Interleaved Practice with Multiple Representations: Analyses with Knowledge Tracing Based Techniques

Martina A. Rau
Human-Computer Interaction Institute
Carnegie Mellon University
marau@cs.cmu.edu

Zachary A. Pardos
Department of Computer Science
Worcester Polytechnic Institute
zpardos@wpi.edu

ABSTRACT

The goal of this paper is to use Knowledge Tracing to augment the results obtained from an experiment that investigated the effects of practice schedules using an intelligent tutoring system for fractions. Specifically, this experiment compared different practice schedules of multiple representations of fractions: representations were presented to students either in an interleaved or in a blocked fashion. The results obtained from posttests demonstrate an advantage of interleaving representations. Using methods derived from Knowledge Tracing, we investigate whether we can replicate the contextual interference effect, an effect commonly found when investigating practice schedules of different task types. Different Knowledge Tracing models were adapted and compared. A model that included practice schedules as a predictor of students' learning was most successful. A comparison of learning rate estimates between conditions shows that even during the acquisition phase, students working with interleaved representations demonstrate higher learning rates. This finding stands in contrast to the commonly found contextual interference effect when interleaving task types. We reflect on the practical and theoretical implications of these findings.

Keywords

Knowledge tracing, intelligent tutoring system, practice schedules, multiple representations, contextual interference.

1. INTRODUCTION

Educational data is highly complex, not only because learning is a complex process, but also because educational materials are complex. Learning materials in realistic educational settings generally cover a number of educational topics and use multiple representations. There is a substantial amount of evidence demonstrating that the use of multiple representations has a significant impact on students' learning [2,12]. When designing educational software that uses multiple representations, designers must decide how to temporally sequence the representations. In particular, it may matter whether representations are presented in a "blocked" manner (e.g., A – A – B – B) or in an interleaved manner (e.g., A – B – A – B). Research on contextual interference shows that interleaving task types leads to better learning results than blocking task types [7]. When working with multiple representations, a relevant question is whether practice with the different representations should be blocked or interleaved.

In the present paper we use log data obtained from an *in vivo* experiment (i.e., a rigorously controlled experiment in a real educational setting) that uses a successful type of intelligent tutoring system to help students learn about fractions while varying the practice schedule of multiple graphical representations. The experiment investigated the effect of practice schedules of graphical representations on students' knowledge of fractions assessed by posttests after they worked with the tutoring system. The goal of the present paper is to augment the findings from the traditional analysis of posttest data by applying a Knowledge Tracing algorithm to the log data. Analyzing student performance during the acquisition phase is particularly interesting when investigating the effects of practice schedules: a common finding is that interleaved practice schedules lead to better long-term retention and to better transfer than blocked schedules, but they often lead to worse performance during the acquisition phase [7]. Knowledge tracing, which tracks student knowledge over time, can be used to investigate learning differences between conditions during the acquisition phase [9].

In order to analyze the effect of practice schedules of multiple graphical representations on students' performance during the acquisition phase, we use a Bayesian Network model based on Knowledge Tracing [5]. Knowledge Tracing uses a two state Hidden Markov Model assumption of learning which uses correct and incorrect responses in students' problem-solving attempts to infer the probability of a student knowing the skill underlying the problem-solving step at hand. Previous research has demonstrated that extensions of knowledge tracing can be used to analyze effects of experimental conditions [9]. We combined this model with several other extensions to Knowledge Tracing to each of the four experimental conditions of the experimental study to investigate differences in model learning rates between the conditions in the Fractions Tutor.

The findings from the present paper are applicable to many other settings. Multiple graphical representations are used in a large variety of domains including science and mathematics. Whether to block or interleave representations is an important practical question in all of these domains.

2. THE FRACTIONS TUTORING SYSTEM

The Fractions Tutor used in the experiment was a type of Cognitive Tutor. Cognitive Tutors are grounded in cognitive theory and artificial intelligence. Cognitive Tutors have been shown to lead to substantial learning gains in a number of studies [6]. We created the tutors used in the present experiment with the Cognitive Tutor Authoring Tools (CTAT [1]. The design of the interfaces and of the interactions students engage in during problem-solving are based on a number of small-scale user studies that we conducted in our laboratory, as well as on Cognitive Task Analysis of the learning domain [3].

