Towards Understanding How to Leverage Sense-making, Induction and Refinement, and Fluency to Improve Robust Learning

Shayan Doroudi[1], Kenneth Holstein[2], Vincent Aleven[2], Emma Brunskill[1]
[1]Computer Science Department, [2]Human-Computer Interaction Institute
Carnegie Mellon University
{shayand, aleven, ebrun}@cs.cmu.edu, kenneth.holstein@gmail.com

ABSTRACT

The field of EDM has focused more on modeling student knowledge than on investigating what sequences of different activity types achieve good learning outcomes. In this paper we consider three activity types, targeting sense-making, induction and refinement, and fluency building. We investigate what mix of the three types might be most effective in supporting robust student learning. To do so, we collected data from students in grades 4 and 5 who completed sequences of activities in largely random order. Students significantly improved from pretest to posttest, suggesting that incorporating all three types can support learning gains. Using hierarchical linear modeling, we found that students who get relatively more fluency problems achieve higher posttest scores. This finding suggests that fluency-building activities are most effective in helping students learn, although our data do not allow us to conclude that fluency alone is sufficient. This work represents a step towards better understanding what combination of different learning mechanisms may best support robust learning.

1. INTRODUCTION

Intelligent tutoring systems (ITSs) have been very effective at enhancing student learning [12, 6]. They typically provide step-level support for complex problem solving such as correctness feedback, next-step hints, and error-specific feedback. ITSs also provide individualized problem selection [11, 3]. It is interesting to consider ITS effectiveness from the perspective of the Knowledge-Learning-Instruction (KLI) framework [5]. KLI posits that three mechanisms of learning—sense-making (SM), induction and refinement (IR), and fluency-building processes—may all be important for robust learning (persistent learning that supports future learning) in any complex domain. However, existing ITSs typically focus only on the IR mechanism through the provision of scaffolded, tutored problem solving. It is possible that providing support for all three learning mechanisms will lead to more robust learning. Supporting the three learning mechanisms would however require a wider range of activity types than typical ITSs offer, to add or enhance support for SM and fluency. Further, it would require that we answer key questions of how and when to provide the different activity types to different learners in an individualized manner, which may itself depend on the student's learning process so far.

In this paper we take a preliminary step towards answering these questions. Fractions Tutor [8] is a web-based intelligent tutoring system for fourth and fifth grade fractions learning. We significantly extended the Fractions Tutor to support all three learning mechanisms. We then collected data from over 600 students with constrained random problem sequences. This allowed us to do a preliminary analysis to understand the contributions of activities targeting the three different learning mechanisms. We did this by fitting a hierarchical linear model (HLM) to our data to see how posttest scores are influenced by the proportion of each activity type in problem sequences as well as looking at the correlation of each activity type with posttest scores. A challenge in drawing conclusions from our data is that the mix of activity types each student was presented with was correlated with the number of problems each student did, but despite this challenge, we show that fluency-building activities are more effective for robust learning.

There has been related work on how to combine two different types of activities, such as worked examples and problem-solving practice [10]. More recent work on MOOCs has analyzed the effectiveness of different activity types chosen by the student (instead of the tutor) [4, 2]. More relevant to the current work is prior work on SM and fluency processes in the Fractions Tutor [8, 9]. While that work also uses hierarchical linear modeling [9], their model includes predictors corresponding to experimental conditions, whereas we have random trajectories with no experimental conditions. Using random sequences gives us the potential to compare a wider variety of relative compositions and sequences of activity types than a standard experimental study.

Finally, prior EDM work has looked at the related problem of how to measure the relative efficacy of different activities [1, 7]. While these works deal with a very similar problem to ours, they differ in at least two main respects from the present work. First, their models consider the efficacy of different activities in performance while being tutored, whereas

