Learning from different styles of narrated-animated solutions among low-performing students

Jose P. Mestre¹²³, Jason W. Morphem², and Gary Gladding¹

¹ University of Illinois at Urbana-Champaign, Department of Physics,
   1110 West Green Street, Urbana, IL 61801
² University of Illinois at Urbana-Champaign, Department of Educational Psychology,
   1310 S. 6th St., Champaign, Illinois 61820
³ University of Illinois at Urbana-Champaign, Beckman Institute for Advanced Science and Technology,
   405 North Mathews, Urbana, IL 61801

Abstract. Students preparing for physics exams must make decisions on how to best prepare, and how to estimate their preparedness. Previous research shows that low-performing students tend to over-predict both their learning and level of preparedness. Providing students with normative feedback has been shown to reduce over-prediction. We present a pilot study where low-performing students in an introductory mechanics course completed a computer-adaptive practice exam, then completed a pretest followed by an intervention involving one of two styles of narrated, animated PowerPoint solutions, and then completed a post-test before taking a midterm exam. Findings suggest that students are able to learn from viewing animated solutions. We discuss the educational implications of our findings.

PACS: 01.40.Fk, 01.40.gb, 01.40.Ha, 01.40.-d

I. INTRODUCTION

We report on a preliminary study based on an ambitious project that combines testing/psychometrics, cognition, and physics education research to help low-performing students better prepare for exams in an introductory mechanics course for scientists and engineers. Research investigating students’ ability to predict their skill level indicates that predictions are much higher that their actual performance levels [1]. Further, inability to predict one’s own skill level differentially affects low-skill and high-skill students; in particular, low-performing students over-predict their competence level significantly, whereas high-performing students slightly under-predict their competence [2]. Consequently, low-skill students walk into exams with a false sense of confidence. Perhaps what is most surprising is that low-performing students’ think they outperformed the majority of their peers even after taking an exam, indicating that they lack the ability to use (as yet ungraded) assessments to judge their own skill level [3].

In physics, a study by Rebello [4] investigated postdictive ability of students in a second-semester calculus-based introductory physics course. Each time after taking 5 exams, students were asked to predict the class average and their own score and offered an incentive to improve their grade if their predictions were accurate. Findings were similar to the studies cited above from other disciplines, namely high-performing students underestimated their performance by 2.9%, and low-performing students overestimated their performance by 8.4%.

In terms of explanations for why low-performing students are poor predictors of their skill level, it might be that low-performers lack metacognitive monitoring ability in specific domains—for example they may be accurate in predicting their skill level in humanity or even certain STEM courses but not in physics. Others have argued that familiarity with the material may lead low-performing students to conclude that they are also competent with the material without any evidence for it [5]. In physics certain study habits, such as looking over worked-out homework or course notes, can lead to familiarity with the material but unless students test themselves by attempting to solve test-like problems they will not be able to accurately assess their preparation for an exam.

The project within which the pilot study was performed has two components. The first was to construct a cognitively-diagnostic computer adaptive testing (CD-CAT) platform that uses computer adaptive testing to diagnose student weaknesses prior to taking a real exam. The second component was to address weaknesses found by the CD-CAT. Thus the approach taken was to begin with a realistic assessment of the students’ skill level, and then to specifically address their weaknesses via an intervention.

II. PILOT STUDY DESIGN

Performance on the first hour exam was used to identify students performing in the bottom third of the class (N=332), and an email was sent inviting them to participate in a “test-preparation study” for exam 2. Volunteers participated in an intervention lasting approximately 2.5 hours, which began by taking a CD-CAT test followed by an intervention. Study details are provided below.

A. Construction of CD-CAT Delivery Platform

A test item-bank was constructed by selecting problems from previously administered second hour exams that adhered...
to psychometric properties appropriate for use in CAT. A CAT administers questions to students according to an algorithm that combines question discrimination and the student’s performance on previous questions, adjusting the level of difficulty of a question accordingly to home in on the students’ skill level. We worked with psychometric experts who had extensive experience with developing CAT exams to develop an item bank of 222 items which had been previously administered in offerings of the course. A web-based delivery platform was also constructed that allowed students to take the CD-CAT prior to the intervention. The CD-CAT platform administers the CD-CAT and provides the student with a predicted score on the second midterm exam (with error bars) based on their performance.

