Desirable Difficulty and Other Predictors of Effective Item Orderings

Steven Tang   Hannah Gogel   Elizabeth McBride   Zachary A. Pardos
University of California, Berkeley
Tolman Hall
Berkeley, CA, USA
{steventang, hgogel, bethmcbr, pardos} @berkeley.edu

ABSTRACT

Online adaptive tutoring systems are increasingly being used in classrooms as a way to provide guided learning for students. Such tutors have the potential to provide tailored feedback based on specific student needs and misunderstandings. Bayesian knowledge tracing (BKT) is used to model student knowledge when knowledge is assumed to be changing throughout a single assessment period. The basic BKT model assumes that the chance a student transitions from ”not knowing” to ”knowing” after each item is the same, with each item in the tutor considered a learning opportunity. It could be the case, however, that learning is actually context sensitive; context in our analysis is the order in which the items were administered. In this paper, we use BKT models to find such context sensitive transition probabilities in a mathematics tutoring system and offer a methodology to test the significance of our model based findings. We employ cross validation techniques to find models where including item ordering context improves predictive capability compared to the base BKT models. We then use regression testing to try to find features that may predict the effectiveness of an item ordering.

Keywords

Item Ordering, Bayesian Knowledge Tracing, Item Difficulty

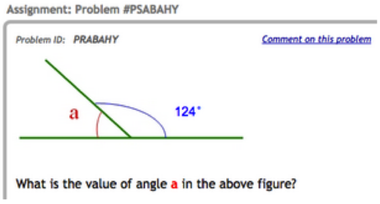
1. INTRODUCTION

Online adaptive tutors are increasingly being used in classrooms as supplements to traditional instruction. Some systems, such as the ASSISTments [4] platform used for middle school math subjects, provide scaffolding or hints to students upon request or when the student answers a question incorrectly. In this paper, we focus on employing the Bayesian knowledge tracing (BKT) model of student learning but with the hypothesis that learning could be *context sensitive*. In this case, the context is the order that items of a particular skill are administered in.

2. BACKGROUND

2.1 ASSISTments Data

The data set analyzed in this paper comes from use of the ASSISTments platform in AY 2012-2013. The data set is publicly available and is rich with information that has been mined by other research projects [7] [9]. In this paper, we focus on the *Skill Builder* sequences used in ASSISTments, where a problem set consists of items given in a random order, generated from a set of templates. Items generated from these templates are assumed to be answerable with knowledge of a single underlying knowledge component (KC). For example, one problem set might contain three item templates. Each template can be populated with a set of numbers to generate an item; thus many different items can be derived from a single template. The number of templates per problem set varies; in this paper, we look at problem sets with between 2 and 6 templates. The number of items delivered to the student depends on the student's performance; in the Skill Builder set, mastery is assumed to occur after three consecutive correct responses. Each template in a Skill Builder sequence has an associated method of assistance; it is either a *hint* template or a *scaffolding* template. Scaffolding templates are bundled with a set of simpler questions to guide the student through the ideas in the item, while hint templates have guiding statements available to assist the students (usually the final hint provides the exact answer to the item).



Assignment: Problem #PSABAHY

Problem ID: PRABAHY

Comment on this problem

What is the value of angle a in the above figure?

Figure 1: Example of an item in the ASSISTments database

3. METHODS AND ANALYSIS

3.1 Bayesian Knowledge Tracing

Bayesian knowledge tracing [3] assumes a binary representation of student knowledge. Figure 2 depicts a BKT model representation as a hidden Markov model (HMM). The basic BKT model is shown inside the dashed portion of the figure. O_1 through O_4 are binary indicators of correctness at opportunities 1 through 4. K_1 through K_4 represent the latent knowledge of the KC (assumed to be 0 or 1) at opportunities 1 through 4. In between each K_i and K_{i+1}, there is an arrow representing a probability of *transition*, or learning. Guess and slip parameters can be assumed to be equal among all items or can be item-specific [10].

3.2 Item Ordering Effects

The Skill Builder sequences in the ASSISTments platform pick from a set of templates at random to generate items for

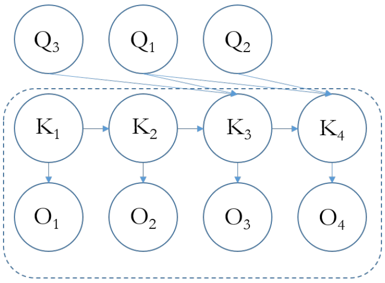


Figure 2: BKT Model. The dashed portion represents the basic BKT model, and the Q nodes represent the item order modification.

the student. However, it is our hypothesis that there may exist pedagogically more advantageous orderings of problems than the default random orders. Data mining and learning analytics techniques have been used to create process models and determine the most effective order of events for learners in online science education [8], as well as for finding patterns where students exhibited patterns of self-regulated learning [6]. Investigating the effects of item ordering can help both researchers and teachers, bridging the gap between educational theory and practice.

