Using Inverse Planning for Personalized Feedback

Anna N. Rafferty
Department of Computer Science
Carleton College,
Northfield, MN 55057 USA
arafferty@carleton.edu

Rachel A. Jansen
Department of Psychology
University of California,
Berkeley, CA 94720 USA
racheljansen@berkeley.edu

Thomas L. Griffiths
Department of Psychology
University of California,
Berkeley, CA 94720 USA
tom_griffiths@berkeley.edu

ABSTRACT

An increasing number of automated models can make inferences about learners' understanding based on their problem solving choices in interactive educational technologies. One potential use of these models is to personalize feedback interventions. We investigate using the output of an inverse planning model to choose feedback activities for learners. The inverse planning model uses the patterns of how a learner solves algebraic equations to estimate her proficiency on several discrete skills. The personalized feedback then focuses on the skill which is least proficient and includes a combination of existing educational content and scaffolded practice. We experimentally tested the effectiveness of personalizing the feedback based on the algorithm's estimate compared to simply providing a random feedback activity. The results show that completing the feedback was associated with performance improvements from pre- to post-test, but that personalized feedback was not associated with reliably more improvement. However, participants who received feedback about a skill that was far from mastery did show reliably more improvement than those who received feedback about an already-mastered skill. This suggests that there is potential in using the inverse planning algorithm to provide more effective learning experiences.

1. INTRODUCTION

Cognitive models of people's learning are often useful for better understanding behavior and can highlight what a particular learner knows and where she may be struggling. There are also an increasing number of educational resources available for learning specific topics, such as online videos, which might be effective for remediating a learner's struggles. However, there can be challenges when trying to close the loop between estimating a model of what someone knows and creating interventions based on that model to address misunderstandings or gaps in knowledge. The model is not a perfect assessment, and many interventions may be effective for a particular learner, making it difficult to determine if personalizing the intervention is valuable. While there are a number of models that have been used to change the behavior of an educational technology, such as providing problems until mastery [2], there has been less of a focus on using models based on behaviors in more open-ended learning environments to guide feedback and remedial interventions in these settings.

We address the problem of closing the loop between a model-based assessment of a learner's algebra skills and the experience the learner has in a web-based algebra activity. The model was an inverse planning model for algebra, which provides an assessment of specific algebra skills based on the pattern of how someone solves equations. While the model provides a profile of what a person may misunderstand, suggesting that it could be used to guide feedback interventions, its estimates also have some error, meaning that it will not perfectly identify misunderstandings for every person. Additionally, the model's assessment is based on a collection of problem solutions, meaning the feedback must be targeted at an overall skill or misunderstanding rather than performance on a specific problem. This differs from many contexts where feedback is provided in interactive educational technologies, but has the potential to facilitate longer interventions about specific concepts or skills. This type of feedback could connect a learner with existing resources about particular concepts, since rather than assisting with a single question, the feedback is remediating a more abstract area of struggle. Thus, we explore how the model's assessment of understanding can be used to provide feedback to learners that targets their misunderstandings.

We investigate this question by designing feedback interventions for specific skills and experimentally testing how people's performance changes from pre- to post-test based on the intervention that they are given. The feedback interventions combined relevant content from existing sources and scaffolded opportunities for practicing a particular algebra skill. In an experiment, we compared performance for people who completed a feedback intervention based on the algorithm's estimate of their skills versus those who completed an intervention that was chosen randomly. We found that both groups showed significant performance improvements from pre- to post-test, but the two groups did not differ in their amount of improvement. However, completing feedback about a skill that one was less proficient in was reliably associated with more improvement than completing feedback about a skill that was near mastery. These results suggest that the algorithm's assessment may be used to di-

rectly improve the educational technology, although there are a number of subtleties in how to do this effectively.

