Yang, J. C. et al. (Eds.) (2018). Proceedings of the 26th International Conference on Computers in Education. Philippines: Asia-Pacific Society for Computers in Education

Modeling the Learning That Takes Place Between Online Assessments

Ryan S. BAKER[a,b*], Sujith M. GOWDA[c] & Eyad SALAMIN[d]
[a]*Teachers College, Columbia University, USA*
[b]*University of Pennsylvania, USA*
[c]*Alpha Data Labs, Inc., USA*
[d]*Alef Education Consultancy LLC, United Arab Emirates*
*ryanshaunbaker@gmail.com

Abstract: Student models for adaptive learning environments and intelligent tutoring systems typically assume a paradigm of use where a student completes exercises or activities, and learns from those exercises or activities. However, many modern systems, including MOOCs, intersperse declarative content or lecture with assessment of the learning from this content. In this paper, we present a variant of a common student modeling algorithm, Bayesian Knowledge Tracing, which assumes that most learning occurs during use of declarative content rather than between exercises. We compare this algorithm's predictive ability to classic Bayesian Knowledge Tracing and another common algorithm, Performance Factors Assessment. We find that our new algorithm, BKT-PL, performs slightly better than algorithms designed for the standard intelligent tutoring paradigm. Moreso, we can use BKT-PL to determine which declarative content is most and least effective, to drive iterative re-design.

Keywords: Student modeling, adaptive learning, Bayesian Knowledge Tracing

1. Introduction

The algorithms designed to model student learning within intelligent tutoring systems and adaptive learning have become increasingly effective in recent years (Khajah et al., 2016; Wilson et al., 2016). While there has been ongoing debate about whether approaches based on machine learning, psychometrics, or interpretable student modeling are superior (e.g. Khajah et al., 2016; Wilson et al., 2016), these algorithms have typically shared key assumptions. One such assumption is the assumption that learning is concurrent with practice; these algorithms have generally assumed that students learn as they are completing exercises, and have assumed that students learn as they are assessed, from features such as hints and feedback (Desmarais & Baker, 2012).

This assumption is a reasonable one for most intelligent tutoring systems and is indeed one of the key ways that student modeling in adaptive learning systems typically differs from the types of tests that are typically the subjects of psychometric assessment (Desmarais & Baker, 2012). However, there is a middle ground between the type of ongoing learning seen in intelligent tutoring systems and the intentional design to avoid learning during assessment seen in psychometric examinations: periodic learning.

In this third case, periodic learning, a student alternates between receiving learning experiences and being assessed as to their learning. This case is common in many widely-used learning systems today. To give one example, the considerable majority of contemporary Massive Online Open Courses (MOOCs) ask students to watch videos or interact with other instructional content, and then provide quizzes which the student completes afterwards. Although this practice has been criticized (Koedinger et al., 2015), an argument can be made for the use of extended instruction of challenging conceptual materials. For example, this practice underlies much of the use of Khan Academy, a widely-used platform which has initial positive evidence for efficacy (e.g. FSG, 2014). It is also seen in the Alef NextGen platform, a platform currently used by learners in the Middle East and New York City, which we will discuss below. Moreso, extended explanations have

been a part of many successful intelligent tutoring systems, from the original LISP Tutor (Anderson, Conrad, & Corbett, 1989) to recent extensions to the ASSISTments system today (Heffernan et al., 2016). These systems represent a hybrid between continual learning/assessment and periodic learning/assessment, suggesting that the standard assumptions of most student modeling algorithms may also not hold in these cases.

In this paper, then, we introduce a new algorithm, BKT-PL (Periodic Learning), which is designed for this situation. BKT-PL assumes that learning is not concurrent with assessment, but instead assumes that periodic learning occurs. We analyze this algorithm in terms of prediction of future student performance, comparing it to algorithms that assume continual learning and assessment, to see whether BKT-PL better represents student knowledge in a periodic learning system. We then analyze BKT-PL's possible applications for understanding the quality of different conceptual content.

1.1 Student Modeling Algorithms

Since the publication of Corbett and Anderson's landmark paper on student knowledge modeling in 1995 (Corbett & Anderson, 1995), the framework they proposed – Bayesian Knowledge Tracing – has remained the most widely-used student knowledge modeling framework in adaptive learning, being used in adaptive learning systems used by hundreds of thousands of students a year (Koedinger & Corbett, 2006). Bayesian Knowledge Tracing has several virtues – reasonably good prediction of future performance within the learning system, effective prediction of future performance outside the learning system, an interpretable estimate of student knowledge, meaningful parameters that can be used to understand the properties of the learning system, and considerable extensibility to handle variant learning situations. Other frameworks perform better than classic BKT on prediction of future performance within the learning system (e.g. Khajah et al., 2016), but no other framework captures each of these elements.