B. Description of Interventions

Students entered the intervention knowing their predicted scores from the CAT, (which were low in general since most students had not begun to study for the exam). There were three parts to the intervention, two 10-question assessments, and animated worked-out solutions to the problems in the two assessments. Both assessments consisted of 10 questions of medium to hard level of difficulty that were typical of questions on the second midterm exam. The two assessments were isomorphic, in that every question in one of the assessments had a very similar question in the other. Figure 1 has an example of two isomorphic questions from the two assessments.

Two somewhat different narrated, animated worked-out solutions were constructed for the problems in the two assessments, which will be referred to as Type 1 and Type 2. Both types consisted of narrated PowerPoint solutions with animations where appropriate that showed a step-by-step approach for solving the problems. Type 1 took a somewhat reflective approach, asking students to think about what concepts and procedures might be required to solve the problem and sometimes deriving results to be used in the solution. Equations were shown with variables and combined when needed; no “plugging in” of numerical values to get answers was ever done, but instead the final numerical answer was provided from the resulting equation and students were left to verify that the answer could be generated from the final equation derived. These solutions took a general approach, such that the solution was not specific to the numerical values used in the problem.

Type 2 solutions were more explicit and employed a “two-column” approach [6-7], with the left column describing the principle, concept or procedure that was being applied sequentially to solve the problem and the right column describing the equation resulting from the left column. Unlike Type 1 solutions, algebraic steps and numerical substitutions were explicitly carried out in Type 2 solutions. These solutions took a more specific approach, such that the solution substituted the numerical values from the problem. Both types of solutions attempted to adhere to multimedia learning principles [8]; for example, text/equations on the screen were not read but rather explained or elaborated, and procedural steps were illustrated in figures with flashing arrows or other highlighting (e.g., when discussing initial and final states in an application of conservation of mechanical energy).

C. Procedure

Fifty-one students volunteered for the study and 46 students completed all components of the study. They were split into two equivalent groups based both on Exam 1 performance and on CD-CAT performance using matched pair random assignment. One of the groups received Type 1 animated solutions and the other group received Type 2. All the students took one of the two assessments as a pre-test and the other as the post-test to ascertain how much learning took place as a result of the animated solutions; the order of the pre- and post-tests were counterbalanced among the two groups.

Study participants were asked to take the CD-CAT online between Friday and Sunday (45-60 minutes needed to complete the CD-CAT), and on Monday or Tuesday they came in for the intervention lasting about 1.5 hours. Students in both groups completed the pre-assessment then were told which of the 10 problems they got wrong. They were asked to at least view the narrated/animated solutions for the problems they got wrong, and were invited to watch all of them to review for the midterm exam. Finally, they completed the post-assessment.
The narrated/animated solutions to the post assessment were made available to the students following the intervention to view as they wished. At the end of the intervention students took a short survey that asked how much time they had spent studying thus far and what methods they have used to study (e.g. reviewing the notes, taking practice exams). The second midterm exam was administered in the evening of Wednesday of that week.

III. ANALYSIS AND RESULTS

Table 1 contains a comparison of the percentage correct scores for the two groups on the pre and post assessments. To determine whether the type of video solution or the assessment order influenced the gains from pre to post, a 2x2x2 (Type x Order x Test) mixed ANOVA with test as the repeated measure was conducted. A main effect of test was significant \( F(1,42) = 31.44, p < .01 \), but not group or order. This indicates that there were significant increases in score from the pre to the post assessment (see Fig. 1), but there were no differences between the groups or assessment order. In addition there was not a significant interaction between group and test \( F(1,42) = .60, p = .59 \) which indicates that no significant differences in the gains were observed between the two intervention styles.

Table 1. Means and SD for Assessments by Intervention Type

<table>
<thead>
<tr>
<th>Intervention</th>
<th>N</th>
<th>Assessment</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1</td>
<td>22</td>
<td>Exam 1</td>
<td>69.5</td>
<td>11.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CAT</td>
<td>52.5</td>
<td>12.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pre</td>
<td>34.1</td>
<td>19.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Post</td>
<td>49.1</td>
<td>23.5</td>
</tr>
<tr>
<td>Type 2</td>
<td>24</td>
<td>Exam 1</td>
<td>72.4</td>
<td>9.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CAT</td>
<td>53.6</td>
<td>13.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pre</td>
<td>39.6</td>
<td>19.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Post</td>
<td>57.9</td>
<td>23.8</td>
</tr>
</tbody>
</table>

While no significant differences were observed in the overall gains from the pre-assessment to the post-assessment, we can use the fact that each question on the post-assessment was isomorphic to a question on the pre-assessment to measure the effect of the solutions on each question individually. We can define the efficacy of a solution for a given question as the fraction of students who answered that question incorrectly on the pre-assessment but correctly on the post-assessment, given that they viewed the solution after completing the pre-assessment. In our study, students who answered a question incorrectly did view the solutions about 88% of the time. We defined “viewing the solution” as spending at least 30% of the narration time for the solution. Figure 3 shows a plot of the solution efficacy for each question for both solution type 1 (S1) and solution type 2 (S2).