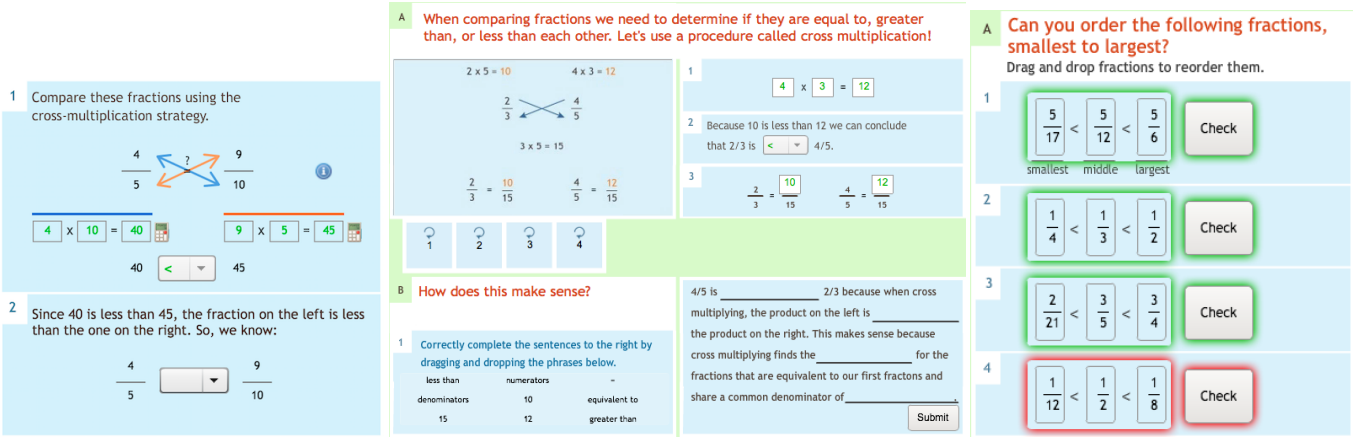

Figure 1: Sample IR (left), SM (center), and fluency (right) activities.

we are interested in robust learning (i.e. performance on a posttest). Second, they only consider which individual activity is best rather than what mix of activities is best. Our modeling approach could in theory suggest optimal mixes of activity types, although we find that in this case the best fitting model reduces to one that can only suggest the relative efficacy of each activity type. It would be worthwhile to compare our findings with the results we can obtain from these models as next steps of our work.

2. METHODS

2.1 Fractions Tutor

For this work, our Fractions Tutor covered topics emphasized in the Common Core[1]: making and naming fractions, fraction equivalence and comparison, and fraction addition. For each topic, we designed three activity types designed to promote each of the KLI learning mechanisms. KLI does not provide strict design guidelines and so we now describe how our designed activities targeting each learning mechanism are in line with KLI's definitions.

Under KLI, IR processes are non-verbal learning processes that improve the accuracy of knowledge [5]. Activities to promote IR processes emphasized procedural learning and practice via fine-grained task decomposition and step-level guidance and feedback, as is typical of ITSs [11]. An IR activity for a procedure for the comparison of two fractions is shown in Figure 1, on the left.

In KLI, SM processes are "explicit, verbally mediated learning in which students attempt to understand or reason" [5]. Our SM activities included instructional videos designed to promote conceptual understanding of targeted fractions topics. The videos were divided into small segments and interspersed with brief supporting problem-solving exercises. Each SM activity concluded with a drag-and-drop fill-in-the-blank question designed to help students self-explain the underlying concepts. An example SM activity for the cross-multiplication procedure is shown in Figure 1 (center). Unlike the IR activities that teach the application of this procedure, the SM activities were designed to help students understand why a certain procedure (e.g., cross-multiplication to compare and order fractions) is effective.

Finally, under KLI, fluency-building processes are non-verbal processes that strengthen memory and enable students to apply their procedural knowledge faster and more fluently [5]. Thus the fluency activities were designed to promote the development of rapid reasoning about fractions and fluent performance on minimally-decomposed problem-solving exercises. Whereas students received support from the tutor via step-level hints in IR activities and video-replays in SM activities, neither were available in fluency activities. See a sample fluency activity in Figure 1, on the right.

2.2 Activity Selection

Since we wish to be able to understand a broader range of activity orderings and mixes rather than a small fixed set, we presented activities to students in a semi-randomized order. A semi-randomized order was chosen as a compromise between two potentially competing objectives. The first is to enhance student learning broadly and for the students that participated in this initial data collection. This objective would push us towards selecting an activity order that draws upon existing research on effective sequencing and satisfies commonly assumed topic orderings. Our second objective is to be able to find effective (potentially adaptive) orderings that may fall outside of the reach of standard procedures. To balance these two competing objectives, we chose to provide students with activity sequences that initially satisfy a prerequisite structure over activity types and topics (designed by the authors). Students could be presented with any activity whose prerequisites had already been presented. This ensured some semantic ordering, e.g. students would not be presented with addition problems before being introduced to the concept of a fraction! However, only a fixed set of 26 problems have prerequisites; once a student finishes the first 26 problems, the student is randomly presented with problems from a large pool of remaining problems.

[1]The Common Core State Standards determine the math curriculum for students from kindergarten through high school in most US states: http://www.corestandards.org/.

