Evaluating the effectiveness of two methods to improve students' problem solving performance after studying an online tutorial

Zhongzhou Chen,[1] Kyle M. Whitcomb,[2] Matthew W. Guthrie,[1] and Chandralekha Singh[2]

[1]*Department of Physics, University of Central Florida, Orlando, FL, 32816*
[2]*Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA, 15260*

An earlier study using a sequence of online learning modules found that a significant fraction of undergraduate students were unable to solve similar new problems after learning from an online problem solving tutorial. The current study examines the effectiveness of two methods to improve students' subsequent problem solving performance. First, an "on-ramp" module designed to help students develop proficiency in relevant basic skills was added prior to the tutorial. Second, a new "transfer" module was added after the tutorial and before the final quiz module. In this new module, half of the students were assigned a compare contrast task in which they were asked to read the solution of a new problem, and compare it to a similar previous problem. The other half of the students were asked to answer several scaffolding questions and read the solution to the same problem. For the on-ramp module, we found that students' performance on subsequent modules were significantly improved over the previous year, and in one of the two sequences we found supporting evidence that the improvement was due to the addition of the on-ramp module. However, neither version of the new transfer module had significant impact on students' performance on the last module, nor were there any significant differences between the two modules. The study demonstrated that mastery-style online homework can serve as an efficient and flexible method for evaluating the effectiveness of new instructional designs.

I. INTRODUCTION

Studies in both general human problem solving and problem solving in physics have long shown that it is difficult for novices to transfer the understanding and skills learned in one problem context to a different, new context [1–3]. In physics, it is well known that novices tend to focus more on superficial differences between problems and pay less attention to deep structural similarities between problems [4]. In an earlier study involving a sequence of three online learning modules, we observed that while most college, introductory level physics students learned to solve a specific problem after engaging with an online problem solving tutorial in the first module, student performance when solving similar problems on two subsequent modules was either unchanged or only slightly improved [5]. In the current study, we tested two strategies for improving student performance on subsequent similar problems following an online tutorial.

First, research in both learning science [6–8] and physics education [9, 10] have shown that explicitly comparing and contrasting multiple examples can be a more effective method to understand common deep structure of the problem compared to studying isolated examples in sequence. The theory would predict that explicitly asking students to compare a new problem to a previously solved similar problem results in better performance on subsequent transfer tasks when compared to only asking students to study the new problem.

Second, students' performance in our previous study could have been negatively impacted by a lack of sufficient mastery of one or more basic skills [11, 12], such as identifying the direction of angular momentum using the right hand rule. Students who are not fluent with those skills may have to devote too many cognitive resources in executing those procedures, leaving insufficient cognitive capacity to process the deeper structure of the solution. Alternatively, students could have learned the correct problem solving procedure, but made a mistake such as a sign error in one of the steps, leading to an incorrect answer. In either case, performance on subsequent similar problems could improve if they had the opportunity to practice and strengthen their basic skills.

Therefore, we will investigate the following two research questions:

RQ 1. Does answering several compare-contrast style questions lead to better performance on a subsequent transfer task compared to completing several guided-tutorial style questions?

RQ 2. Does the addition of an "on-ramp" module designed to develop basic procedural skills improve students' performance on subsequent problem solving tasks?

Both research questions will be answered by analyzing data collected from students' interaction with a sequence of online learning modules, which are assigned as regular homework for students to complete over the time span of two weeks. Since students' performance on online homework problems could be impacted by other extraneous factors, such as answer copying from various sources [13, 14], we will also take certain data analysis measures to estimate whether the observed results are biased by certain extraneous factors, as explained in section III.