The only significant differences that were observed between the two solution types are seen in questions 9 and 10. These questions involve the situations shown in Fig. 1 which involve both gravitational and spring potential energy as well as motion through a region containing friction. Different solution strategies were used in these questions. In solution type 1, the questions were broken down into solving for the motion in the three distinct regions sequentially. In solution type 2, the work-energy theorem was used to solve the problems in one step. Our interpretation of this result is that one step solutions are easier for students to learn than multi-step solutions, likely due to reduced memory load. For the remaining 8 questions, identical solution strategies were used in each solution type. If we sum the results from these 8 questions, we determine that the difference in the average solution efficacy for the two types is 1% ± 7%. Consequently, we can say that when the same solution strategy is used, there is no difference in the effectiveness of the solution types at the ± 7% level.

Finally, we address the question of whether our interventions led to an improvement on the midterm. Because there were no significant differences in learning with the two types of animated solutions, and because we had no control over how much studying students did after leaving our intervention, we combined the two groups and compared their scores on the second midterm to those of students who performed in the bottom one-third on the first midterm from which we drew our study participants. The 46 participants in our study averaged 71.0 ± 1.5 on Exam 1 whereas the non-intervention students averaged 71.9 ± 0.6. Since the Exam 1 data was non-normally distributed a Mann-Whitney U-test...
was conducted. A significant difference between the groups was not detected on Exam 1 ($z = .48, p = .63$). The scores on exam 2 were normally distributed so an independent sample t-test was conducted. The 46 participants in our study averaged 67.1 ± 2.3 whereas the non-intervention students averaged 59.6 ± 0.9. Participants who received an intervention scored higher on the second midterm exam ($t(322) = 3.01, p < .01$). The effect size ($d = .48$) indicates that this difference represents a medium effect.

### IV. DISCUSSION

Although the results look promising, with both interventions resulting in pre to post improvement on the assessments and improvement on the second midterm exam relative to similar students in the class using z-scores, there are more questions remaining to be answered. Because we are targeting low-performing students, there are challenges in finding both study strategies and interventions to help them.

Within the controlled portion of the study, the only significant difference that emerged between the two solution types occurred in questions in which the solution strategies used by the two types were different. In those questions in which the same solution strategies were used, we saw no evidence for any difference in performance. We note however, that our assessments were made only on isomorphic questions. It remains an open question as to whether there is a difference in the effectiveness of these solution types when the assessments require transfer of learning to non-isomorphic questions, either of a conceptual or of a quantitative nature. We plan to develop studies in the future that address these issues.

Another challenge is to get these underperforming students to start studying earlier on. The survey we administered following the intervention indicated that students had studied only 4 hours for an exam that would take place within 1-2 days’ time. This means that if they were going to devote more time to studying, it was going to be in a “cramming” mode.

Cognitive research indicates that distributed practice over longer periods of time is better for long-term learning than massed practice (cramming) [9].

We are planning future studies to address these and additional questions. For example, we are interested in exploring two cycles of CD-CAT+Intervention. The hope is that this will encourage the type of problem-solving-intensive studying needed for these students (as opposed to reviewing past work in the course, which builds familiarity with the material but does not evaluate competence), but also to build realistic confidence by showing that incremental improvements ensue from the first to the second CD-CAT administration and by showing additional pre-post gains in the mini-tests after each intervention. The CD-CAT should not only serve as a “reality-check” for students who may be overconfident due to familiarity with the material, but can also serve to motivate students to take a more active problem solving approach to their studying. We also would like to explore if the two types of narrated/animated solutions help different ability students differentially. For example, Type 1 solutions ask the student to fill in the gaps in terms of the algebra, and spending the effort to do so might help high-performing students learn the material better, whereas such activities might tax working memory for low-performing students.

### ACKNOWLEDGEMENTS

The authors are grateful to Hyeon-Ah Kang and Hua-Hua Chang for developing the CD-CAT algorithm and Gregory Fabry for developing the web-based platform for the CD-CAT.