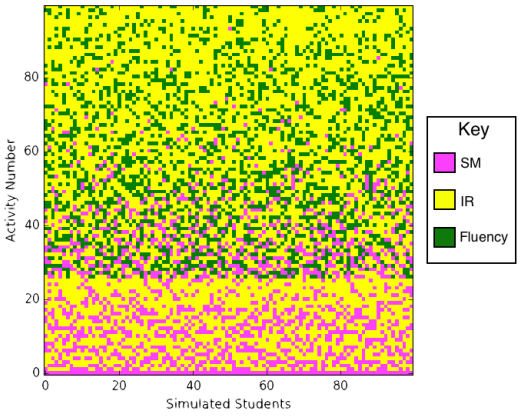

Figure 2: Simulation of potential activity type orderings. Each column represents a sequence of activity types for a student who was given 100 problems.

2.3 Data Collection

We collected data from students using the tutor in eight schools spanning two school districts. Students took a pretest, used the tutor for several sessions, and then took a posttest. The pretest and posttest consisted of 16 items covering conceptual and procedural understanding over skills involved with the three topics. Items were developed by building off of Common Core standards and prior assessment items developed for the Fractions Tutor. For our data analysis we used data from students who started each of the pretest and posttest (639 students).

3. ANALYSIS AND RESULTS

Our ultimate objective for this initial analysis was (1) to evaluate if the new tutor helped students improve their understanding of the material, and (2) to determine what static mix of activity types (SM, IR, and fluency) has the most effective learning outcomes.

3.1 Learning Gains

The mean pretest score is 5.82 ± 3.19 and the mean post test score is 8.23 ± 2.78 (both out of 16). Students significantly improved from pretest to posttest (paired t-test, $t(638) = 27.67$, $p < 10^{-110}$). The effect size was $d = 1.09$, which is considered a large effect size. These results demonstrate that our assortment of activity types can support learning gains, even when those activities are largely randomized.

3.2 Correlation of Variables

Exploratory data analysis revealed a substantial variation in both the number of activities done (mean: 49.6 ± 30.9) and the amount of time students had with the tutor (mean: 183.2 ± 82.3 minutes). Due to the prerequisite structure and semi-randomized ordering used, the number of activities and amount of time spent on the tutor influenced the relative proportion of each activity type that the students completed. To see this we can look at a set of possible simulated sequences that could have been given to students: Figure 2 shows 100 such sequences of 100 problems each. We can observe that students completing 26 problems or less would only receive SM and IR problems. In addition, because the total number of SM activities was fewer than the other two types of activities, if a student did a very large number of activities, the fraction of activities he/she completed would eventually be dominated by IR and fluency.

Predictor	Pearson's r	Partial Pearson's r	p-value
SM	-0.48	-0.15	$5.8 * 10^{-4}$
IR	0.26	-0.033	1
Fluency	0.44	0.18	$5.0 * 10^{-6}$

Table 1: Pearson's r between proportion of problem types and posttest scores, along with partial correlation coefficients when controlling for the number of problems done and amount of time spent on the tutor and Bonferroni corrected p-values for the partial correlations. Predictor variables represent the proportion of problems done by the student that were SM, IR, or F.

To help tease apart the strong correlation between the number of problems and the distribution of activity types completed, we computed the partial correlation between the proportion of problems belonging to each activity type and the posttest score, controlling for both the total number of problems done as well as the amount of time spent by the student. The results are shown in Table 1.

The decrease in magnitude between the raw correlation and partial correlation for each activity type tells us that the number of problems done and total time spent on the tutor accounts for some of the correlation with post test, as expected. More interestingly, the proportion of fluency problems is significantly positively correlated with the posttest scores even after considering the number of problems done and time spent. This suggests that having relatively more fluency problems is beneficial for students, beyond the fact that the students who did more fluency problems tend to have completed more problems; we will verify this with our hierarchical linear modeling. On the other hand, the proportion of SM problems is significantly negatively correlated with the posttest score even after accounting for time and number of problems.

To limit the extent to which students who got more time tended towards a certain mix of activity types, we restricted our subsequent analysis to only those students from one school district who had 150-200 minutes of tutor time in between pretest and posttest (resulting in 268 students).

3.3 Impact of Activity Proportions

The second key issue we wished to investigate was how student learning may be influenced by the mix of different activity types that they complete. To address this issue, we used hierarchical linear modeling to predict posttest scores as a function of the mix of SM, IR and fluency problems that a student completed. In the analysis below, we consider two-level HLMs that treat the class the student is from as a level-2 variable. Using a two-level model resulted in a better fit than just using linear regression. (We tried adding school as a level-3 variable, but this did not improve the

Predictor	Coefficient	p-value
Intercept	12.97	$5.4 * 10^{-9}$
Pretest Score	0.59	$< 1.0 * 10^{-15}$
Proportion SM	-11.20	$9.0 * 10^{-8}$
Proportion IR	-7.67	.021