2. BACKGROUND

There has been a great deal of previous work related to assessing student understanding and providing interactive feedback to improve understanding. In our work, we are most interested in techniques where a student's actions or choices are used as part of the assessment of understanding, such as in open-ended learning environments (OELEs). OELEs are often used in science education, as they can provide opportunities for students to generate and test their own hypotheses [6]. Educational data mining has been used to better understand what behaviors are associated with learning in some of these environments, such as Betty's Brain [4], and these environments may provide feedback to students about their progress (e.g., [12]). Data mining is also used in these environments for assessments of skills, especially those like experimentation that are more difficult to measure in other environments [3]. However, it is rarer for the data mining to be used directly to inform feedback to students, and the feedback that is provided is frequently in the form of a short hint or suggestion about what to do. In mathematics education, there exist several systems, such as the Cognitive Tutor [2], that maintain a model of student learning and use this to adapt instruction, such as providing more problems on an unmastered skill; typically these systems assess student knowledge based on final answers rather than on what actions are taken to generate a solution. In both the science and mathematics systems the type of adaptive feedback differs from our focus on providing a somewhat longer session of feedback focused on re-teaching a particular skill.

While formative feedback to learners is an effective way to improve understanding and help create a more integrated base of knowledge [13], the problem of determining what type of feedback will be most effective is an area of active research. Much of the previous work on feedback in mathematics tutors has focused on progressively more informative hints (e.g., [5]). More holistic information based on assessments of skills may be provided to students, such as when making a learner model "open" to the learner [1], but this is not necessarily paired with feedback or interventions to remediate understanding. Research about teachers' responses to student work in educational technologies has found that teachers may customize their instruction in a variety of ways to adjust to student misunderstandings [8]. Based on this work, we were interested in how more holistic feedback that focuses on a particular skill that a student is struggling with, rather than a specific problem, might affect learning.

3. INVERSE PLANNING

In order to get a holistic assessment of a learner's algebra skills based on observing their behavior, we used a Bayesian inverse planning approach [11]. Bayesian inverse planning takes as input a set of step-by-step actions from a learner, and outputs a posterior distribution over possible levels of proficiency for various skills. This approach allows us to interpret people's patterns of behaviors while they solve algebraic equations in a relatively freeform interface. In this interface, shown in Figure 1, learners have the ability to enter step-by-step solutions to equations, with no constraints on whether individual steps are correct before entering a new

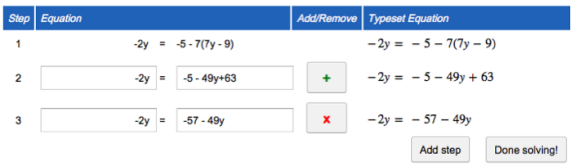

Figure 1: A screenshot of the step-by-step interface for solving algebraic equations. The user may solve the problem using any steps she chooses and record them in the interface.

step. The Bayesian inverse planning algorithm uses both the mathematical correctness of each step and the way it moves the learner towards the solution to diagnose proficiency; the model is substantially similar to that described in [10]. We provide a brief overview of the algorithm and its underlying model of problem solving.

Bayesian inverse planning is based on a generative model: it models how likely a person would be to choose each possible solution step if she had a particular understanding of algebra, and then uses this model to infer what understanding is most likely to have resulted in the observed solutions. To create this generative model, we need to specify how choices about solution steps are made as well as specifying the representation of possible understandings. Inverse planning treats algebraic equation solving as a Markov decision process (MDP), in which people choose actions to bring them closer to the goal of solving an equation with as few steps as possible. With each action, the person moves from one (partially solved) equation to another. In an MDP, the value $Q_h(s, a)$ of taking an action a given that the current equation is s can be approximated using dynamic programming. This long-term value is dependent on the person's understanding of algebra, denoted as h, since that understanding may change what actions she believes are possible or what next equation she generates from the current equation. We model people as following a noisy optimal policy when choosing actions: $p(a|s) \propto \exp(\beta \cdot Q_h(s, a))$, where β controls the level of noise. Intuitively, this policy assumes people tend to choose actions that they think will help them solve the problem efficiently but they do not do so deterministically. The parameter β is estimated for each individual, as described below.

In this model, understanding is represented by the level of proficiency for several skills. For each skill, the proficiency indicates whether the person generally applies the skill correctly or if she makes a particular type of error. The different levels of proficiencies form a *hypothesis space* of possible algebra understandings. The hypothesis space was based on past education and psychology research and consists of parameters for six skills (see [10] for details): moving terms, dividing by the coefficient of a term, applying the distributive property, combining terms, arithmetic, and planning. The first four parameters relate to specific rules of manipulating algebraic equations, while the latter two apply more broadly.