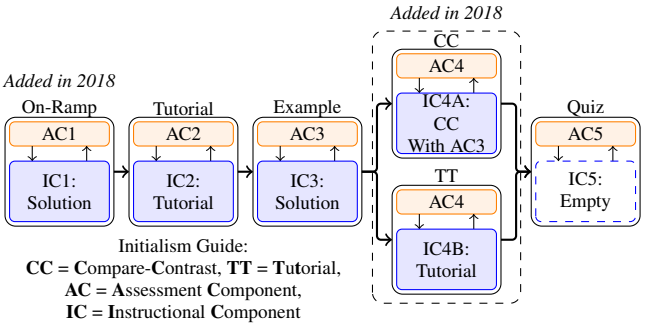

FIG. 1. The sequence of Online Learning Modules (OLMs) designed for this experiment. Each OLM contains an assessment component (AC) and instructional component (IC). Students are required to make at least one attempt on the AC first, then are allowed to view the IC, and then make subsequent attempts on the AC. Modules 1 and 4 were added for the 2018 implementation. Further, students are randomly assigned into groups that receive two different versions of the IC in module 4, a compare-contrast (CC) task and a tutorial (TT).

II. METHODS

A. OLM Sequence Structure

The study was conducted using online learning modules (OLMs) [5, 15, 16] implemented on the Obojobo platform [17] developed by the Center for Distributed Learning at the University of Central Florida (UCF). Each OLM contains an assessment component (AC) and an instructional component (IC). Students have 5 attempts on the AC which contains 1-2 multiple-choice problems, but must make at least one attempt before being allowed to access the IC. The IC contains instructional text and figures, and practice questions which will reveal the solution after a student submitted an answer. A student must either pass the AC or use up all 5 attempts on one module before being allowed access the next module in an OLM sequence. Students' interaction with each OLM can be roughly divided into three stages. In the pre-study (Pre) stage, students make one or more attempts on the AC prior to accessing the IC. Those who failed during the Pre stage can study the IC during the Study stage, before going into the post-study (Post) stage to make additional attempts on the AC. In 80% of the cases students study the IC of each module only once. In the rest of the cases, their two longest study events were combined in to one study event, and the rest of shorter events neglected from analysis.

Although each student is given a total of 5 attempts on each module, in the current study we count a student as passing the module if the student correctly answers all problems in the AC within 2 attempts. More specifically, students are considered to pass in the Pre stage if they answered correctly within

the first 2 attempts prior to accessing the IC, and pass in the Post stage if they answered correctly within 2 attempts after accessing the IC, regardless of how many attempts they took during the Pre stage.

B. Study Setup

In Fall 2017, two sequences each containing 3 OLMs (specifically, modules 2, 3, and 5 in Fig. 1) were assigned to students enrolled in the calculus based introductory physics class at UCF. The ACs of each OLM contain one problem that can be solved using the same physics principle. The IC of the first OLM contains an online tutorial for the problem in the AC, developed by DeVore and Singh [18, 19]. The IC of the second OLM contains a worked solution to the AC problem, and the IC of the last OLM is empty since the last module is intended to serve the role of a quiz. The first sequence is on rotational kinematics (RK), involving Atwood machine type problems with blocks hanging from massive pulleys. The second sequence is on conservation of angular momentum (AM), involving angular collision problems such as a girl jumping onto a merry-go-round.

The two OLM sequences were modified as follows and implemented again in Fall 2018 by the same instructor. **To investigate RQ1**, we added a new module (module 4 in Fig. 1) in the 2018 implementation between the worked-example module and the quiz module. The AC for this module consists of a new problem that shares the same deep structure as the problem in the previous module, but differs in surface features. The IC for this module comes in two different formats. The compare-contrast (CC) format asks students to first study the worked solutions of the problem in the AC of the current module and the one in the previous module, then answer 2-3 practice questions comparing the similarities and differences in the solutions of both problems. Each question asks students to select from a list of physics equations the ones that can either be applied to both problems, or are only applicable to one of the problems. In the tutorial (TT) format, the problem in the AC is broken down into 2-3 tutorial-style practice problems, and the solution to each of the practice problems combine to form the complete solution of the AC problem in the current module. The student population is divided into two groups with matching average score on a previous midterm exam. Each group is presented with one format of the IC in the RK sequence, and the other format in the AM sequence. RQ1 can be answered by comparing the performance of students in the two groups on the AC of the Quiz module.