The Fractions Tutor included three interactive graphical representations of fractions (circles, rectangles, and number lines) and covered a comprehensive set of task types ranging from identifying fractions from graphical representations, creating graphical representations, reconstructing the unit of unit fractions and of proper fractions, identifying improper fractions from graphical representations, and creating graphical representations of improper fractions. Students solved each problem by interacting both with fractions symbols and with the interactive graphical representations. As is common with Cognitive Tutors, students received error feedback and hints on all steps. In

addition, each problem included conceptually oriented prompts to help students relate the graphical representations to the symbolic notation of fractions. These prompts were shown to be effective in an earlier study with the Fractions Tutor [12].

3. EXPERIMENT AND DATA

T	Blocked	Moderate	Interleaved	Increased
1	c-c-c-c-c-c	c-c-c-r-r-r	c-r-n-c-r-n	c-c-c-c-c-c
2	c-c-c-c-c-c	r-r-r-n-n-n	c-r-n-c-r-n	c-c-c-c-c-c
3	c-c-c-c-c-c	c-c-c-r-r-r	c-r-n-c-r-n	r-r-r-r-r-r
4	c-c-c-c-c-c	r-r-r-n-n-n	c-r-n-c-r-n	r-r-r-r-r-r
5	c-c-c-c-c-c	n-n-n-c-c-c	c-r-n-c-r-n	n-n-n-n-n-n
6	c-c-c-c-c-c	c-c-c-r-r-r	c-r-n-c-r-n	n-n-n-n-n-n
1	r-r-r-r-r-r	r-r-r-n-n-n	c-r-n-c-r-n	r-r-r-n-n-n
2	r-r-r-r-r-r	n-n-n-c-c-c	c-r-n-c-r-n	n-n-n-c-c-c
…	…	…	…	…
1	n-n-n-n-n-n	n-n-n-c-c-c	c-r-n-c-r-n	c-r-n-c-r-n
2	n-n-n-n-n-n	c-c-c-r-r-r	c-r-n-c-r-n	c-r-n-c-r-n
…	…	…	…	…

Table I. Practice schedule for each condition for all six task types (T). Each T was revisited three times. Students worked on nine problems per T. Each letter stands for one tutor problem and its representation: circle (c), rectangle (r), or number line (n).

The data used in this paper is based on an experimental study conducted with the Fractions Tutor during the end of the school year of 2009/2010. A total of 527 4th- and 5th-grade students from six different schools (31 classes) in the Pittsburgh area participated in the study during their regular mathematics instruction. We excluded students who missed at least one test day, and who completed less than 67% of all tutor problems (to ensure that students in the blocked condition encountered all three representations). This results in a total of N = 230 (n = 63 in blocked, n = 53 in moderate, n = 52 in fully interleaved, n = 62 in increased). Students worked with the Fractions Tutor for about 5h.

Table I illustrates the practice schedules of task types and representations for the experimental conditions. In all conditions, students worked on the same sequence of task types and revisited each task type three times. Students were randomly assigned to one of four conditions. In the blocked condition, students switched between the graphical representations after 36 problems. In the moderate condition, students switched representations after every three or six problems. In the fully interleaved condition, students switched representations after each problem. In the increased condition, the length of the blocks was gradually reduced from twelve problems at the beginning to a single problem at the end. To account for possible effects of the order of graphical representations, the order in which students encountered graphical representations was also randomized.

For the experiment, students' knowledge of fractions was assessed at three test times: before their work with the Fractions Tutor, immediately after, and one week after students finished working with the Fractions Tutor. The tests included four knowledge types: area model problems (i.e., problems that involved circles and rectangles), number line problems, conceptual transfer, and procedural transfer. The results from the test data (described in more detail by [11]) showed that the fully interleaved condition performed significantly better than the blocked condition, the moderately interleaved, and the increasingly interleaved conditions on conceptual transfer at the delayed posttest. Furthermore, there was a marginally significant advantage for the increasingly interleaved condition compared to the blocked, moderately interleaved, and fully interleaved conditions on number line items at both the immediate and the delayed posttests.

The analyses presented in the current paper are based on the tutor log data obtained from the Fractions Tutor. The log data provide the number of correct steps at a student's first attempt at solving a step in the tutor, the number of attempts until a step was correctly solved, the number of hints requested per step, and the time students spent per attempt.

4. BAYESIAN MODEL

We evaluated four Bayesian models based on the experiment log data. Two of the models were created for the purpose of analyzing the learning rates of the conditions in the experiment while the other two were used as baseline models to gauge the relative predictive performance of the new models.