The BKT model could be extended to model a transition probability per particular item ordering. For example, one student might receive items from templates in the order of (3, 1, 2, ...) while another student might receive items from templates in the order of (1, 3, 2, ...). Over a number of such permutations, the BKT model could estimate a separate transition probability associated with items in the order (3, 1) as opposed to (1, 3). Figure 2 depicts how this new model might be formulated as an HMM, where items in the order of (3, 1, 2) are seen by the student. Note that the probability of knowledge at K_3 is influenced by seeing question 3 followed by question 1. Other students will be given items in different and random orders, allowing for all possible combinations of item order pairs to be analyzed. This model is drawn from work by Pardos and Heffernan [9]. We extend this work by finding significant improvements in predictive accuracy with the item order model by looking at the mean absolute errors produced by both the basic BKT and the item order model.

3.3 BKT model fitting

Among the Skill Builder response sets (SBs) from the 2012-2013 ASSISTments data set, we only looked at sets with more than 2000 student responses, more than 250 students, and between 2 and 6 (inclusive) templates. There were 112 Skill Builders that met these criteria, with 130,496 student response streams and 606,948 responses. Two BKT models, estimated using the XBKT code base, were fit to each of the 112 SBs. The first model was standard BKT (baseline), where every item was assumed to have the same transition probability. In our standard BKT model, every template type was allowed to have its own guess and slip parameters. The second model allowed for both different guess and slips per template and different transition probabilities based on the previous two items administered. We enabled different guess and slips per template for our baseline model so that any difference between models would be attributed to the different item order learning transitions. Additionally, we modeled a transition probability for each template specifically when that template was the first item administered in the sequence.

3.4 CV prediction to identify item orders of interest

To obtain statistical confidence in the generalization of a certain item ordering to unobserved students, we performed 5-fold cross validation (CV) on the data. This process starts by fitting both base and item order BKT models on a randomly selected 80% of student response data, and then using the trained models to predict student responses in the held out 20%, called the test set.

By comparing the predicted responses to the actual responses, Mean Absolute Errors (MAE) were obtained for both the base and the item order models. The error rates were then compared using a paired t-test for each possible item order. Out of the 1789 possible item orders among all Skill Builder problem sets, 605 item orders were found to have statistically significant error differences between the two predictive models at the .05 level. Among the 605 item orders, 157 had their responses predicted better by the base BKT model (by an average rate of .0138), while the remaining 448 item orders had their responses predicted better when using the item order model (by an average rate of .0173). It is important to note that the item orders in this section include ordering situations where the same template is administered twice in a row. The result that a portion of the item orders had better response prediction when using the base BKT model is not surprising, considering that each addition of a single new template to an SB increases the number of potential item orders dramatically. Thus, as the number of templates increases, the number of responses per item order decreases, resulting in less data per parameter for the model to learn from. The occurrence of 448 item orders whose responses were better predicted by the item order model suggests that the item order model could be able to uncover effective (or ineffective) item orderings.

Figure 3 shows the distribution of learn rates from both the basic and the item order BKT models. In the basic BKT model, a learn rate represents the rate at which a student is expected to learn (if they did not already know it) the latent knowledge component after seeing any item. In the item order BKT model, learn rates are modeled per item order pair, thus representing the rate a student is expected to learn a knowledge component after seeing a particular order of two items. The combination of the item order model with the cross validation approach provides a procedure that can determine when the item order model provides more accurate predictions compared to the base BKT model. Such a procedure can reveal when an item ordering might be considered effective or ineffective.

3.5 Regression analysis

Regression analyses (212,858 student responses) were run on the 448 item orders found to be significantly better fitting from the cross validation approach in order to find predictors of the item order learn rates. For the regression analyses,

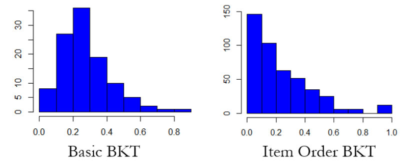


Figure 3: Distribution of learn rates

we extracted template level features from both templates in an item order. Features included are: average time to first response (milliseconds), percent correct on first problem attempt, average number of attempts, problem type (text response or radio button/multiple choice), difference in time to first response between Template A and Template B (where Template A is the first item in an ordering), difference in percent correct between Template A and Template B, whether Template A offered hints or scaffolding as assistance, and the *individual* learn rates for Templates A and B.