To investigate RQ2, we require students to complete an "on-ramp" module (module 1 in Fig. 1) prior to accessing the tutorial module, with the intention to develop or refresh one or more basic procedural skills necessary to solving the problem in the tutorial module. For the RK sequence, the on-ramp module presents students with two of the simplest form of Atwood machine problems, involving one or two blocks hanging at the same radius from a single massive pulley. Those problems are intended for students to focus on practicing writing down multiple parallel equations, including writing down Newton's second law for both the hanging blocks and the pulley and stating the constraint conditions, without being distracted by a more complicated problem setup. The problems are placed in the AC of the module, while their solutions are in the IC. For the AM sequence, a common student difficulty is to correctly identify and calculate both the magnitude and the sign of the angular momentum of an object traveling in a straight line about a point that lies off of the line. Therefore, the on-ramp module involves assessment and practice problems focusing on developing that particular skill. RQ2 can be answered by measuring students' performance on their AC attempts on the following two modules (tutorial and worked example), especially in the Pre stage of each module, and comparing the performance of the Fall 2018 student population with that of the Fall 2017 population, who did not receive the on-ramp module.

Data analysis, statistical testing and visual analysis are conducted using R [20] and the tidyverse package [21].

III. RESULTS

To answer RQ1, we calculated and compared the passing rate on the Quiz module between Fall 2018 students subjected to the CC and TT conditions on module 4, and found no statistically significant differences between the two groups in either sequence. Considering that some students may not have fully engaged with the IC of module 4, we conducted the same comparison among students who spent sufficient amount of time studying the IC of module 4. Since the distribution of study duration on the IC is approximately log-normal, we considered two definitions for "sufficient" time on the IC: 1) spending more than the mean of log study duration and 2) spending more than one standard deviation below the mean of log study duration. In neither case did we find any statistically significant differences in the passing rate of the Quiz module between the two conditions.

To reveal the effects of the on-ramp module and answer RQ2, we first plotted passing rates on the Pre and Post stage AC attempts of each module in Fig. 2, for students in both Fall 2017 and Fall 2018 semesters. The Pre passing rates are calculated based on all students who attempted the module, while the Post passing rates are calculated only for those who failed in Pre (i.e., the pass rates reported for Post are not cumulative). Note that since modules 1 and 4 were added in 2018, the 2017 population does not have data for those modules. Furthermore, since the IC of the Quiz module was empty, we did not distinguish between Pre and Post stage attempts when analyzing its passing rates. Lastly, since we did not find any differences between the CC and TT conditions on module 4, the two groups were combined in this analysis.

As shown in Fig. 2, Fall 2018 students had higher passing rates than Fall 2017 students in both the Pre and Post stages of every module of each OLM sequence. To examine the significance of the differences, we conducted Fisher's exact test on each pair of passing rates common to both 2017 and 2018. Ten such tests were conducted and the statistically significant results shown on the left side of Table I. After performing p-

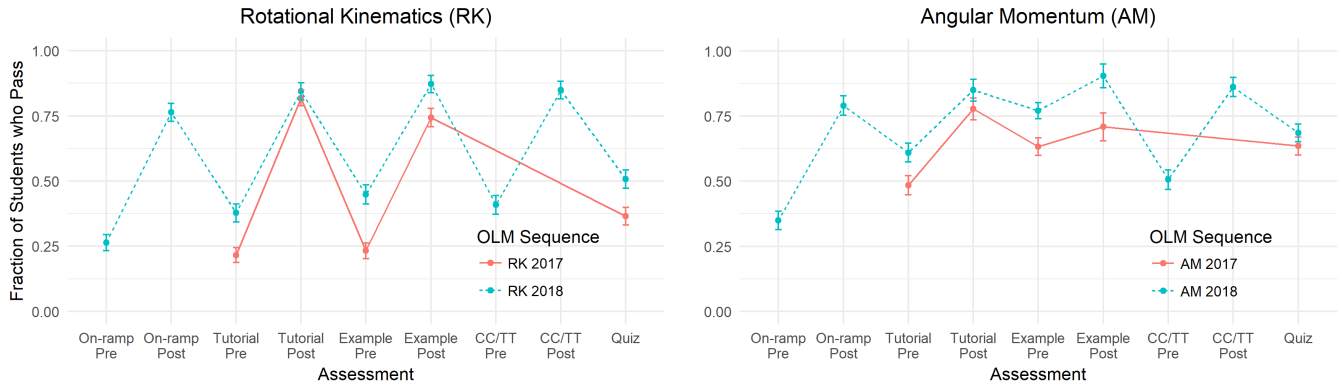

FIG. 2. The passing rates on each assessment for the two OLM sequences. Passing rates are calculated as the fraction of students who attempt the assessment and pass within two attempts. Passing rates on Post assessments are calculated only for those who did not pass in Pre.

