PHYSICAL REVIEW SPECIAL TOPICS - PHYSICS EDUCATION RESEARCH **4**, 010102 (2008)

# Measuring student learning with item response theory

Young-Jin Lee, David J. Palazzo, Rasil Warnakulasooriya, and David E. Pritchard
*Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
(Received 9 March 2007; published 31 January 2008)

We investigate short-term learning from hints and feedback in a Web-based physics tutoring system. Both the skill of students and the difficulty and discrimination of items were determined by applying item response theory (IRT) to the first answers of students who are working on for-credit homework items in an introductory Newtonian physics course. We show that after tutoring a shifted logistic item response function with lower discrimination fits the students' second responses to an item previously answered incorrectly. Student skill decreased by 1.0 standard deviation when students used no tutoring between their (incorrect) first and second attempts, which we attribute to "item-wrong bias." On average, using hints or feedback increased students' skill by 0.8 standard deviation. A skill increase of 1.9 standard deviation was observed when hints were requested after viewing, but prior to attempting to answer, a particular item. The skill changes measured in this way will enable the use of IRT to assess students based on their second attempt in a tutoring environment.

DOI: 10.1103/PhysRevSTPER.4.010102 PACS number(s): 01.40.Fk, 01.40.G−, 01.50.ht

## INTRODUCTION

This work stands at the intersection of two current trends in education: interactive learning environments and formative assessment of students during learning. The usefulness of learning environments that give relevant "when needed" guidance has been emphasized by many learning theories. For instance, cognitive apprenticeship asserts that teachers (experts) should provide "scaffolding" to students (novices) to help them solve items that they otherwise could not solve by themselves.[1] Advances in computer technology are enabling more students to experience individualized tutoring and simulation environments. An example is the widely used MASTERINGPHYSICS homework tutorial system,[2] used to collect the data in this experiment.

Recent National Academy studies,[3] among others, have called for more formative assessment, and for extending formative and summative assessment from testing environments to everyday learning environments. This presents a significant challenge for item response theory (IRT) because this widely used method of psychometric analysis assumes both that students have only one attempt at a given item, and furthermore that students do not learn while working through the items (i.e., while taking the test).

In this paper we propose and study a method to extend IRT to assess students in a learning environment where they are allowed multiple attempts at answering a particular item, often with learning activities in between. There are three major motivations for this study: quantifying the effectiveness of the learning activities, extending the range of applicability of IRT into the "multiple attempts with intervening learning" regime, and providing better formative assessment of students during an interactive learning process. Quantifying the effectiveness of the various tutoring material in the tutor program offers the hope that we can improve educational effectiveness by revising the tutoring material appropriately. Subsequently, one might revise the format or the grading algorithm to encourage students to follow along tutoring trajectories that maximize their learning outcome. Additionally, one might strengthen the content along the most effective or most popular learning routes. The extension of IRT to this new regime may also increase the reliability of computer-given standard tests, since giving students a second attempt after a wrong answer is straightforward in these environments and provides another gradable interaction at the cost of little additional student time. Finally, being able to reliably assess a student while (s)he is being tutored can provide formative as well as summative assessment without the need for interrupting the learning process for testing or some other form of assessment.

This paper presents a study of the effect of various forms of learning assistance on problem-solving performance in the Web-based homework tutor MASTERINGPHYSICS, which uses a Socratic tutoring style. (We use "problem" in the colloquial sense—most are like end of chapter problems in a typical introductory physics textbook as shown in the top of Fig. 1). This Socratic tutor provides help in several ways as students work toward the correct solution. If the student submits a wrong answer (even in the free response mode studied here), the program makes a useful response to that particular answer about half of the time (see the bottom left of Fig. 1). A list of hints is available upon student request at any time during the solution process. These hints are one of three types: a list of steps, declarative statements, and procedural subtasks. Declarative hints provide a description of what should be done, taken note of, or used to solve the item at hand (see the bottom middle of Fig. 1), while procedural subtasks ask students to work on simpler subtasks that are helpful steps toward solving the current item (see the bottom right of Fig. 1). In addition, MASTERINGPHYSICS may also ask follow-up questions and give comments to help the student understand the import of the just-obtained solution. To enable constructivist learning, students can open hints and subtasks, and answer problems and items in any order.