In our first model, stepwise regression was used regressing item order learn rate on these features (R^2=.17, F = 46.19, p<.01). The only features that were found to be significant at the .05 level were the learn rate of Template A, which had a negative effect on item order learn rate for the pair (β = -0.13, p = .01), and the learn rate of Template B, which had a positive effect (β = .502, p < .01) in the model.

Our second model only included item orderings where Template A was a scaffolding problem (R^2=.37, F = 11.47, p < .01). All of the features from the first model were included except for problem type due to lack of variation. Features unique to scaffolding problems were added as potential predictors: problem type of the associated sub-questions and percentage of scaffolding problems (including sub-questions) answered correctly. Average attempts on Template A (β = .93, p < .01) and the learn rate for Template B (β = .58, p < .01) had a positive effect on the item order learn rate. When the scaffolding for Template A consisted of text responses, the learn rate of the ordering decreased (β = -.13, p < .01).

The third model was fit using only orders where Template A was a hint item (R^2=.22, F = 20.94, p < .01). Hint features included percentage of students who went through all the hints on Template A and average amount of template hints seen. Average number of attempts on Template A (β = .27, p < .01), average milliseconds to first response on Template A (β = < .01, p = .03), percentage of students who accessed all of the hints on Template A (β = .71, p < .01), learn rate of Template A (β = -0.16, p <.01), and the learn rate for Template B (β = .43, p < .01) were significant predictors.

Regression analyses were also conducted to look for feature predictors of individual template learn rates for the 321 individual templates included in these 448 orderings. Percent correct on the template (β = .31, SE = .1, p < .01) and the item requiring a text response (β = .14, SE = .04, p < .01) were significant predictors (R^2=.06, F = 10.84, p < .01).

The primary unexpected result from the regression findings is that a *lower* learn rate of Template A predicts a higher learn rate for the ordering. It is important to note that this effect may be due to constraints in our current model. The individual learn rate of Template A is calculated when Template A occurs as the first item in a problem set presented to a student. That Template A is also included as part of an item ordering pair made up of the first and second items in the administered problem set. If the learn parameter for Template A is high, the knowledge component is already known (and has already been learned) by the time we consider the learn rate for the ordering including Template A. However, this phenomenon does not occur for template B of the item ordering, as Template B would not be the first template seen by the student in this case. In order to alleviate the discrepancy between the correlations, single template learn rates should be calculated from all template occurrences throughout administration in future work.

3.6 Desirable difficulty

In previous proof-of-concept work [11], a qualitative analysis was performed to examine what might make certain item orderings more effective than other item orderings. One feature of item pairs that became obvious was that not all items had exactly the same level of difficulty. In addition, some effective orderings contain a harder item first whereas other effective orderings contain an easier item first. One potential hypothesis that can help explain this difference in item ordering and difficulty is that of "desirable difficulties". In a series of studies, Bjork and colleagues determined that some challenges to performance during learning activities may actually contribute to greater learning [1] [2] [5]. By introducing "desirable difficulties" that help learners engage in the active processing of information, learning tasks that may be perceived as challenging or inefficient may prove more beneficial in the long run than those completed with high fluency.

In the case of item orderings where the first problem is more difficult than the second, the first (more difficult) problem may introduce a desirable difficulty, leading the student to learn more than they would with an easier problem. This learning then carries over into the second problem in the pair, thus leading to a higher overall rate of learning. This hypothesis works towards explaining our finding that a lower learn rate of the first template predicts a higher learn rate for an item ordering. When the first problem is easier than the second, this might be an instance where the material is better learned through a gentler or simpler introduction, as perhaps the second problem might be more difficult than is "desirable". In this case, a student would not properly learn from the more difficult problem unless it were preceded by an easier problem that would serve as a scaffold.