TABLE I. Findings of student performance differences from 2017 to 2018. Of the 30 Fisher's exact tests performed comparing performance of students in 2017 to those in 2018, the 10 below were the only results significant at $p < 0.05$. Due to the large number of tests performed, we also report the p-values adjusted by the Holm-Bonferroni method.

Population	Assessment	2017 Pass Rate	2018 Pass Rate	p	p_{adj}	1st Attempt Duration	Assessment	2017 Pass Rate	2018 Pass Rate	p	p_{adj}
All	RK Tutorial Pre	22%	38%	< 0.01	0.01	> 30 s	RK Tutorial Pre	20%	40%	< 0.01	0.02
All	RK Example Pre	23%	45%	< 0.01	< 0.01	> 30 s	RK Example Pre	22%	51%	< 0.01	< 0.01
All	RK Quiz	36%	51%	< 0.01	0.12	> 30 s	RK Quiz	36%	52%	0.01	0.28
All	AM Tutorial Pre	49%	61%	0.02	0.48	≤ 30 s	AM Tutorial Pre	42%	68%	< 0.01	0.06
All	AM Example Pre	63%	77%	< 0.01	0.12	≤ 30 s	AM Example Pre	48%	75%	< 0.01	0.05

value adjustment [22] to account for elevated Type II error caused by conducting multiple tests, the difference between 2018 and 2017 students remained statistically significant for the Pre stage passing rates of the Tutorial and Example modules in the RK sequence. However, the difference on the Quiz module is no longer statistically significant. To ensure that the observed differences were not caused by the student population in 2018 being in general stronger than the student population in 2017, we checked their scores on four common problems given on a classroom exam that was administered shortly before students were assigned the modules. On none of the problems did 2018 students outperform the 2017 students.

As mentioned in section I, one extraneous factor that could have contributed to the observed differences is that some students may have obtained the answers to the problems from other sources. A thorough and complete investigation of such phenomena, such as in [13] is far beyond the scope of the current paper. However, we can still make a less precise estimation by assuming that students who spent less than 30 seconds are much less likely to have actually solved the problem. If the observed improvement in performance were due to increase in problem solving skills, then the difference should not be observed among students who spent less than 30 seconds on their attempts.

We performed the same analysis on students who spent ≤ 30 s and > 30 s on their AC attempts separately, and listed the statistically significant results at the $\alpha = 0.05$ level before p-value adjustment on the right side of Table I. On both RK Tutorial and RK Example modules, the differences in Pre stage passing rates remain significant for the > 30 s population, but not the ≤ 30 s population. On the other hand, for the same two modules in the AM sequence, the differences were only marginally significant for the ≤ 30 s population after p-value adjustment.

IV. DISCUSSION

For RQ1, we found no statistically significant differences in subsequent Quiz module performance between students subjected to the CC and TT conditions. There are several possible explanations for this observation. First, unlike many previous studies that relied on explicit self-explanation [6, 9, 10], the current implementation of the compare and contrast tasks via multiple choice problems may not be sufficient to engage many students in an authentic and productive compare-contrast process. Second, the current compare-contrast problems focused on identifying the correct mathematical expression applicable in each situation which may have encouraged rote memorization and a "plug-and-chug"

approach. Future implementations may explore more effective ways to focus on contrasting surface feature differences and comparing deep structure similarities. Finally, the current study design incorporated the Quiz module as the sole assessment for detecting potential differences. The problems in the Quiz module could have insufficient discrimination power to detect any differences.