Given the complexity of student interactions with the feedback and hint structure, some simplification is necessary. Our simplification is to lump interactions into one of two groups: (1) received (useful) wrong answer feedback or (2) engaged hints and subtasks. We then considered various paths from start to second attempt. Using IRT, we first determine the skill of students and the difficulty of items based

 

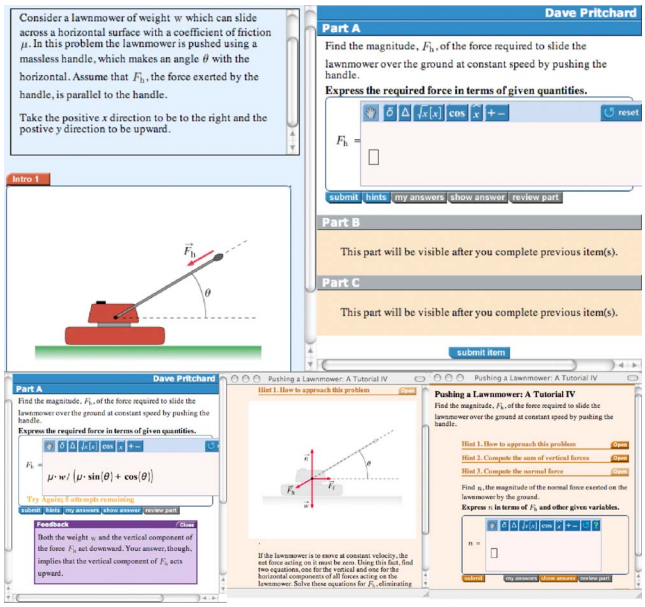


FIG. 1. (Color) Solving a "pushing a lawnmower" problem with MASTERINGPHYSICS (top); wrong answer feedback (bottom left), declarative hint (bottom middle), and procedural subtask (bottom right).

solely on the first response of the students to each item (in the spirit of IRT, we exempt students requesting a hint before attempting an answer from this analysis). Second responses on each item were then analyzed to find the changes in skill of students who had followed a particular path through the hints and feedback.

## METHOD

We first employed IRT using a two-parameter logistic model in which the probability for student s (who has skill $s_s$) to get an item $i$ correct, denoted as $P^{s,i}_{\text{correct}}$, is assumed to be

$$P^{s,i}_{\text{correct}} = \frac{1}{1+e^{-\alpha_i(s_s-d_i)}}. \tag{1}$$

In Eq. (1), each item is parametrized by a discrimination coefficient $\alpha_i$ and a difficulty coefficient $d_i$;[4–6] no allowance for guessing seems necessary for the free response answer type involved here. (To lower cognitive load for those unfamiliar with standard IRT notation, we have used $s$ and $d$ for skill and difficulty, respectively, rather than the conventional $\theta$ and $b$.) The difficulty coefficient $d_i$ on the skill axis is the point for which the predicted probability of correct response, $P^{s,i}_{\text{correct}}$, equals 1/2. The discrimination coefficient $\alpha_i$ is sometimes called a slope parameter because it is related to the slope of the linear portion of a logistic curve. The larger a discrimination coefficient is (in other words, the steeper a logistic curve is), the better the item discriminates (distinguishes) between students of low vs high skill.

The commercial IRT program[7] BILOG-MG was used to build this model from the first responses to 15 homework

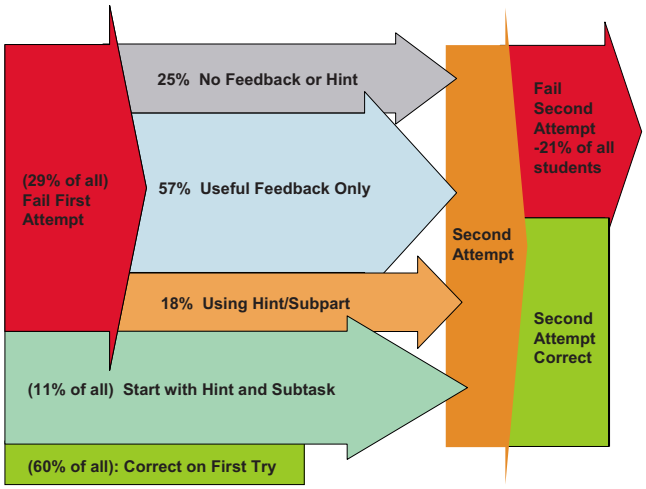


FIG. 2. (Color) Transition diagram showing several possible paths involving the first two answers.