4.1 Learning Analysis Models

One of the simplifying assumptions made by the standard Bayesian Knowledge Tracing model [5] is that there is a probability that a student will transition from the unlearned to the learned knowledge state at each opportunity regardless of the particular problem just encountered or practice schedule of the student. Our model hypothesis corresponds to the hypothesis of the experiment that different practice schedules within a task type may be more or less effective at allowing students to acquire the skill being practiced. Thus, we depart from the Knowledge Tracing assumption of a single learning rate per skill and instead fit a separate learning rate for each of the four different practice schedule conditions defined in the experiment.

To model different learning rates within Knowledge Tracing, we adapted modeling techniques from prior work which evaluated the learning value of different forms of tutoring in (non-experiment) log data of an intelligent tutor [9]. Different representations of fractions are expected to result in different degrees of difficulty in solving the tutor problem [4]. In our condition and representation analysis model we used techniques from KT-IDEM [10] to model different guess and slips for problems depending on the representation used in the tutor problem.

We employed two models which served as benchmarks for model fit and designed two novel models for evaluating learning differences among the experiment conditions. We compared four Bayesian models all of which were based around Knowledge Tracing. Figure 1 provides an overview of the different models that we compared. The Prior-Per-Student Model [8] includes the students' individualized prior knowledge, the Condition-Analysis Model includes students' prior knowledge and models the effect of experimental condition (C). Finally, the Condition-Representation-Analysis Model incorporates students' prior knowledge (S), condition (C), and the graphical representation encountered by each student in each problem (R).

4.2 Model Fitting Procedure

In order to determine model fit by task type, we analyzed the log data by tasks type. For the evaluation of predictive performance, reported in the next section, a 5-fold cross-validation at the student level was used. For the reporting of learning rates by practice schedule, all data was used to train the model.

The parameters in all four models were fit using the Expectation Maximization algorithm implemented in Kevin Murphy's Bayes Net Toolbox. For the Condition-Representation-Analysis Model the number of parameters fit per task was 12 (2 prior + 4 learn rate + 3 guess + 3 slip). Probabilities of knowledge are fixed at 1 if the skill was already known, $\boldsymbol{P}(\boldsymbol{L_{n-1}})$ = 1, to represent a zero chance of forgetting, an assumption made in standard KT.

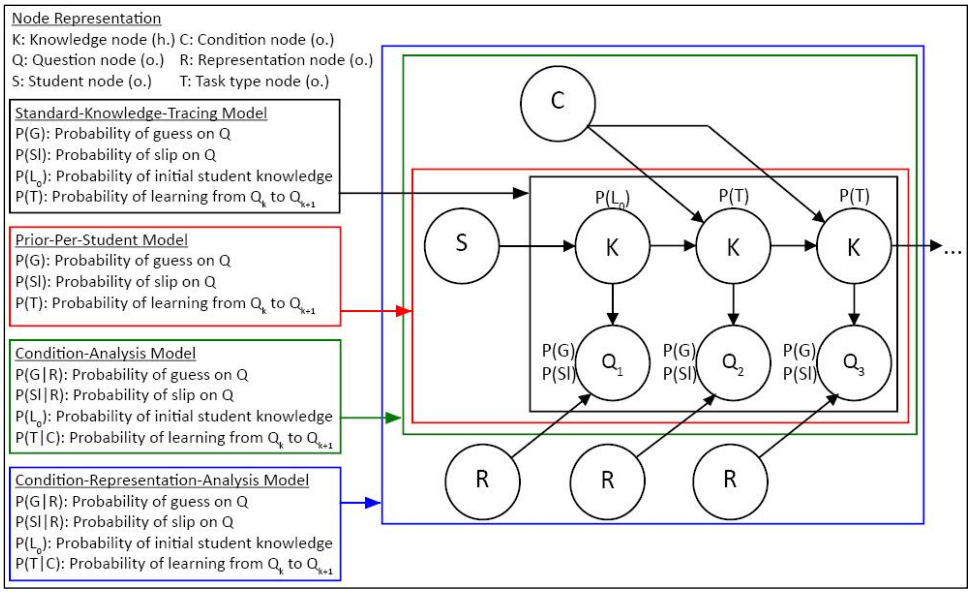

Fig. 1. Overview of the four different Bayesian Networks tested, with observed (o.) and hidden (h.) nodes.

