized as lines from left to right, for constructed and selected response problems. The outcome quantities are first-attempt success rate $s_1$, cumulative success rate $S$, abandon rate $A$, average number of attempts $\overline{m}$, average fraction of available attempts for success $\overline{m}_S/M$, and average fraction of available attempts before abandoning $\overline{m}_A/M$. Control (black, dashed) and modified (red or blue, solid) groups are shown separately.**

## 5 DISCUSSION

The increase in $M$ for some problems in this course was motivated by student feedback that attempts were too few. This suggested a difference between the perceptions of the course designers and the experiences of the students themselves, namely that the designers underestimated the attempts needed for students to comfortably and correctly answer the graded problems in a formative context. Lacking a controlled trial, and with clear evidence that the course run itself has substantial effects on problem performance outcomes, the data analyzed here nevertheless provide some clues about how increasing available attempts might have both positive and negative effects.

First, it is worth remarking on the background effects of course run (2021 vs. 2022). The general trend in problem performance on the control problems is one of lower success rates ($s_1$ and $S$) despite slightly higher average attempt numbers $\overline{m}$. These changes from 2021 to 2022 within the control and modified groups are not statistically significant, but they do align with broad course-level performance differences like lower participation. Structural changes to the course, in particular an increase in daily class length, may have played a part.

Even so, students in the 2022 run actually achieved higher success rates (and therefore higher grades) on the **CR** problems. While the effect was modest and not statistically significant, it might constitute a positive outcome in the eyes of the students. However, there seem to be associated costs; the first-attempt success rate $s_1$ decreased while the abandon rate $A$ increased. This suggests that a subset of students may be making less effortful tries or giving up in greater numbers now that there are nominally more chances to succeed, a phenomenon reported elsewhere [8]. Even though the average number of attempts increased, those additional attempts may not have been productive. From this perspective, the case for increasing $M$ for **CR** problems is not compelling.

On the **SR** side, the cumulative success rate $S$ experienced a significant positive effect, but $s_1$ and $A$ showed no noticeable effects across the control and modified groups. Compared to the **CR** situation, at least, this may imply a disconnect between effort and success, which would be in line with discussions of guessing and "gaming the system" in other work [3, 12].

## 6 CONCLUSIONS

The intent of the 8.02 course designers in choosing a given $M$ could be articulated as providing students with enough attempts to allow for mistakes and keep stress low, but also with few enough to keep tries effortful and discourage unproductive practices like guessing. Students themselves called for increased attempts in the course's formative assessment, and implementing that change gave rise to the research questions in this work. In answer to **RQ1**, while it appears that increasing attempts likely led to higher grades for the students, several inferred negative effects suggest that the problems may have diminished in quality as formative assessment for learning. Better characterizing this give-and-take in the formative context and addressing **RQ2** are goals for future work.

## ACKNOWLEDGMENTS

The author is grateful for feedback on methodology, analysis, and visualization from colleagues in the MIT Physics Education Group and the MIT Digital Learning Lab, most especially the guidance of Michelle Tomasik.

## REFERENCES

[1] Leona S. Aiken, Stephen G. West, and Raymond R. Reno. 1991. *Multiple Regression: Testing and Interpreting Interactions.* Sage Publications. 38–46 pages.

[2] Giora Alexandron, Mary Ellen Wiltrout, Aviram Berg, and José A. Ruipérez-Valiente. 2020. Assessment That Matters: Balancing Reliability and Learner-Centered Pedagogy in MOOC Assessment. In *Proceedings of the Tenth International Conference on Learning Analytics and Knowledge* (Frankfurt, Germany) *(LAK '20)*. Association for Computing Machinery, New York, NY, USA, 512–517. https://doi.org/10.1145/3375462.3375464

[3] Yigal Attali. 2015. Effects of multiple-try feedback and question type during mathematics problem solving on performance in similar problems. *Computers and Education* 86 (Aug. 2015), 260–267. https://doi.org/10.1016/j.compedu.2015.08.011

[4] Yigal Attali and Don Powers. 2009. Immediate Feedback and Opportunity to Revise Answers to Open-Ended Questions. *Educational and Psychological Measurement* 70, 1 (March 2009), 22–35. https://doi.org/10.1177/0013164409332231

[5] Scott Bonham, Robert Beichner, and Duane Deardorff. 2001. Online homework: Does it make a difference? *The Physics Teacher* 39 (June 2001). https://doi.org/10.1119/1.1375468

[6] Yehudit Judy Dori and John Belcher. 2005. How Does Technology-Enabled Active Learning Affect Undergraduate Students' Understanding of Electromagnetism Concepts? *Journal of the Learning Sciences* 14, 2 (2005), 243–279. https://doi.org/10.1207/s15327809jls1402_3

[7] Peter Dourmashkin, Michelle Tomasik, and Saif Rayyan. 2020. *The TEAL Physics Project at MIT.* Springer International Publishing, Cham, 499–520. https://doi.org/10.1007/978-3-030-33600-4_31

[8] Gerd Kortemeyer. 2015. An empirical study of the effect of granting multiple tries for online homework. *American Journal of Physics* 83 (June 2015). https://doi.org/10.1119/1.4922256

[9] Paula Magalhaes, Diogo Ferreira, Jennifer Cunha, and Pedro Rosario. 2020. Online vs traditional homework: A systematic review on the benefits to students' performance. *Computers and Education* 152 (July 2020). https://doi.org/10.1016/j.compedu.2020.103869

[10] David J. Nicol and Debra Macfarlane-Dick. 2006. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education* 31, 2 (2006), 199–218. https://doi.org/10.1080/03075070600572090 arXiv:https://doi.org/10.1080/03075070600572090

[11] National Institute of Standards and Technology. 2012. NIST/SEMATECH e-Handbook of Statistical Methods. https://doi.org/10.18434/M32189

[12] M Taylor Rhodes and Jeffrey K Sarbaum. 2015. Online Homework Management Systems: Should We Allow Multiple Attempts? *American Economist* 60, 2 (Sept. 2015), 120–131. https://doi.org/10.1177/056943451506000203

[13] Susu Zhang, Yoav Bergner, Jack DiTrapani, and Minjeon Jeon. 2021. Modeling the interaction between resilience and ability in assessments with allowances for multiple attempts. *Computers in Human Behavior* 122 (Sept. 2021). https://doi.org/10.1016/j.chb.2021.106847