problems with 58 total items by 142 students who took an introductory Newtonian physics course in the fall 2004 semester at Massachusetts Institute of Technology. All 58 items were free response questions (all requiring a symbolic response). Students were allowed to submit up to six responses, being provided with feedback and/or electing some (or no) intervening learning activities before each subsequent attempt at answering the main item. After the solution is obtained, the program may make follow-up comments or ask follow-up questions (which are treated like any other items).

Obviously, the variety of elective learning material implies that students' learning paths through the tutorial help can be quite complicated. Figure 2 shows a transition diagram specifying four common paths students take that involve some learning. Most tutoring paths involve a second attempt preceded by an incorrect first attempt followed by no useful feedback or hint (no feedback or hint); receiving only feedback (useful feedback only); and using hints and/or subtasks (using hint and/or subtask). We characterize students' paths through the hints only by whether they used declarative hints or opened any subtasks (completion of these is elective). The last path considered is using hints and subtasks before their first attempt (start with hint and subtask). We are interested to see how each of these particular paths affects students' skill on their second attempt, and expect that it will quantify the utility of hints and feedback to students.

For each of the paths in Table I (the paths are shown graphically in Fig. 2), we compute the change in skill on each item ($\delta s_i$) by fitting the end-of-path response data to Eq. (2),

$$\sum_{s=1}^{N_s} \frac{1}{1+e^{-\alpha_i[(s_s+\delta s_i)-d_i]}} = N_{\text{correct 2}}, \tag{2}$$

where the skill of students ($s_s$) and the item parameters of items ($\alpha_i$, $d_i$) have been previously found from first response data as indicated above. $N_{\text{correct2}}$ in Eq. (2) is the number of students who got the item correct out of the $N_s$ who followed that particular trajectory (e.g. one of those listed in Table I).

TABLE I. Statistics of changes in skill after various interactions with MASTERINGPHYSICS.

| MASTERINGPHYSICS interactions | When $\alpha_i$ is fixed to the value from first attempt data | | | When $\alpha_i$ is determined from the second attempt data | | |
|---|---|---|---|---|---|---|
| | $\alpha_i$ | $\delta s_i$ | $\chi^2_\nu$ | $\alpha_i$ | $\delta s_i$ | $\chi^2_\nu$ |
| No feedback or hint | 0.81 | $-0.66 \pm 0.14$ | 4.79 | 0.66 | $-1.02 \pm 0.22$ | 0.96 |
| Useful feedback only | 0.82 | $0.26 \pm 0.08$ | 2.94 | 0.55 | $0.34 \pm 0.10$ | 0.44 |
| Using hint and/on subtask | 1.02 | $0.56 \pm 0.11$ | 4.66 | 0.51 | $0.62 \pm 0.18$ | 0.95 |
| Starting with hint and/on subtask | 0.95′ | $1.20 \pm 0.10$ | 5.13 | 0.49 | $1.87 \pm 0.16$ | 0.96 |

This amounts to fitting these "second attempt" data to a similar logistic function as the first attempt, the only change being a shift toward lower or higher skill by the amount $\delta s_i$.

## JUSTIFICATION OF MODEL

A sample of binned data used in this procedure is shown in Fig. 3. The data points of Fig. 3 were obtained by binning observed second attempts of students who received some meaningful feedback from the tutor after their erroneous first attempt, but did not use any hints or subtasks before making their second attempt. The solid line shows the expected $P^{\text{correct}}$ from their skill determined as described from previous problems (before they used the meaningful feedback provided by the tutor), the dashed line represents the best logistic fit to their second attempts (after they received meaningful feedback to their particular incorrect first answer) with the discrimination parameter $\alpha_i$ fixed to the value obtained from the first response data, and the dotted line represents the best logistic fit to the second attempts when the discrimination parameter $\alpha_i$ is determined from the second response data.