Each of the four algebra-specific parameters indicates whether the person is prone to a particular type of error or "mal-

rule" [9]. For moving terms, the mal-rule is failing to flip the sign of a term when moving it from one side of the equation to another; the inferred parameter is the probability of not following this mal-rule when moving a term. For dividing by the coefficient of a term, the mal-rule is multiplying rather than dividing (i.e., not using the reciprocal), and for applying the distributive property, the mal-rule is only distributing the coefficient to the first term rather than all terms. Both of these parameters, like moving terms, are probabilities. For combining terms, the mal-rule is combining unlike terms, such as a variable and a constant. This parameter is binary: the person either considers combining unlike terms when choosing actions or she does not.

The final two parameters for the hypothesis space are the arithmetic parameter and the planning parameter. The arithmetic parameter is the probability that a person accurately computes a calculation. The planning parameter is the parameter β in the noisily optimal policy: higher values for this parameter indicate very high probability of choosing the most efficient action for moving towards a solution, while values close to zero indicate very different choices from those expected by the model, such as choosing an action that does not make progress towards the solution or giving up prior to reaching a solution. This parameter is the only parameter not targeted for feedback, as a mixture of cognitive and motivational feedback might be most effective for improving planning and lessening the rate of non-answers.

The parameters above form a six-dimensional, continuous hypothesis space $\mathcal{H}$, where each point in the space represents one possible set of skill proficiencies h. Given this hypothesis space, the posterior distribution after observing the person's problem solutions D is calculated using Bayes' rule: for each $h \in \mathcal{H}$, $p(h|d) \propto p(h) \prod_{d \in D} p(d|h)$, where $p(h)$ is the prior distribution over the hypothesis space and $p(d|h)$ is the likelihood that the person would produce the observed step-by-step solution if she had the skill levels indicated in h. The prior favors higher levels of proficiency; intuitively, this means that the algorithm favors the part of the hypothesis space indicating normative algebra understanding unless it observes evidence in the solutions that non-normative steps are being taken. Because the hypothesis space is continuous, the posterior distribution cannot be calculated exactly. Instead, Markov chain Monte Carlo (MCMC) methods are used to compute an approximate posterior distribution. As shown in Figure 2, the resulting posterior distribution indicates both the most likely proficiency for each skill as well as the algorithm's confidence. In the figure, both the parameter for moving terms and the distributive property are close to one, but the estimate for moving terms is more certain; there is also a lower estimated proficiency for arithmetic than for the other skills. In order to use the posterior distribution for feedback, we calculate the mean value of the posterior on each skill dimension (shown as green lines in Figure 2).

4. FEEDBACK DESIGN

Given the output of the inverse planning algorithm, our goal was to "close the loop" by providing learners with a feedback activity that could help to remediate their understanding of a particular skill. In an attempt to minimize differences in feedback effectiveness due to quality rather than topic, all of the feedback interventions followed the same pattern. First,

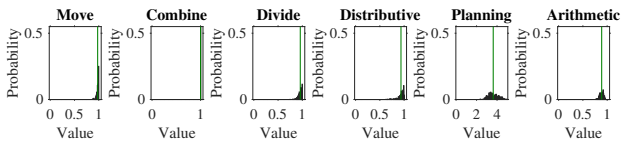

Figure 2: The inverse planning algorithm's assessment for a learner from the experiment. Each plot shows the posterior distribution for one skill. Larger values are closer to mastery.

an overview screen showed the learner two skills: the skill closest to mastery and the skill she would receive feedback about. In both cases, she was shown her proficiency level as a colored bar and a short description of the skill was provided. The bottom of the page told her that she would be learning more about the second skill that was shown; we refer to this skill as the *feedback skill*. On the next page, learners were shown a 2–5 minute embedded video about the feedback skill. Since there already exist a large number of freely availably educational videos, we aimed to connect learners to a relevant resource rather than create new tutorial content. All videos were sourced from Khan Academy[1], and were chosen because they targeted one of the five skills.