Table 2: The coefficients of the HLM and their significance with a Bonferroni correction for doing four t-tests. (Satterthwaite approximations were used to compute the degrees of freedom.)

fit.) After trying a variety of models, we found that the best fitting model (in terms of cross-validated RMSE) was one of the simplest. The best model used only three predictor variables: pretest score, proportion of SM problems, and proportion IR problems. (Note that proportion of fluency problems is not a necessary predictor since the three proportions sum to one.) The coefficients for the level-1 variables of the HLM and their p-values are given in Table 2. We see the coefficient for the proportion of SM and the coefficient for the proportion of IR were significant and negative. Thus our model suggests fluency is the most effective activity type (since minimizing the proportion of IR and SM maximizes the posttest score) followed by IR, which agrees with our partial correlation analysis. The apparent lack of efficacy of SM problems may be because these items were substantially more time consuming for students to complete than the two other activity types. Thus even if SM problems are useful, their relative effectiveness per time spent may be lower than more active problems. This is also supported by recent results on the benefit of learning by doing [4].

If our model generalized to all possible sequences, it would suggest that students should do as many fluency problems as possible and not do any SM or IR problems. To allow for non-trivial mixes of activity types, the model would need to include interaction terms between the proportions of different activity types. Such models had statistically insignificant coefficients and worse fits than the model presented.

Nonetheless, it is important to note that any student who did fluency problems in our study necessarily also did SM and IR problems due to the prerequisite structure. Therefore we cannot reliably evaluate the value of a sequence consisting of only a single activity type using our model; such a sequence is *very* different from sequences the students actually received. Rather, the conclusion we can draw from our model is that if we were able to provide additional tutoring to students who already did many problems using our tutor, we should probably just give them more fluency problems.

Notice that our model includes no term for the total number of problems a student did (which we know correlates well with the posttest score). When adding such a term to our model, the fit was worse and the coefficient for that term was both small and statistically insignificant. This implies that the proportion of fluency problems is a better predictor than the number of problems a student did!

4. CONCLUSION

We have extended an existing ITS to include activity types that support all three learning mechanisms posited by the Knowledge-Learning-Instruction framework. In a large-scale classroom study, our ITS had learning gains with a large effect size. A preliminary analysis indicates that students who have a high percentage of fluency problems have the largest posttest scores, suggesting that fluency-building activities are most effective in helping students learn. However, many open questions remain. To what extent are SM and IR problems necessary? Does the appropriate mix of activity types differ for different topics (e.g. making fractions vs. fractions addition)? We hope to address these questions as we work towards our goal of learning personalized policies that best support robust student learning.

5. ACKNOWLEDGEMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A130215 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Dept. of Education.

6. REFERENCES

[1] J. E. Beck and J. Mostow. How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. In *Proc. of ITS*, pages 353–362, 2008.

[2] J. Champaign, K. F. Colvin, A. Liu, C. Fredericks, D. Seaton, and D. E. Pritchard. Correlating skill and improvement in 2 MOOCs with a student's time on tasks. In *Proc. of L@S*, pages 11–20, 2014.

[3] A. Corbett, M. McLaughlin, and K. C. Scarpinatto. modeling student knowledge: cognitive tutors in high school and college. *UMUI*, 10:81–108, 2000.

[4] K. Koedinger. Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In *Proc. of L@S*, 2015.

[5] K. R. Koedinger, A. T. Corbett, and C. Perfetti. The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, 36(5):757–798, 4 2012.

[6] J. F. Pane, B. A. Griffin, D. F. McCaffrey, and R. Karam. Effectiveness of Cognitive Tutor Algebra I at scale. *Educational Eval. & Policy Analysis*, 2013.

[7] Z. A. Pardos and N. T. Heffernan. Detecting the learning value of items in a randomized problem set. In *Proc. of AIED*, 2009.

[8] M. A. Rau, V. Aleven, and N. Rummel. Complementary effects of sense-making and fluency-building support for connection making: A matter of sequence? In *AIED*, 2013.

[9] M. A. Rau, V. Aleven, N. Rummel, and S. Rohrbach. Sense making alone doesnt́ do it: Fluency matters too! its support for robust learning with multiple representations. In *Proc. of ITS*, pages 174–184, 2012.

[10] R. J. Salden, K. R. Koedinger, A. Renkl, V. Aleven, and B. M. McLaren. Accounting for beneficial effects of worked examples in tutored problem solving. *Educational Psychology Review*, 22(4):379–392, 2010.

[11] K. VanLehn. The behavior of tutoring systems. *IJAIED*, 16(3):227–265, 2006.

[12] K. VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221, 2011.