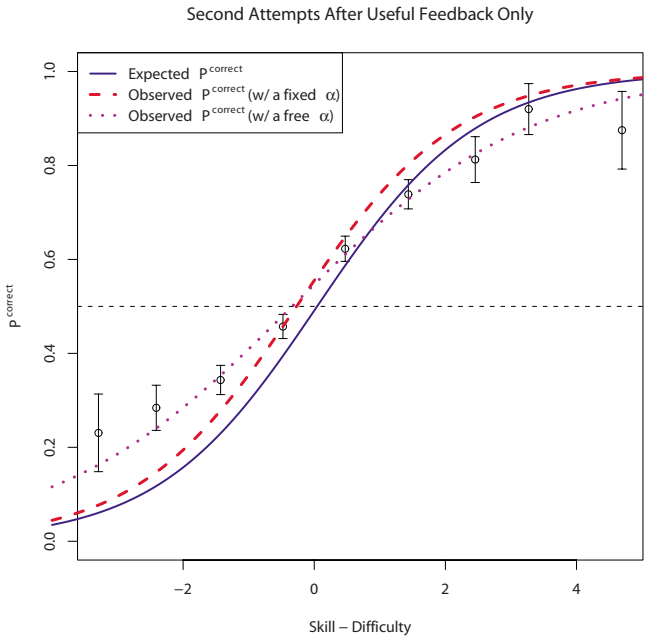


FIG. 3. (Color) Logistic fits to the second attempt of students after they received useful feedback from the tutor. The discrimination has decreased from 0.82 to 0.55.

The data in Fig. 3 are fitted by the logistic curve of Eq. (2). This shows that the overall skill of the students is the prime determinant of whether they can answer the posed question or not, even after receiving feedback. Since the feedback differs depending on the specific error, the learning value clearly is different for different students, which may well explain why the discrimination is lower after the feedback.

The fits here are typical and justify two fundamental points about the model: (1) The model fits all four data sets within error ($\chi^2_\nu < 1$) if $\alpha_i$ is varied. (2) The model does not fit the data if $\alpha_i$ is not varied.

This procedure is equivalent to fitting the difficulty of an item using only attempts made after the particular learning trajectory, and then setting $\delta s_i$ equal to the *decrease* in difficulty. We prefer to represent the learning as a path- (and skill)-dependent increase in skill because it is more consistent with the postulates of IRT. The difficulty of a free response item remains independent of the student or his (her) recent preparation; the increased probability of responding correctly is most logically attributed to a skill change that is local for that item. (Certainly the overall skill of the student does not change by one or two standard deviations due to five minutes of tutoring on one item.)

For a multiple choice item, the item would obviously be easier given that one of the attractive distractors has been eliminated on the first attempt; nevertheless, this change in difficulty of the item could be absorbed into $\delta s_i$ unless this skill change depends on the particular distractor chosen on the first attempt. (This paper uses only free responses to items requesting a symbolic response, so this point is moot.) The key point here is that we assume a skill increase due to the tutoring interactions that depends on both the type of trajectory and the effectiveness of the help available in that item along that particular trajectory.

In determining the error bars of data points in each bin, we assume that the number of students in each bin who got the item correct follows a binomial distribution. Under this assumption, the error bar of each data point is given as $\sqrt{p(1-p)/N_{s*}}$ where $p$ is $N_{\text{correct2}*}/N_{s*}$, $N_{\text{correct2}*}$ is the number of students in the current bin who got the item correct, and $N_{s*}$ is the total number of students in that bin. Consequently, if every student who followed a particular route in a

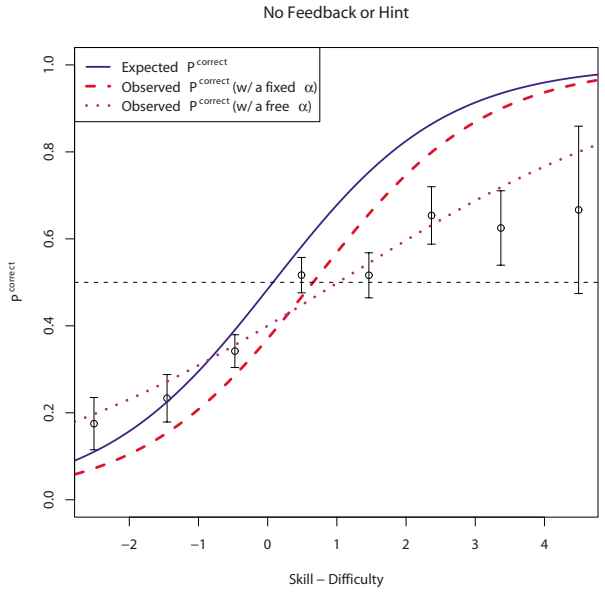


FIG. 4. (Color) Logistic fits to the second attempt of students when students neither received feedback from the tutor nor used hint or subtask.