After the video, several stages of scaffolded practice were provided. For the four skills related to algebraic rules, the scaffolded practice began with four problems to highlight the core skill being practiced. For example, only the feedback focused on correctly applying the distributive property included practice on the distributive property. For these problems, the learner's steps were checked for correctness with each new step. If a mathematical error was detected, the step was highlighted and she was asked to fix it before continuing. After each problem, the learner was told the correct answer. Following these problems, eight problems were provided that still focused on the feedback skill, but checking of correctness was only provided after the learned submitted her answer. At that point, steps with errors were highlighted and the learner was given the opportunity to review them before continuing. These problems thus targeted the feedback skill, but included slightly less immediate assistance than the first set of problems. For the feedback targeting arithmetic, all twelve practice problems were arithmetic computations to complete rather than algebraic equations. Finally, all feedback finished with twelve algebra problems that were not specialized based on the feedback skill, with the intention for people to practice in context what they had learned from the skill-specific problems. The interface for these problems was the same as when doing general problem solving on the website: people had the opportunity to enter individual problem steps, and they were told whether they were correct before moving to the next problem.

5. EXPERIMENT

When we designed the feedback, our goal was to personalize what feedback someone was given based on the algorithm's assessment of their skills by assigning the person to complete feedback on their least proficient skill. While it is intuitively plausible that personalized feedback based on

[1]http://www.khanacademy.org/

this assessment might be more helpful than non-personalized feedback, there are several reasons to be skeptical. First, the algorithm's diagnosis is an approximation: there is error both in the MCMC estimate, and in the model itself. In general, the algorithm can interpret most problem solutions [10], but some people's behavior may be poorly fit by the model, resulting in poor accuracy for an individual. Additionally, the algorithm does not account for learning within the period that the skills are being assessed and depending on the person's behavior, there may be some skills about which we have very limited information. For example, a person might solve only a few problems using the distributive property, giving a relatively large confidence interval for possible skill proficiencies. A second concern about personalizing feedback is that learners who are struggling may be struggling in many skills. In that case, it may be that the personalization is unnecessary: most students who benefit from one feedback activity would also benefit from any of the other feedback activities. Thus, we ran an experiment to test whether the feedback activities were associated with learning and to examine whether personalized feedback produced larger learning gains than feedback that was not personalized based on the algorithm's assessment.

5.1 Methods

Participants. 200 participants in the USA were recruited from Amazon's Mechanical Turk (AMT) and compensated $4 for session 1, $6 for session 2, and $8 for session 3. Participants had taken an algebra course and had not completed college math classes beyond algebra.

Stimuli. Participants completed a multiple-choice assessment, solved algebra problems on a website, and responded to several surveys. The twelve question multiple-choice assessment was based on College Board ACCUPLACER® tests used for math placement in many postsecondary institutions[7]. The questions were substantially similar to the Elementary Algebra questions used in [11], but the numbers were changed to create two versions of the assessment.

All problem solving on the website used a similar interface to that shown in Figure 1. In sessions 1 and 3, learners were told whether or not they were correct immediately after submitting a problem. During the feedback in session 2, the interface behaved as described in the previous section.

In the surveys, participants indicated their demographics as well as prior math class experience. They also completed 18 questions focused on the usability of the website and the perceived helpfulness of the feedback.

Procedure. Participants completed three sessions, separated by at least one day. In the first session, all participants solved 24 equations on the algebra website, followed by the multiple-choice questions about elementary algebra topics. The website included a short tutorial about how to use the interface, and the 24 problems were generated based on templates. For example, one template was a constant plus a variable equal to a constant. The constants and coefficients for variables were generated randomly, but all participants shared the same templates. After a participant completed all problems on the website, the diagnosis for that participant was computed automatically by the inverse planning algorithm, and results were stored in the database for the participant's next session. Participants were randomly assigned to receive version one of the multiple-choice questions or version two; these versions were identical except for changes to the exact numbers used in the problems.

