Measuring the effectiveness of online problem-solving tutorials by multi-level knowledge transfer

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

[1]*Department of Physics, University of Central Florida, Orlando, FL, 32828*
[2]*Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA, 15260*

This study presents a new method for assessing the effectiveness of instructional resources using online learning technology that provides much richer information than a traditional summative assessment. By requiring students to complete a sequence of problem solving and learning activities in a given order, this new method not only measures students' ability to directly transfer learning to a new problem, but also their ability to learn from additional resources, or the "preparation for future learning" effect. We used this method to evaluate the quality of two problem solving tutorials, and found that both tutorials significantly benefit transfer to nearly identical problems, but only one facilitates transfer to a further distance. Moreover, we found evidence suggesting that one tutorial prepared students with lower prior knowledge to learn as effectively from a following worked example as students with higher prior knowledge.

I. INTRODUCTION

One of the major goals of physics instruction is to help students develop a robust understanding of physics knowledge and be able to transfer this understanding to solve problems in a different context. However, selecting good transfer problems to assess the effectiveness of instruction can be a challenging task for instructors. If the problem is too similar to the instruction, or in other words, if the transfer distance is too small, then students may be able to solve it by rote memorization or by using a plug and chug approach. If the transfer distance is too large, then many students with moderate learning gains may not be able to solve the problem. More importantly, in many cases engaging with learning or problem solving activities may not directly lead to improvements on transfer tasks [1, 2]. Rather, the benefit may be reflected in so-called "preparation for future learning" (PFL), where students are better prepared to learn more effectively from follow-up instructional materials and perform better on future transfer tasks [3].

The difficulty may be partly alleviated by including multiple problems at different transfer distances on the same assessment, but this method also has several drawbacks. First, longer exams may reduce students' test taking effort, especially when used as a not-for-credit pre-test [4]. Second, on summative assessments students receive no feedback on their performance, and therefore a small mistake may lead to wrong answers on all similar problems, especially in computer graded multiple-choice assessments. Moreover, traditional summative assessments still cannot detect the effects of PFL.

Modern online learning technology enables instructors to design more creative forms of assessments that extend beyond the traditional definition of a "quiz" or an "exam." On one hand, students can be presented with a set of activities, including both learning and problem solving, that they must complete in a given order. On the other hand, students' behavior data such as time-on-task can be combined with performance scores to provide richer and more accurate information on the state of their knowledge and skill. In this study, we will utilize those advantages to design and test a new form of assessment that aims at remediating the aforementioned drawbacks of traditional summative assessments.

This new form of assessment is based on an online instructional design called Online Learning for Mastery (OLM) that implements the "mastery learning" strategy [5] in an online environment. The OLM design breaks down each topic into a sequence of modules each containing both learning resources and a short mastery test. Students can move onto the next module after passing the mastery test of the current one, and each student can proceed at their own pace through the sequence [6].

In the current study, we will assess the effectiveness of two online physics problem solving tutorials [7, 8] using a sequence of three OLM modules. The design not only measures students' performance on direct transfer tasks, but also provides students with an additional opportunity to learn from a worked example before attempting a second transfer task, which can be treated as a simple form of PFL.

This exploratory study will seek to answer the following research questions:

RQ 1. Can this new form of assessment provide richer information on students' learning than a single summative assessment by measuring both direct transfer and PFL?

RQ 2. How do students' learning efforts on the tutorial correlate with their performance on different problem solving tasks?

RQ 3. To what extent can students who learned from the tutorials perform at a similar level as those who already knew how to solve this type of problem?

II. METHODS

A. Design of the OLM module sequence

Each OLM module contains an instructional component (IC) and an assessment component (AC) (Fig. 1). A key feature of the design is that upon accessing a new module, students are first required to make at least one attempt at the AC before being given access to the IC. Each student can have multiple attempts on the AC, and can move onto the next

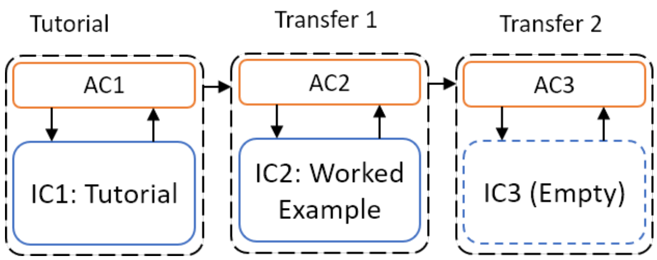

FIG. 1. The design of an OLM sequence consisting of three modules. Students can move on to the next module if they pass the AC, regardless of whether they have accessed the IC.

module whenever they pass the AC. On each new attempt, a slightly different version of the assessment problem is presented to the student. After the initial attempt, students can either study the IC or make additional attempts on the AC. The IC of OLM modules can contain a variety of learning resources including text, figures, videos, and practice problems with hints and solutions. During each AC attempt the IC is temporarily blocked from access.