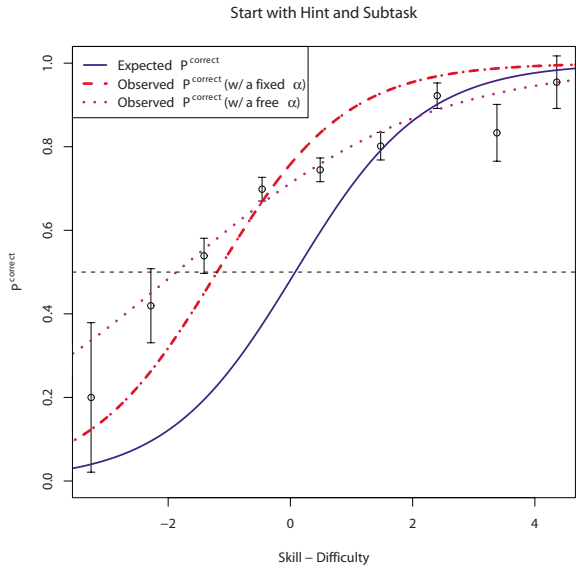


FIG. 5. (Color) Logistic fits to the first attempt of students used hint and/or subtask before their first attempt.

particular bin gets an item correct or wrong in the subsequent attempt, that bin is assigned a zero error. In such extreme cases, we used $p=1/(3N_{s*})$ instead, as suggested by Kim *et al.*, to assign nonzero error bars.[8]

## RESULTS AND DISCUSSION

The techniques described above were used to fit data along the other paths indicated in Table I (see Figs. 3–5).

First of all, it is noteworthy that students making a second attempt to answer without receiving either useful feedback or requesting hints showed a skill of $-1.02\pm0.22$ standard deviation below expectation (see Fig. 4). Surely receiving no useful information does not make these students less skillful. Therefore we attribute this decrease to "item-wrong bias." Any item is particularly easy or difficult for some students (this is a cause of testing error); the sample of students who fail to answer an item correctly on their first attempt is obviously biased toward students for whom that item is especially difficult. The observation that students with sufficient skill to do the item (i.e., with $s-d>0$) trend well below initial (IRT-based) expectation on this second attempt may well indicate that many started with, and retain, some significant confusion.

Significantly, receiving useful feedback or using hint(s) and/or subtask(s) provided by the tutor increased the skill of the students who had initially answered incorrectly.[9] The effect size ranged from 0.34 to 0.62 standard deviation, with an average $0.44\pm0.10$ standard deviation. Their actual learning probably significantly exceeds this, since without useful tutoring this group should have (negative) item-wrong bias for the same reason as the sample just discussed. In addition, the data indicate that consulting hint(s) and/or subtask(s) seems to be more beneficial than receiving only feedback from the tutor ($p=0.12$ using the two-parameter fit, but $p=0.02$ if the fit constrains the discrimination). The IRT result that the discrimination is reduced can be stated more informatively as follows: while the second-attempters whose skill exceeds the difficulty of the item get back to the level initially expected of them, those with less skill (i.e., $s-d<0$) benefit most—their probability of answering correctly is very significantly increased over initial expectation by the feedback.

We were surprised that students who elected to consult the hints (at some penalty to their score) before submitting their first answer[10] showed the largest increase in their skill, effect size $1.87\pm0.16$ standard deviation (see Fig. 5). One possible explanation is that these students have enough metacognitive ability to recognize that they need help solely from inspecting the item (rather than having to make a mistake to learn this) and that enables them to use hints and subtasks until they are reasonably certain that they have the solution. This interpretation is buttressed by the fact that students whose *a priori* chance of answering correctly is less than 50% (i.e. $s-d<0$) show an increased probability of answering correctly on their second attempt by 0.3 or more, a factor of 2–5.