Classic BKT assumes that, at any given time, a student either knows a skill or does not know that skill. The student starts with an initial probability $P(L_0)$ of knowing the skill. At any opportunity to demonstrates that skill (an "item"), the student can either get the item correct or incorrect. If the student knows the skill, their probability of correctness is governed by the probability that they slip and make an error despite knowing the skill: P(S). If the student does not know the skill, their probability of correctness is governed by the probability that they guess and produce a correct response despite not knowing the skill: P(G). If the student does not know the skill, they have a certain probability P(T) of learning the skill in the course of going from one problem to the next. The student's knowledge and behavior are estimated using the following set of formulas, based on Bayes' Theorem as shown in Figure 1:

$$P(L_{n-1}|Correct_n) = \frac{P(L_{n-1}) * (1 - P(S))}{P(L_{n-1}) * \big(1 - P(S)\big) + \ (1 - P(L_{n-1})) * (P(G))}$$

$$P(L_{n-1}|Incorrect_n) = \frac{P(L_{n-1}) * P(S)}{P(L_{n-1}) * P(S) + \ (1 - P(L_{n-1})) * (1 - P(G))}$$

$$P(L_n|Action_n) = \ P(L_{n-1}|Action_n) + \big((1 - P(L_{n-1}|Action_n)\big) * P(T))$$

Figure 1. Bayesian Knowledge Tracing algorithms – used both in the original formulation of BKT and in the extension BKT-PL introduced in this paper

One of the key virtues of BKT is its extensibility. Modifications to the underlying form of BKT and the semantics of its parameters have been used for a range of applications – to modify BKT for a specific MOOC with virtual labs and complex multi-step homework (Pardos et al., 2013), to add item properties (Pardos & Heffernan, 2011), to estimate the moment of learning (Baker, Goldstein, & Heffernan, 2010), and to estimate the probability that a student error represents

carelessness (San Pedro, Baker, & Rodrigo, 2011). We take advantage of BKT's extensibility within the current paper, modifying BKT as follows.

Classic BKT assumes that each problem or problem step encountered offers an opportunity to learn that skill – P(T). However, as discussed above, within many online learning systems learning can be expected to occur primarily between sets of problems. In many of these contexts, feedback is not given until a set of problems is completed, and instruction is offered wholly through videos or other conceptual content between problem sets (sets of problems given together, without any intervening instructional activities).

BKT-PL modifies BKT to take this different situation into account. In BKT-PL, we use the exact same algorithms as in classic BKT, shown in Figure 1, but we change the situation in which one equation is applied. Specifically, in BKT-PL the third equation of Figure 1 – the equation adjusting knowledge based on P(T) -- is applied when a problem set is completed rather than when a problem is completed. In other words, BKT-PL assumes that students learn between problem sets, from conceptual instruction, rather than through completing problems. This shift in how BKT is applied allows us to explicitly model – and investigate – the differences in learning rates associated with the conceptual content associated with different skills.

Additionally, when assessment and instruction are separated, one can assume that the problems included will be of differing difficulty levels; to address this, BKT-PL adopts a separate guess and slip parameter for each problem, following other work to assess problem difficulty within BKT (Pardos & Heffernan, 2011).

We compare the performance of this modified algorithm, entitled BKT-PeriodicLearning (BKT-PL), to both classic BKT and the second most-common student knowledge modeling algorithm, PFA (Performance Factors Analysis; Pavlik et al., 2009), which is based on logistic regression and estimates the degree of improvement associated with both correct and incorrect answers. We do not include recent algorithms which do not attempt to make interpretable estimations of parameters, as they cannot be used to determine which content is more and less effective. For instance, DKT (Deep Knowledge Tracing; Khajah et al., 2016) is not investigated, as it performs comparably to BKT variants and lacks the interpretability that the other approaches investigated afford.

1.2 Learning System

We study this system in the context of the Alef NextGen learning environment, a holistic education technology system. Alef NextGen is now in use by students in public sector schools in Abu Dhabi, in the United Arab Emirates, with students in New York City charter schools slated to begin use of the system in Fall 2018. Whereas many adaptive learning environments focus solely on practice of procedural skill, Alef NextGen alternates between conceptual instruction (in the form of multimedia, videos, and texts) and opportunities to practice the skills being learned, shown in Figures 2 and 3. These opportunities to practice the skill take the form of a set of problems to complete, referred to as a problem set. In order to promote greater reflection on the material being practiced, Alef NextGen utilizes delayed feedback, where students complete every item in the problem set and then receive feedback on each of those items, detailing whether and how each item is incorrect. Items within Alef NextGen can receive partial credit as well as full credit.

While students work with Alef NextGen, teachers receive real-time data. As students work through any given Alef lesson there are around 15 data collection points that provide the teacher with real-time data to group and remediate students during class time (cf. Miller et al., 2015). Periodic change management sessions are also held by Alef to assist teachers in building a rapid feedback loop between the data from the platform and their instructional strategies. Teachers are also supported in this through the provision of a series of offline experiential learning kits, usually consisting of manipulatives, simulations, and other hands-on activities.

Alef NextGen currently contains content on 402 topics for 6^{th}-8^{th} grades, spanning six core subjects ranging from Social Studies to Science to the Arabic language, taught in both English and Arabic. Within this paper, we focus on content from 49 topics in middle school mathematics, focusing on mathematics because the item/skill mapping is most straightforward for this content.