In the second session, participants completed one of the feedback activities. They were randomly assigned to either *targeted* or *random* feedback. Those receiving targeted feedback completed the feedback activity for the skill which the algorithm estimated they had least proficiency; those receiving random feedback completed one of the five feedback activities selected uniformly at random.

In the third session, participants again solved 24 equations on the algebra website, followed by the multiple-choice questions about elementary algebra topics. Just as in the first session, participants all completed problems on the website that used the same templates. For the multiple-choice questions, each participant completed the version of the questions that they did not complete in the first session. Finally, participants ended the third session by completing the demographics and usability surveys.

5.2 Results

82% of participants completed all three parts of the experiment in a single session. Several participants were removed due to technical problems, such as needing to restart the computer during a session and thus losing their place in the activity. The results that follow include only the 164 participants who completed all parts of the experiment.

Responses to our demographics questions suggest that participants came in with varying levels of mathematics background and that for most, significant time had passed since they had last studied algebra in school. 98% of participants reported what previous math classes they had taken, in college or in high school. 62% of those who responded had taken no math classes beyond geometry (typically at a high school level); the remaining participants had taken trigonometry, pre-calculus, or calculus at a high school level. A number of participants who reported taking one of these higher-level courses in high school also reported taking a college algebra class. Thus, we would expect all participants to have prior experience with solving equations, but to be likely to have some gaps in their knowledge.

We first examined changes in participants' performance between the first session, before getting feedback, and the final session, after getting feedback. Results from the first session confirmed that participants were on average far from ceiling on the task: they correctly answered an average of 7.2 multiple-choice questions out of a total of 12, and correctly answered an average of 12.4 out of the 24 algebra problems on the website. There was a small increase in the number of multiple-choice questions answered correctly in the final session. Using a repeated-measures ANOVA with factors for time of test, condition, and a random factor for participant, we found that this main effect was reliable ($F(162, 1) = 15.7$, $p < .001$), but there was no interaction between condition (targeted versus random feedback) and time of test. Given that many of the questions focused on skills that were not directly targeted by our intervention, in-

cluding some quadratic equations and linear inequalities, it is not surprising that we see only a small improvement from the first to the final session. The increase in performance was somewhat larger for the algebra equations solved on the website: participants correctly answered 23% more problems correctly, for a mean of 16.6 problems correct in the final session. We again used a repeated-measures ANOVA with factors for time of test, condition, and a random factor for participant to analyze the reliability of this finding, and found that there was a main effect for time of test ($F(162,1) = 89.9$, $p < .001$), but no interaction between time of test and condition.

To better understand why there was no interaction between condition and the amount of improvement, we examined the estimated proficiency level of the skills for which feedback was given. On average, the targeted condition selected skills that had lower levels of proficiency (average proficiency level of 0.56 versus 0.88; $t(162) = 7.03$, $p < .001$), indicating that in many cases, there were large differences between the least mastered skill and a random skill. However, there were a number of participants in the random condition who received feedback about a skill with which they were struggling as well as participants in the targeted condition who did not have any skills that were far from mastered. To test whether participants who received feedback that was more appropriate for them improved more than participants who received feedback that was less appropriate for them, we divided all participants into two categories: those who received feedback about a skill that was estimated to be less than a proficiency level of 0.85 (an *unmastered* skill) and those who received feedback about a skill that was at a proficiency level greater than or equal to 0.85 (a *mastered* skill). This criterion categorizes 46% of participants as receiving feedback about an unmastered skill. As shown in Figure 3, participants who received feedback about an unmastered skill improved more than those who received feedback about a mastered skill. A repeated-measures ANOVA with factors for whether the feedback skill was already mastered, time of test, and a random factor for the participant showed that there was a main effect of time of test as well as an interaction between time of test and whether the feedback skill was already mastered ($F(162,1) = 9.42$, $p < .01$). To ensure that this result was not simply due to the cutoff level we chose for mastery, we also examined a categorization based on mastery level 0.9, and found the same trends ($F(162,1) = 46$, $p < .05$). While these results must be interpreted with some caution, as participants were not randomly assigned to the two categories, they suggest that receiving feedback that the algorithm indicates is more appropriate can result in greater improvements in performance.