Using data from the BKT model to examine this hypothesis, we looked at how the difference between prior knowledge (at the start of an SB) and the percent correct on a template (as a proxy for template difficulty) compared to the probability of learning using regression. Finding *no difference* between a student's prior knowledge and the percent correct for a given template might show when an item has an "appropriate" difficulty. In this case, the difficulty of the item closely matches the prior knowledge of the student. Pedagogically, for an item to help the student learn, the difference between the student's prior knowledge and the item difficulty should be negative; in other words, the difficulty of the item should

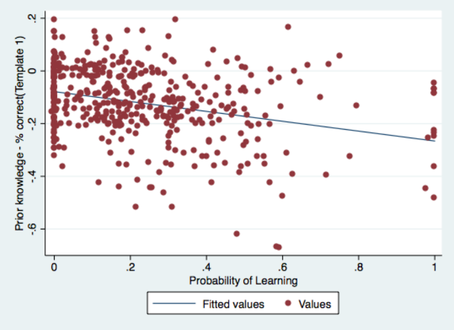


Figure 4: Scatterplot using template A data

be above the level of the student's prior knowledge to promote learning.

Regressing the difference between prior knowledge and item difficulty (percent correct) on the probability of learning showed statistical significance at the 0.01 level. This statistical significance held when using the difference between prior knowledge at the beginning of an SB and the percent correct on the first item in a pair (Template A), as well as the difference between prior knowledge and the percent correct for the second item in the pair (Template B). Using the percent correct for Template A to find the difference between the student's prior knowledge and the item difficulty had a correlation of -0.3039 with the probability of learning, while using Template B had a -0.2146 correlation with the probability of learning. These correlations are both relatively high, showing enough relationship between the variables to warrant further exploration in this area.

Similar to the correlations, the regressions were also run using percent correct from Template A and from Template B in the difference between prior knowledge and item difficulty. For Template A the coefficient for regressing the difference between prior knowledge and item difficulty (percent correct) on the probability of learning was -0.187 (R^2 =0.09, F=45.39); using template B, the coefficient was -0.120 (R^2= 0.046, F=21.53). The negative correlations, as well as negative coefficients in each of the regressions, show that the more negative the difference between prior knowledge and item difficulty becomes (the larger the difference between these two variables in the right direction for a "desirable difficulty"), the greater the probability of learning becomes. A scatterplot showing the relationship between these variables can be seen in Figure 4.

4. LIMITATIONS AND FUTURE WORK

The findings from this paper suggest that the item order BKT model combined with the use of a cross-validation technique show promise in uncovering learning mechanisms not apparent when just the base BKT model is used. The cross-validation approach confirmed that some item order models had better predictive capabilities compared to the base BKT models. Thus, statistically reliable suggestions can be made about item order delivery, and more research into item ordering is warranted, especially using such a cross-validation approach.

The results from the regression were somewhat surprising, where a lower individual learn rate from the first template in an ordering predicted a higher overall learn rate for the ordering. We hypothesize that this could be due to a constraint in our item order model, where individual learn rates of templates were modeled using only instances of that item when it appeared as the first item in a sequence. This hypothesis can be investigated in future research using a modified item order model.

5. REFERENCES

[1] M. Anderson, J. Neely, E. Bjork, and R. Bjork. *Memory, Chapter 6: Interference and inhibition in memory retrieval.* Academic Press, 1996.

[2] R. W. Christina and R. A. Bjork. *In D. Druckman and R. A. Bjork (Eds.), In the mind's eye: Enhancing human performance: Optimizing long-term retention and transfer.* Washington, DC: National Academy Press, 1991.

[3] A. T. Corbett and J. R. Anderson. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *The Journal of User Modeling and User-Adapted Interaction*, 4:253–278, 1995.

[4] M. Feng, N. Heffernan, and K. R. Koedinger. Addressing the assessment challenge in an Intelligent Tutoring System that tutors as it assesses. *The Journal of User Modeling and User-Adapted Interaction*, 19:243–266.

[5] V. Halamish and R. Bjork. When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37.

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

[7] J. Ocumpaugh, R. Baker, G. S., N. Heffernan, and C. Heffernan. Population validity for Educational Data Mining models: A case study in affect detection. *British Journal of Educational Technology*, 45(3):487–501, 2014.

[8] L. M. P. Reimann and M. Bannert. e-Research and learning theory: What do sequence and process mining methods contribute? *British Journal of Educational Technology*, 45:528–540, 2014.

[9] Z. A. Pardos and N. T. Heffernan. Determining the significance of item order in randomized problem sets. In *Proc. of the 2nd International Conference on Educational Data Mining*, 2009.

[10] Z. A. Pardos and N. T. Heffernan. KT-IDEM: Introducing item difficulty to the knowledge tracing model. In *Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization (UMAP)*, 2010.

[11] S. Tang, E. McBride, H. Gogel, and Z. A. Pardos. Item ordering effects with qualitative explanations using online adaptive tutoring data. In *Proceedings of Works-in-progress at the second ACM conference on Learning@ scale*, 2015.