Due to platform limitations, in the current study a student can move on to the next module after making an attempt on the AC of the current module. However, students were not explicitly informed of this fact, and were strongly encouraged to complete one module (by either passing the AC or depleting all available attempts) before moving on to the next module. Almost all students followed this order.

B. Measuring transfer through a sequence of OLM modules

The structure of the three module sequence is illustrated in Fig. 1. The ACs of all three modules (AC1 - AC3) contain one problem each. The three problems, perceived by experts as having similar difficulties, can be solved by applying exactly the same physics equation(s), yet are very different in surface features. The IC of the first module (IC1) contains the online tutorial implemented as a sequence of practice problems, teaching students how to solve the problem in AC1. The IC of the second module (IC2) contains a detailed worked example of the same problem used in AC2. IC3 is essentially empty except for one sentence stating that no hint or solution is given for this module.

We will refer to students' attempts on AC1 and AC2 prior to accessing IC1 and IC2, respectively, as "pre-learning" attempts (or PRE for short), and attempts after accessing the IC as "post-learning" (POST). Those attempts serve as de-facto pre- and post-tests for IC1 and IC2. Note that we do not distinguish between the attempts on AC3 since IC3 is empty. As students move through the modules in order, AC1-PRE serves as the pre-test for the entire sequence, while AC1-POST measures students' ability to solve an essentially identical problem after studying the tutorial, which can be viewed as transfer at near "zero" distance. AC2-PRE serves as the first transfer task, and AC2-POST examines students' "zero-distance" transfer from the worked example (IC2). For those students who studied IC2, AC3 serves as a second transfer task that measures the learning outcome from IC2. For students who studied both IC1 and IC2, their performance on AC3 can be seen as a measure of the PFL effect. We will refer to the three modules as "Tutorial" (Tut.), "Transfer 1" (Trans.1), "Transfer 2" (Trans. 2).

C. Study Setup

Two OLM sequences were created and assigned as part of weekly homework for a college introductory physics course on Newtonian mechanics at the University of Central Florida. The IC1 of each sequence contained a research based interactive tutorial developed by DeVore and Singh [7, 8], which guides students through solving a challenging problem on **rotational kinematics (RK)** and **conservation of angular momentum (AM)**, respectively. The student body consisted of 22% female students and 46% minority students.

Students received 2 points each for solving the problem in AC1 and AC2, and received 1 point for solving the problem in AC3. No credit was assigned for interacting with the IC. The problem bank of each AC contained 4 isomorphic problems that were different only by numbers. Each student was given 5 attempts on each AC, and on the 5th attempt the student receive the 1st isomorphic problem for a second time.

The OLM modules were created and hosted on UCF's award-winning open source online learning objects platform, Obojobo, developed by the Learning System and Technology (LS&T) team at the Center for Distributed Learning [9]

D. Data Collection and Analysis

Student's performance on AC1 and AC2 are divided into four categories: 1. **Pass on PRE**: Students who pass within two pre-learning attempts 2. **Pass on POST**: Students who passed within two post-learning attempts. 3. **Fail on POST**: Students who did not pass within two post-learning attempts. 4. **Other**: Other types of outcomes such as taking three or more attempts on the AC before studying the IC. Few students ($\sim$10%) belong to the Other category.

Students' performance on AC3 is divided into **PASS** (pass within 2 attempts) and **FAIL** (pass outside of 2 attempts or eventually failed).

Student's interaction with the IC is captured by the duration of their longest study session (LSS). A "study session" is the sum of all interactions with the IC that took place between two AC attempts. Only in about 5% of the cases did a student make a second study session that was at least 30% as long as the longest one. In those cases, the duration of the second session was added to the first one and the AC attempt in between was neglected.

III. RESULTS

The number of students who made at least one attempt on the AC of each module is largely constant for both OLM sequences, as shown in Table I. To measure students' problem solving performance we plot in Fig. 2 the passing rates for

TABLE I. The number of students who made at least one attempt on each AC.