5. EVALUATION RESULTS

	Model	RMSE	AUC
1	Condition-Representation-Analysis Model	0.3427	0.6528
2	Standard-Knowledge-Tracing Model	0.3445	0.6181
3	Condition-Analysis Model	0.3466	0.5509
4	Prior-Per-Student Model	0.3469	0.5604

Table II. Cross-validated prediction results summary of the four models using RMSE and AUC metrics

To evaluate the predictive accuracy of each of the student models mentioned in section 2, we conducted a 5-fold cross-validation at the student level. By cross-validating at the student level we can have greater confidence that the resulting models and their assumptions about learning will generalize to new groups of students. The metric used to evaluate the models is root mean squared error (RMSE) and Area Under the Curve (AUC). Lower RMSE equals better prediction accuracy. For AUC, a score of 0.50 represents a model that is predicting no better than chance. An AUC of 1 is a perfect prediction.

As shown in Table II, the Condition-Representation-Analysis Model has the lowest RMSE with .3427 as well as the best AUC. We conclude that the Bayesian Network that includes students' prior knowledge (S), experimental condition (C), and representations used for a certain problem (R) provides the best model fit. All predictions were statistically significantly different from one other by a paired t-test of squared errors.

Table III provides the learning rates obtained from the Condition-Representation-Analysis Model for each condition for each of the task types that the tutoring system covered. Overall, the learning rate estimates align with the results obtained from the posttest data: the interleaved condition demonstrates higher learning rates than the blocked condition. The task types were as follows: (1) identifying fractions from representations, (2) making representations of fractions, (3) reconstructing the unit from unit fractions, and (4) reconstructing the unit from proper fractions. On task type (5) identifying improper fractions from representations and (6) making representations of improper fractions.

TT	Blocked	Moderate	Interleaved	Increased
1	0.0061	0.0061	0.0080	0.0072
2	0.0019	0.0032	0.0065	0.0036
3	0.0149	0.0059	0.0337	0.0030
4	0.0037	0.0022	0.0035	0.0014
5	0.0108	0.0220	0.0124	0.0130
6	0.0043	0.0107	0.0078	0.0090
Overall	0.0062	0.0056	0.0120	0.0062

Table III. Learning rates by task type (TT) and condition from the Condition-Representation Analysis Model.

The learning rates by task type provide more specific information on the nature of the differences between conditions in learning rates. For all but the fourth task type, the fully interleaved condition demonstrates a higher learning rate than the blocked condition. This difference was statistically significant for tasks 1, 2 and 3 ($p < 0.05$) and moderately significant for task 5 ($p = 0.06$). The same binomial test as was used in [9] was employed here to test for significance. The interleaved condition achieved the highest overall learning rate which was twice that of any other condition. This was despite having the second lowest percent correct among responses in the acquisition phase.

6. DISCUSSION

The findings from the Bayesian Networks support and augment the findings from the posttest data in several ways. First, the finding that the Condition-Representation Analysis Model provides the best fit to the log data confirms the overall finding from the posttest data of the experiment that practice schedules of multiple representations matter. It also highlights that per item level parameters are greatly beneficial, especially when the problem opportunities involve different cognitive operations, such as solving problems with different representations. Furthermore, the finding that the representation used in a tutor problem is a useful predictor of learning confirms that different graphical representations provide different conceptual views on fractions in a way that influences how students understand fractions [4].

Second, the learning rate estimates per condition support the finding from the posttest data that interleaved practice schedules of multiple graphical representations of fractions lead to better learning than blocked practice schedules. This finding is interesting, because the literature on contextual interference shows that interleaved practice schedules often impair performance during the acquisition phase [7]. It is assumed that temporal variation between consecutive problems interferes with immediate performance since students have to adapt their problem-solving procedures each time they encounter a new task. This interference leads to higher processing demands and lower performance during the acquisition phase, but results in better long-term retention and transfer performance later on. Hence, one might expect that higher learning gains in the interleaved condition become apparent only in the posttest data, but not during the acquisition phase, because they might be "masked" by impaired performance due to interference. Our findings show, however, that an intervention that is assumed to lead to impaired performance during the acquisition phase nonetheless leads to a learning advantage that is not only detectable in higher posttest performance but also during the acquisition phase using our experiment adapted Bayesian model. Bayesian Network analyses allowed us to detect learning gains that may be too subtle to detect during the acquisition phase when relying on performance. We believe this was able to be achieved thanks to the item level modeling that distinguished learning from variation in problem difficulty.