A majority of students who failed their first attempt and used any form of tutoring were able to solve items on their second attempt, as reflected in their skill increase by $0.43\pm0.10$ standard deviation. This may be quantified by noting that on first attempt for all items 60% were answered correctly. After tutoring, 59% of the second attempts will be correct, hence by now 84% of the students will have answered the item correctly. Those who have not continue using the tutoring until eventually ~95% of all students have answered correctly. In this work, student background information such as gender or student major was not considered. However, previous research shows that such contextual factors do not correlate with student performance on MASTERINGPHYSICS homework.[11] Thus, it would be reasonable to assume that the learning effect reported in this work is also not associated with such contextual factors.

This paper does not address the question of whether the skill increase is transferable to subsequent problems. However, research on knowledge transfer in this environment shows that students need 15.4% fewer hints and submit 11.4% fewer wrong answers when working a subsequent problem on the same topics as the previous one.[12,13] We estimate that this would be consistent with a learning effect of the increased student skill by about 0.5 standard deviation (but just on items concerning that detailed topic).

## IMPLICATIONS

The generalization of IRT developed here has several potential applications beyond studying the effectiveness of hints and subtasks. These include refining tutorial content based on its performance and also assessing students in a learning environment.

Data like those presented here inform the tutorial authors about the efficacy of the tutoring material along each learning trajectory through each individual item because the skill changes found, $\delta s_i$, have this specificity. With some study and experience, these should allow the content author to evaluate the effectiveness of the tutorial help along each path relative to tutoring in other items. For this purpose (aiding the content author) we note that our approach—assigning a skill increase to each trajectory through the tutoring—might better be replaced by an analysis in which each particular wrong answer response and each particular hint and subtask is regarded as a separate learning element. Then a multiple regression approach on a large enough sample could find the skill increase attributable to each hint and subtask, allowing even more specificity in evaluating the tutoring elements for each tutoring item.

The major assessment payoff of our generalized IRT is greater reliability, achieved by making inferences about a student's skill from each attempt at solution, rather than from only their first attempt (as conventional IRT does). Assume that a large number of students have calibrated a particular trajectory and thereby determined $\delta s_i$ for that trajectory. Then for subsequent students, their second attempt at this item after having passed along that trajectory may be regarded as a "first" attempt at a new item of known difficulty (i.e., $d_{\text{new}}=d_i-\delta s_i$) and discrimination. Then Eq. (2) may be used in the same manner as Eq. (1) was used to determine or to update the student's skill. Our use of "update" here implies that, in a learning environment, a student's skill is changing over time, and hence it is desirable to have a way to determine a "current skill" rather than just a final skill (as IRT normally provides after a test).

It also seems possible to use our approach to increase the reliability of computer-administered examinations. If they require free responses, the above techniques will work (the only trajectory being a second attempt without benefit of specific feedback or hints). For a multiple choice exam, the item is arguably less difficult after the student has been informed that the distractor that he selected is incorrect. Nevertheless, if each particular distractor is regarded as a distinct learning trajectory, then Eq. (2) may be applied to determine $\delta s_i$, allowing the skill of future students to be assessed along that trajectory. This is because Eq. (2) (like all of IRT) depends only on the difference between skill and difficulty; whether $\delta s_i$ is regarded as an increase in skill or a decrease in difficulty is immaterial.

## CONCLUSIONS

We have presented a mathematical model and procedure to generalize IRT to measure students' skill change due to learning that occurs between successive attempts to answer a single item. We showed that this model, using a shifted logistic function, accounted within error for response after feedback. We showed that the skill change depended on the students' path through the tutoring available in an online tutorial environment, with effect sizes from −1.02 to +1.87. We argued that the surprisingly large −1.02 standard deviation skill change of students who attempted to solve the item without using any hint or feedback from the tutor did not represent any sort of "unlearning," but probably reflected item-wrong bias associated with being selected because their first response was incorrect. This certainly indicates that simply telling a student that they are wrong does not help them perform very much better on their second attempt.