Based on the fact that proficiency level influenced the effectiveness of the feedback, we examined the distribution of proficiencies for individual participants. We were interested in whether participants tended to have all skills at a similar level or whether they usually had some skills that were mastered and some that were unmastered. As shown in Figure 4, 35% of participants were at mastery for all skills, where mastery is defined as proficiency of at least 0.85, and 14% of participants were not at mastery for any skills. The remaining 51% of participants who had some mastered skills and some unmastered skills are arguably those that might most benefit from targeted rather than random feedback. A repeated-measures ANOVA with factors for time of test, condition, and a random factor for participant shows that there is a significant interaction between time of test and condition when restricting the data to these participants: as shown in Figure 5, those who completed targeted feedback improved almost twice as much those who completed random feedback (average improvement 5.3 versus 3.0; $F(82,1) = 5.64$, $p < .05$).[2] This suggests that inverse planning can provide a benefit for these participants: it allows us to determine what skill(s) will be appropriate targets for feedback.

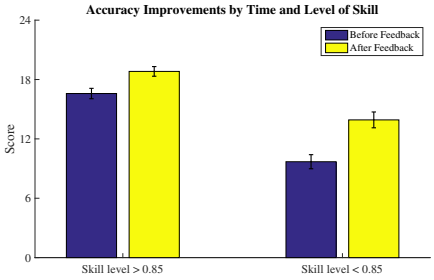

Figure 3: Improvement from first to last session in accuracy on website problems, categorizing participants based on prior level of proficiency in feedback skill. Participants who received feedback about an unmastered skill improved more from the first to the final session than those who received feedback about a mastered skill.

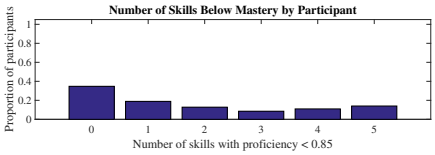

Figure 4: Count of the number of unmastered skills by participant.

6. DISCUSSION

Our goal in the feedback design and the experiment was to evaluate the benefit of connecting the holistic assessment and the feedback activities. While many of the feedback problems provided practice on multiple skills, since multiple skills are required to solve the algebraic equations, there was specialization in our feedback based on the algorithm's assessment. Our results show that overall, participants' performance improved after completing the feedback activities. The effects of personalization on the size of this effect were mixed: across all participants, feedback targeted at someone's weakest skill was not associated with reliably more improvement than feedback about a random skill, but restricted to those who had some mastered and some unmastered skills, we observed more improvement for those receiving the targeted feedback compared to those receiving the random feedback. This suggests that there is promise in using the inverse planning algorithm's assessment to connect

[2]With mastery level set at 0.9, this effect is marginally significant (average improvement 4.2 versus 2.7; $F(103,1) = 3.33$, $p = .07$).

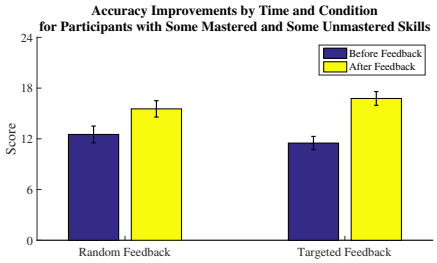

Figure 5: Improvement from first to last session in accuracy on website problems, restricted to participants with some unmastered and some mastered skills. Participants show reliable improvement, and participants who received targeted feedback tended to improve more than those who received random feedback.

learners to relevant resources and personalize feedback activities, although further investigation is needed to determine ways to make this personalization even more effective.

There are several limitations of this work. First, our population of AMT workers may not be typical of algebra learners. These people were paid to participate in the study, and may differ in motivation and background from those who would use the website by choice. However, their varied backgrounds may be typical of adult learners who are trying to surmount barriers such as algebra at the community college level, a group we are particularly interested in reaching. Second, this experiment does not separate whether the content of a feedback intervention is helpful from whether the targeting of that feedback is accurate. We intend to further evaluate these two components to better understand what the maximum benefit of this type of feedback would be if targeting was perfectly accurate, but any evaluation of the overall effectiveness of the knowledge diagnosis-feedback loop must acknowledge that inaccuracies in the diagnosis may lead to the personalization being less effective.