Sequence	Module	Assessment	N
RK	Tutorial	AC1	210
	Transfer 1	AC2	209
	Transfer 2	AC3	203
AM	Tutorial	AC1	202
	Transfer 1	AC2	200
	Transfer 2	AC3	200

each AC, distinguishing between PRE and POST attempts. We included both Pass on PRE and Pass on POST in the passing rate of AC-POST, to reflect the total fraction of students who were able to solve the AC problem after being given access to the IC.

We compared the difference between each pair of passing percentage using McNemar's exact test. All POST attempts have significantly higher passing rates than PRE attempts ($p < 0.05$). For the AM sequence, both AC2-PRE and AC3 have significantly higher passing rate than AC1-PRE, but the difference between those two are not significant. For the RK sequence, the passing rate on AC3 is significantly higher than both AC1-PRE and AC2-PRE ($p < 0.05$).

To understand how students' learning time on ICs correlates with their performance on subsequent AC attempts, we first investigated the distribution of students' learning time (LSS duration). For the ICs of the first two modules in each sequence, the distributions of LSS duration are not significantly different from a log-normal distribution (Shapiro-Wilk normality test), except for RK-IC1. For the current study, we used a simple cut-off by labeling students with log LSS duration less than one standard deviation below the mean as "brief learners" and the rest of the population as "normal learners" (see example in Fig. 3).

We tested the correlation between students' IC learning behavior ("brief" or "normal") and their passing rate on following ACs using Fisher's exact test. The resulting p-values are listed in Table II. We note that IC learning behavior is only significantly correlated with students' performance on AC-POST of the same module (except for AM-IC1 and AC1-POST, which is marginally significant at $p = 0.07$). Correlation with performance on transfer or PFL tasks were not significant.

FIG. 2. Passing rates for each AC on both OLM sequences.

TABLE II. p-values from Fisher's exact test on 2×2 contingency tables measuring the correlation between study time ("brief" or "normal" learners for the ICs) and assessment outcome ("pass" or "fail" on the ACs). Statistically significant results are bolded. * indicates $p < 0.05$ and ** indicates $p < 0.01$.

	RK		AM	
AC	IC1	IC2	IC1	IC2
AC1 PRE	—	0.48	—	0.10
AC1 POST	**0.03***	0.22	0.07	0.48
AC2 PRE	1.00	—	0.37	—
AC2 POST	0.75	**0.00****	0.23	**0.00****
AC3	0.81	0.24	0.55	0.42

To answer RQ3, we look at two groups of students who interacted differently with the Tutorial modules: "Pass AC1-PRE" are those who passed AC1 on pre-learning attempts and skipped IC1, and "Pass AC1-POST" are the "normal learners" on IC1 who passed AC1 after learning. We plot their performance on the following transfer task (AC2-PRE) and the PFL task (AC3) in Fig. 4. In addition, we selected from both groups those who are "normal learners" on the worked example (IC2), and compared their performance on the PFL task (AC3).

As shown in Fig. 4, in both sequences the students who passed AC1 on pre-learning attempts still performed significantly better than those who learned to solve AC1 from IC1. The difference is much smaller on AC3 for the RK sequence, but remains significant for the AM sequence. However, for the two sub-populations who both studied IC2, the difference in AC3 performance (PFL task) is much smaller.

IV. DISCUSSION

For both sequences, students had high success rates on "zero-distance" transfer tasks (AC1-POST, AC2-POST). The fact that normal learners significantly outperformed brief-learners on those tasks further verifies that studying the tu-

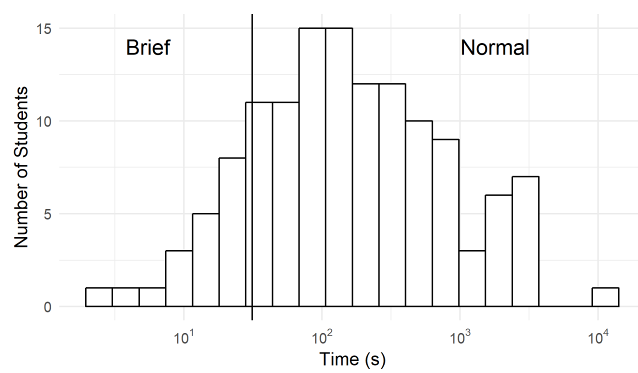

FIG. 3. A histogram of the log-scaled LSS duration for RK-IC2. The vertical bar (31 s) denotes one standard deviation below the log mean for this module, which is used as a cutoff between the "brief" and "normal" learners.