For RQ2, the consistent and significant increase in passing rate from 2017 to 2018 in the Pre stages of the RK Tutorial and RK Example modules serve as evidence for the benefit of the on-ramp module, since students have not yet accessed the IC of the two modules in the Pre stage. The fact that the improvement is significant for the > 30 s group and not for the ≤ 30 s group further suggests that the improvement is likely due to actual increase in problem solving ability, rather than extraneous factors. On the other hand, the difference in the AM sequence is not only less significant, but also mainly due to the increased passing rate in the ≤ 30 s group. This observation suggests that the benefit of an on-ramp module is not uniform, and can depend on various other factors such as content, assessment and the implementation.

Another noteworthy observation is that, despite the addition of two new modules, the improvement in passing rates over the 2017 population on the RK Quiz module is marginal at best, and insignificant for the AM Quiz module. For the RK sequence, it is likely that Atwood machine problems remain very challenging for the student population involved, since the RK Quiz module passing rate remains at 50%, despite the improvements seen on previous modules. For the AM sequence, however, it could be that the problems are not challenging enough, since the AM Quiz module passing rate in 2017 was already at 75%, leaving little room for further improvement. In addition, there is a noticeable drop in passing rate on AM Module 4 Pre, compared to the Pre stage of the two preceding modules. More careful analysis shows that most wrong answers concentrated on one of the distractors which is written in a potentially confusing way. This might have negatively impacted the effectiveness of AM Module 4.

Overall, our observations suggest that even when instructional resources are created based on well established principles from learning science, their effectiveness can be highly sensitive to implementation details and other factors such as the difficulty of the content. This observation calls for new methods that can quickly evaluate the effectiveness of new instructional materials and pedagogical innovations in order to reliably improve the quality of instruction. The current study demonstrated two such methods using OLM sequences and analysis of student log data. The first is a controlled AB experiment conducted in a single course, similar to the design in two previous studies [5, 23]. The second method involves implementing improved instructional design in the same class for consecutive years.

Compared to conventional clinical or classroom experiments, both new methods are much easier to implement and replicate, allow more flexibility in study design, and provide rich information on students' learning behavior that could support in depth data analysis. The first method has the advantage of reducing the number of extraneous impacting factors, but it is also more disruptive for students and more logistically complicated to implement. It also resulted in a smaller sample size since each group contains only half of the student population. The second method provides twice the sample size and is far less intrusive to implement as part of normal instruction. However, the results are subjected to the influence of more extraneous factors, which require sophisticated data analysis procedures to validate.

A. Future directions

Several follow up analysis can be carried out in the future to gain new insight into the current results and address some of the shortcomings of the current study. First of all, the current analysis for most part did not take into account the variation in students' level of engagement with the instructional material, as some students spent significantly less time studying the IC than others. Future analysis could probe the relationship between the level of engagement with the IC and students' performance on subsequent problem-solving task.

Secondly, since the passing rate on the Pre stages for each module are used as a measure of the effectiveness of instruction, future studies should investigate what fraction of students took the initial attempts seriously before accessing the IC . Preliminary analysis on the duration of the attempts has shown that some students seem to consistently submit a random guess on their first attempt, likely because of knowing that the IC will guide them towards solving the AC problem. Since this type of behavior significantly reduces the validity of using Pre stage attempts as assessments, future studies need to explore methods to encourage serious problem solving, such as providing small amounts of credit incentives.

Finally, an important topic for future analysis is to examine whether the observed benefits of new interventions are uniform across the student population or selectively benefit students of, e.g., certain demographic backgrounds.

ACKNOWLEDGMENTS

The authors would like to thank the Learning Systems and Technology team at UCF for developing the Obojobo platform. This research is partly supported by NSF Award No. DUE 1845436.

[1] H. S. Broudy, Types of knowledge and purposes of education, in *Schooling and the Acquisition of Knowledge*, edited by R. C. Anderson, R. J. Spiro, and W. E. Montague (Routledge, 1977) pp. 1–17.