In future work, there are a number of ways we will explore how to design more effective personalized feedback and investigate variations in how to use the algorithm for personalization. Our intervention was relatively short, with most participants taking about an hour for the session in which feedback was provided. One might expect the effects of personalization to be cumulative, with targeted feedback being most helpful when learning over a longer period; in that case, the targeting could be used to remediate the same skill multiple times if struggles were still evident or to recognize that say, one session of feedback had resulted in several skills reaching mastery and skipping the already mastered skills. Such longer interventions are likely to have larger effects, and may highlight whether targeted feedback is overall more effective or whether there is a subset of participants for which targeting makes a difference. Another area to explore is providing the profile generated by the inverse planning algorithm to the learner and using this in conjunction with targeted feedback, random feedback, or feedback chosen by the learner. The current system provides learners with the algorithm's assessment of several of their skills, but it does not allow them to make choices about what feedback they receive. Choice might be useful for those not well-modeled by the algorithm or in cases where several non-mastered skills have been identified; however, it is also possible that struggling learners are unable to understand the possible types of feedback in order to make a good choice. Finally, there are several ways we might adjust how the algorithm's output is linked to feedback. The diagnosis includes information about the algorithm's certainty. This might be used to focus on skills that we are confident are unmastered. Additionally, the algorithm outputs a diagnosis of planning efficiency, but this was not used for feedback. Low levels of this parameter can be indicative of someone who frequently gives up or who is not well fit by the model. In either situation, it may not be appropriate to simply give feedback about the least proficient skill. Overall, the results in this paper serve as first steps for a larger investigation into how to effectively close the loop between holistic assessments of misunderstandings and guiding personalized feedback interventions for learners.

Acknowledgements. This work was funded by NSF grant number DRL-1420732 to Thomas L. Griffiths. Thanks go to Jonathan Brodie and Sam Vinitsky for programming parts of the feedback.

7. REFERENCES

[1] S. Bull and J. Kay. Student models that invite the learner in: The SMILI open learner modelling framework. *International Journal of Artificial Intelligence in Education*, 17(2):89–120, 2007.

[2] A. T. Corbett, K. R. Koedinger, and W. Hadley. *Cognitive tutors: From the research classroom to all classrooms*, pages 235–263. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 2001.

[3] J. D. Gobert, M. Sao Pedro, J. Raziuddin, and R. S. Baker. From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, 22(4):521–563, 2013.

[4] J. S. Kinnebrew, K. M. Loretz, and G. Biswas. A contextualized, differential sequence mining method to derive students' learning behavior patterns. *JEDM-Journal of Educational Data Mining*, 5(1):190–219, 2013.

[5] K. R. Koedinger and V. Aleven. Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, 19(3):239–264, 2007.

[6] S. M. Land. Cognitive requirements for learning with open-ended learning environments. *Educational Technology Research and Development*, 48(3):61–78, 2000.

[7] K. D. Mattern and S. Packman. Predictive validity of accuplacer scores for course placement: A meta-analysis. Technical report, College Board, December 2000.

[8] C. F. Matuk, M. C. Linn, and B.-S. Eylon. Technology to support teachers using evidence from student work to customize technology-enhanced inquiry units. *Instructional Science*, 43(2):229–257, 2015.

[9] S. Payne and H. Squibb. Algebra mal-rules and cognitive accounts of error. *Cognitive Science*, 14(3):445–481, 1990.

[10] A. N. Rafferty and T. L. Griffiths. Interpreting freeform equation solving. In *Artificial Intelligence in Education*, pages 387–397. Springer International Publishing, 2015.

[11] A. N. Rafferty, M. M. LaMar, and T. L. Griffiths. Inferring learners' knowledge from their actions. *Cognitive Science*, 39(3):584–618, 2015.

[12] J. R. Segedy, J. S. Kinnebrew, and G. Biswas. The effect of contextualized conversational feedback in a complex open-ended learning environment. *Educational Technology Research and Development*, 61(1):71–89, 2013.

[13] V. Shute. Focus on formative feedback. *Review of Educational Research*, 78(1):153–189, 2008.