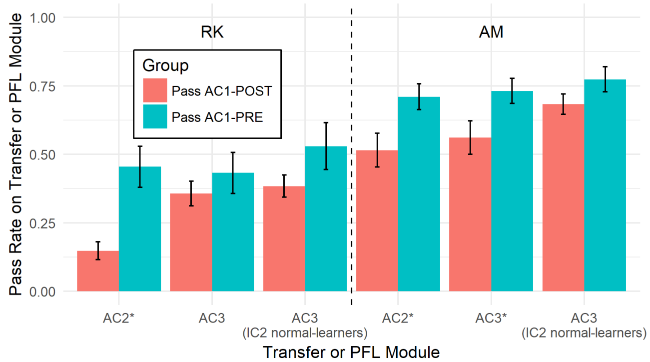

FIG. 4. Passing rates on AC2 and AC3, comparing students who passed AC1 on PRE with those who passed AC1 on POST after studying IC1. The third set in each sequence is plotted for a subset of students who also studied IC2 and passed AC2-POST. Statistically significantly differences are marked with *.

torial helped students learn to solve the particular problem being taught. However, the impact of the tutorial on transfer and PFL tasks is much smaller.

For the RK sequence, students' passing rate on the first transfer task is identical to that of the pre-test, while their performance on the PFL task is significantly higher. However the passing percentages on all three ACs are very low, and not significantly different from a random guessing rate. Overall, the results suggest that the RK problems are challenging for the current student population, and that the tutorial alone did little to facilitate transfer, which is consistent with previous research [8].

For the easier AM sequence, students' overall performance on both transfer tasks (AC2-PRE, AC3) is significantly better than on the pre-test (AC1-PRE), indicating that some students were indeed able to transfer their learning to a new problem. However, the performance of "brief-learners" are indistinguishable from that of the "normal-learners." One possible explanation is that students' prior knowledge plays a more important role in transfer tasks, since "brief-learners" on IC1 could include students with stronger prior knowledge who only needed to skim through the tutorials. This explanation is also supported by the fact that those who could pass AC1 before learning the tutorial (strong prior knowledge) consistently performed better on both transfer tasks than those who learned how to solve AC1 from the tutorial (Fig. 4). Finally, a noteworthy observation is that on the PFL task (AC3), those who studied both the tutorial and the worked example performed similarly to those who passed AC1 on pre-learning attempts and only studied the worked example. This could imply that studying the AM tutorial prepared students with weaker prior knowledge to learn from the worked example as much as their peers with stronger prior knowledge. Additional future studies are needed to verify this implication.

Overall, we found that the three-module design can provide richer information than a single summative assessment using either one of the AC problems or all three problems together. With this design, we can not only detect the performance difference between near and far transfer tasks, but also detect more subtle benefits of the tutorial on PFL tasks. The incorporation of students' learning time as part of the analysis further improves the reliability of this assessment method.

Finally, there are a few shortcomings in the current study that should be addressed in future studies. First, we assumed that the three AC problems in the same sequence have similar difficulties based on expert opinion. This needs to be verified in future studies by switching the order of the problems. Second, we did not account for the effect of answer copying or random guessing behavior among students, which can have an impact on the results. Future studies can either be conducted in a proctored setting, or exclude students who spent an unusually short time answering the AC problems [10]. Third, a few students studied the IC after passing the AC on some modules. Although there were only a couple of cases, they should be properly considered in more careful analysis in the future.

ACKNOWLEDGMENTS

We would like to thank, Dr. Francisca Yonekura, Dr. Ian Turgeon, and Dr. Zachary Berry from the Center for Distributed Learning at UCF for developing the Obojobo platform and providing technical support for data collection and analysis.

[1] H. S. Broudy, 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, 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, Review of Research in Education **24**, 61 (1999).

[4] S. L. Wise, Applied Measurement in Education **19**, 95 (2006).

[5] J. H. Block and R. B. Burns, Review of Research in Education **4**, 3 (1976).

[6] Z. Chen, S. Lee, and G. Garrido, in *11th International Educational Data Mining Conference* (Buffalo, NY, to appear).

[7] C. Singh and D. Haileselassie, Journal of College Science Teaching **39**, 42 (2010).

[8] S. DeVore, E. Marshman, and C. Singh, Phys. Rev. Phys. Educ. Res. **13**, 010127 (2017).

[9] C. Bishop, F. Yonekura, and P. Moskal, in *Proceedings of the Association of College & Research Libraries 2013* (2013) pp. 466–471.

[10] Z. Chen, G. Garrido, Z. Berry, I. Turgeon, and F. Yonekura, in *Physics Education Research Conference 2017* (Cincinnati, OH, 2017) pp. 84–87.