[2] D. K. Detterman, The case for the prosecution: Transfer as an epiphenomenon, in *Transfer on Trial: Intelligence, Cognition, and Instruction*, edited by D. K. Detterman and R. J. Sternberg (Ablex Publishing, 1993) pp. 1–24.

[3] J. D. Bransford and D. L. Schwartz, Rethinking transfer: A simple proposal with multiple implications, Review of Research in Education **24**, 61 (1999).

[4] M. T. H. Chi, R. Glaser, and E. Rees, Expertise in problem solving: Advances in the psychology of human intelligence, in *Advances in the Psychology of Human Intelligence*, Vol. 1, edited by R. J. Sternberg (1982) pp. 1–75.

[5] Z. Chen, K. M. Whitcomb, and C. Singh, Measuring the effectiveness of online problem-solving tutorials by multi-level knowledge transfer, in *Physics Education Research Conference 2018*, PER Conference, edited by A. Traxler, Y. Cao, and S. Wolf (Physics Education Research Topical Group and the American Association of Physics Teachers, Washington, DC, 2018).

[6] D. Gentner, J. Loewenstein, and L. Thompson, Learning and transfer: A general role for analogical encoding, Journal of Educational Psychology **95**, 393 (2003).

[7] D. L. Schwartz, C. C. Chase, M. A. Oppezzo, and D. B. Chin, Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer, Journal of Educational Psychology **103**, 759 (2011).

[8] J. Roelle and K. Berthold, Effects of comparing contrasting cases and inventing on learning from subsequent instructional explanations, Instructional Science **44**, 147 (2016).

[9] R. Badeau, D. R. White, B. Ibrahim, L. Ding, and A. F. Heckler, What works with worked examples: Extending self-explanation and analogical comparison to synthesis problems, Physical Review Physics Education Research **13**, 1 (2017).

[10] E. Kuo and C. E. Wieman, Toward instructional design principles: Inducing Faraday's law with contrasting cases, Physical Review Physics Education Research **12**, 010128 (2016).

[11] B. D. Mikula and A. F. Heckler, Framework and implementation for improving physics essential skills via computer-based practice: Vector math, Physical Review Physics Education Research **13**, 010122 (2017).

[12] N. T. Young and A. F. Heckler, Observed hierarchy of student proficiency with period, frequency, and angular frequency, Physical Review Physics Education Research **14**, 10104 (2018).

[13] G. Alexandron, J. A. Ruiperez-Valiente, Z. Chen, Pedro J. Muñoz-Merino, and D. E. Pritchard, Copying @ scale: Using harvesting accounts for collecting correct answers in a MOOC, Computers & Education **108**, 96 (2017).

[14] *Chegg*, Chegg Inc., Santa Clara, CA (2019).

[15] Z. Chen, G. Garrido, Z. Berry, I. Turgeon, and F. Yonekura, Designing online learning modules to conduct pre- and post-testing at high frequency, in *Physics Education Research Conference 2017*, PER Conference (Physics Education Research Topical Group and the American Association of Physics Teachers, Cincinnati, OH, 2017) pp. 84–87.

[16] Z. Chen, S. Lee, and G. Garrido, Re-designing the structure of online courses to empower educational data mining, International Educational Data Mining Society (2018).

[17] Z. Berry, I. Turgeon, and F. Yonekura, *Obojobo Next* (2019).

[18] S. DeVore, E. Marshman, and C. Singh, Challenge of engaging all students via self-paced interactive electronic learning tutorials for introductory physics, Physical Review Physics Education Research **13**, 010127 (2017).

[19] C. Singh and D. Haileselassie, Developing problem-solving skills of students taking introductory physics via web-based tutorials, Journal of College Science Teaching **39**, 42 (2010).

[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2019).

[21] H. Wickham, *tidyverse: Easily Install and Load the 'Tidyverse'* (2017), R package version 1.2.1.

[22] Y. Benjamini and Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the Royal Statistical Society: Series B (Methodological) **57**, 289 (1995).

[23] Z. Chen, N. Demirci, Y.-J. Choi, and D. E. Pritchard, To draw or not to draw? Examining the necessity of problem diagrams using massive open online course experiments, Physical Review Physics Education Research **13**, 010110 (2017).