Our findings make a strong positive statement about the overall effectiveness of the tutoring of MASTERINGPHYSICS in helping students who answer incorrectly on their first attempt. Students who either spontaneously or by request receive such help perform 0.84 standard deviation better on their second attempt. (More help is available for those who get it wrong the second time.) We have shown only dramatic correlations between the skill change and the learning route selected by the students (or selected by the computer for spontaneous feedback). We do not see that particular students strongly favor the most effective "tutoring before first attempt" path so it appears that the learning correlates with the path and not the student. However, the path selected may indicate the student's readiness to learn (e.g., we suggested that selecting tutoring first showed metacognition in action). Still, while some paths may be selected by those "ready to learn," the available paths did indeed help the students reach the correct solution.

Another significant finding is that less skillful students benefited far more from the tutoring than more skillful ones. Figure 3 shows that tutoring increased $P^{\text{correct}}$ significantly only for students attempting items whose difficulty exceeds their skill. Figure 3 shows improvements of $\sim$1.9 to $\sim$4.0 error bars ($p$ values from $\sim$0.02 to $1.3\times10^{-4}$) for all points in this region. Students who decide to use hints and subtasks first improve even more.

We argue that the strong improvement from tutoring measured here, and the fact that it helps weaker students more, shows that hints and feedback from MASTERINGPHYSICS effectively play the role of "as needed" scaffolding. This tutoring is available when needed, enables 59% of those students who answer incorrectly on their first attempt to find their

way to the solution on their very next attempt, and helps about 2/3 of those answering incorrectly on their second attempt to the solution before exhausting their remaining four attempts.

**ACKNOWLEDGMENTS**

We are grateful to MIT and to NSF Grant No. DUE-0231268 for supporting this work.

[1] A. Collins, J. S. Brown, and S. E. Newman, Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics, in *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser*, edited by L. B. Resnick (Lawrence Erlbaum, Hillsdale, NJ, 1990), p. 453.

[2] http://www.masteringphysics.com/ was made by Effective Educational Technologies, a company started by the family of one of the authors (D.E.P.). It has been purchased by Pearson Education.

[3] N. Chudowsky, R. Glaser, and J. W. Pellegrino, *Knowing What Students Know: The Science and Design of Educational Assessment* (National Academies Press, Washington, DC, 2001).

[4] F. B. Baker and S.-H. Kim, *Item Response Theory: Parameter Estimation Techniques* (Marcel Dekker, New York, 2004).

[5] R. K. Hambleton and H. Swaminathan, *Item Response Theory: Principles and Applications* (Kluwer Nijhoff, Boston, 1985).

[6] F. M. Lord, *Applications of Item Response Theory to Practical Testing Problems* (Erlbaum, Mahwah, NJ, 1980).

[7] Computer code BILOG-MG (Version 3.0) (Assessment Systems Corporation; St. Paul, MN, 2003).

[8] S. Kim, F. B. Baker, and M. J. Subkoviak, The $1/kn$ rules in the minimum logit chi-square estimation procedure when small samples are used, Br. J. Math. Stat. Psychol. **42**, 113 (1989).

[9] We also examined if subtasks were more beneficial than hints, considering the fact that subtasks provide much more specific information than hints, but found no statistical evidence.

[10] This "hints first" route was followed in 11.0% of all cases, and the median, mode, and maximum use of this route by students were 8.6%, 0.01%, and 71%, respectively.

[11] E.-S. Morote and D. E. Pritchard, Technology closes the gap between students' individual skills and background difference, in *Proceedings of the Society for Information Technology and Teacher Education: International Conference, Atlanta, GA* edited by C. Crowford, D. A. Willis, R. Carlsen, I. Gibson, K. McFerrin, J. Price, and R. Weber (AACE, Chesapeake, VA, 2004), p. 826.

[12] R. Warnakulasooriya and D. E. Pritchard, Time to completion reveals problem-solving transfer, in *Proceedings of the Physics Education Research Conference, Sacramento, CA*, edited by J. Max, P. Heron, and S. Franklin (AIP, Melville, NY, 2004), p. 205.

[13] R. Warnakulasooriya, D. J. Palazzo, and D. E. Pritchard, Evidence of problem-solving transfer in web-based Socratic tutor, in *Proceedings of the Physics Education Research Conference, Salt Lake City, UT*, edited by P. Heron, L. McCullough, and J. Max (AIP, Melville, NY, 2005), p. 41.