Finally, the differences between learning rate estimates between task types yield important insights into the effectiveness of the tutor task types that will help improve the tutoring system in future iterations. Bayesian Network analysis provides us with a useful tool that can help us evaluate this iterative improvement of the tutoring system at a much finer grain size than through the traditional analysis of posttest data. This technique also allowed for analysis to be accomplished without pre or post test data.

The results from the Bayesian Network analysis presented in this paper yield interesting insights that are both of theoretical and practical significance. Our results confirm the finding from our previous experiment [18] that interleaving representations leads to better learning than blocking representations and extend the finding by demonstrating that the advantage of interleaved practice is apparent also during the acquisition phase. This finding is of practical relevance as it demonstrates that face-value methods, such as percent correct during the acquisition phase, do not provide sufficient information to evaluate an educational intervention. Since many domains use multiple graphical representations to augment instructional materials, we believe that our findings have the potential to generalize across a wide range of learning materials. Furthermore, the analysis of learning rates by condition allows us to identify parts of the Fractions Tutor curriculum that need to be improved as they do not seem to help students learn. Bayesian Network analyses can help us make sense of the complex educational data that we obtain from the rich settings in which education takes place, and hence, help us understand complex learning processes.

7. ACKNOWLEDGMENTS

This work was supported by the National Science Foundation, REESE-21851-1-1121307, "Graduates in K-12 Education" Fellowship, award number DGE0742503, and by the Pittsburgh Science of Learning Center which is funded by the National Science Foundation, award number SBE-0354420. In addition, we would like to thank Haya Shamir, Ken Koedinger, John Stamper, and the students and teachers who participated in our study.

8. REFERENCES

[1] Aleven, V., Mclaren, B. M., Sewall, J., & Koedinger, K. R. 2009. A new paradigm for intelligent tutoring systems: Example-tracing tutors. International Journal of Artificial Intelligence in Education, 19(2), 105-154.

[2] Ainsworth, S., Bibby, P., & Wood, D. 2002. Examining the effects of different multiple representational systems in learning primary mathematics. Journal of the Learning Sciences, 11(1), 25-61.

[3] Baker, R. S. J. D., Corbett, A. T., & Koedinger, K. R. 2007. The difficulty factors approach to the design of lessons in intelligent tutor curricula. International Journal of Artificial Intelligence in Education, 17(4), 341-369.

[4] Charalambous, C. Y., & Pitta-Pantazi, D. 2007. Drawing on a Theoretical Model to Study Students' Understandings of Fractions. Educational Studies in Mathematics, 64(3), 293-316.

[5] Corbett, A. T., & Anderson, J. R. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4, 253-278.

[6] Corbett, A. T., Koedinger, K., & Hadley, W. S. 2001. Cognitive tutors: From the research classroom to all classrooms. In Technology enhanced learning: Opportunities for change. (pp. 235-263). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

[7] De Croock, M. B. M., Van Merrienboer, J. J. G., & Paas, F. G. W. C. 1998. High versus low contextual interference in simulation-based training of troubleshooting skills: Effects on transfer performance and invested mental effort. Computers in Human Behavior, 14(2), 249-267.

[8] Pardos, Z. & Heffernan, N. (2010) Navigating the parameter space of Bayesian Knowledge Tracing models: Visualizations of the convergence ofthe Expectation Maximization algorithm. In Baker et al. (Eds.) Proceedings of the 3rd International Conference on Educational Data Mining (pp. 161-170).

[9] Pardos, Z. A., Dailey, M. D., Heffernan, N. T. (2010) Learning what works in ITS from non-traditional randomized controlled trial data. In Aleven et al. (Eds.) Proceedings of the 10th International Conference on ITS2010 (pp. 41-50).

[10] Pardos, Z. & Heffernan, N. (2011) KT-IDEM: Introducing Item Difficulty to the Knowledge Tracing Model. In Konstant et al. (Eds.) In Proceedings of the 20th International Conference on UMAP 2011 (pp. 243-254).

[11] Rau, M. A., Aleven, V., Tunc-Pekkan, Z., Pacilio, L., & Rummel, N. (accepted). How to schedule multiple graphical representations? A classroom experiment with an intelligent tutoring system for fractions. To appear in the proceedings of ICLS 2012.

[12] Rau, M. A., Aleven, V., & Rummel, N. 2009. Intelligent Tutoring Systems with Multiple Representations and Self-Explanation Prompts Support Learning of Fractions. In Dimitrova et al. (Eds.), Proceedings of the 14th International Conference on AIED (pp. 441-448).