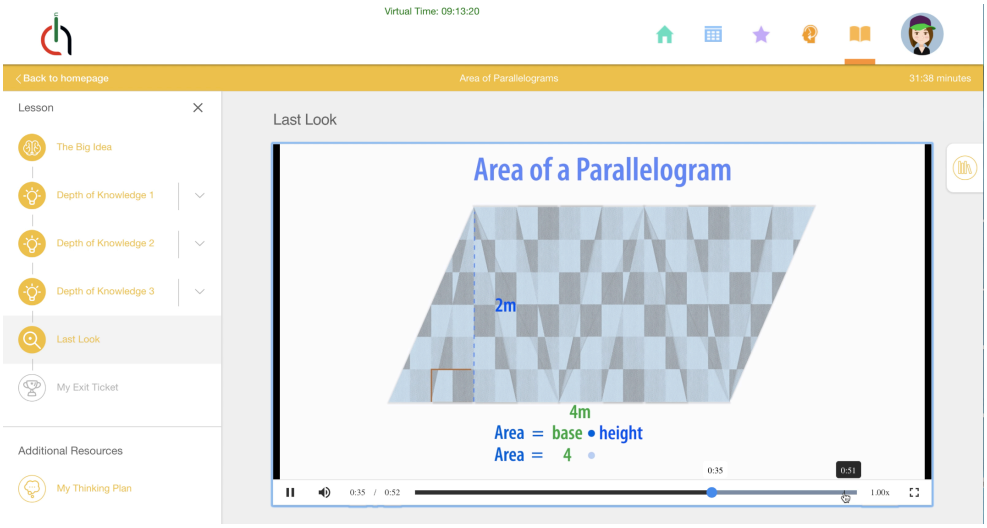

Figure 2. Alef NextGen video instruction

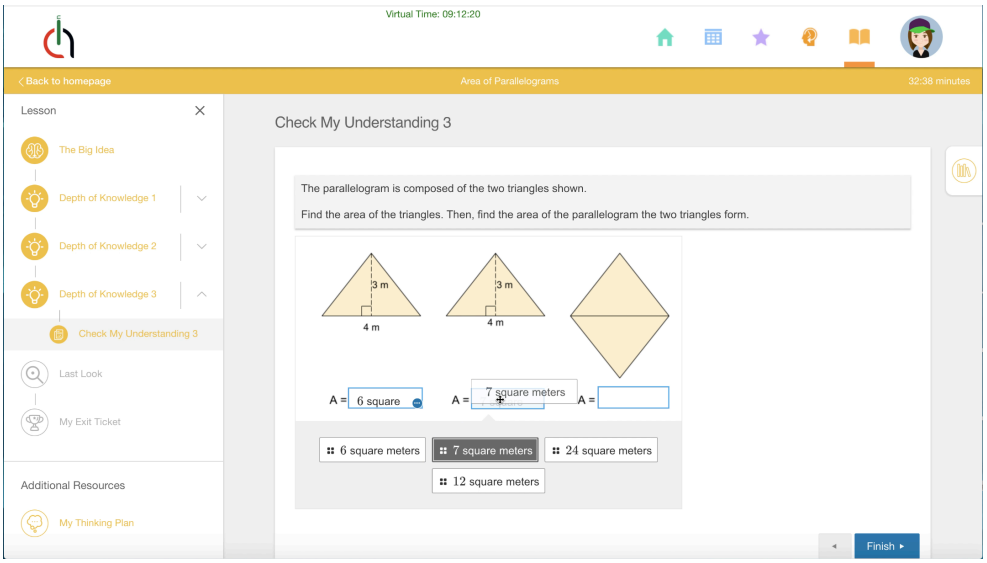

Figure 3. Alef NextGen assessment of problem-solving skill

These several differences in design suggests that existing student knowledge frameworks may be less effective for Alef NextGen than an approach tailored to Alef NextGen. In addition, by explicitly modeling the learning that takes place within the video and conceptual instruction, we may be ab0le to understand which conceptual instruction is most effective and least effective at promoting learning. Alef's content is built on the principle that a student should receive multiple opportunities to encounter the same concepts, experiencing the content, concepts, and skills in multiple fashions. As such, it may be that some ways content is presented will be more effective than others, for a specific content area. Determining which content is more and less effective for specific content can foster the iterative enhancement of Alef NextGen's video and other conceptual content. Across Alef's entire body of content, investigating these issues has the potential to generate broadly usable knowledge about learning in this type of learning activity.

2. Methods

We obtained data from 298 students, who completed a total of 45, 135 assessment activities within Alef NextGen's content for 6th-grade mathematics, aligned with United Arab Emirates Ministry of Education standards.

We then compared the performance of these algorithms at predicting next-problem student correctness, within existing data. For all approaches, we built the models on 70% of the students and tested on the other 30% of the students. We evaluated models in terms of AUC ROC (also referred to as A' or Wilcoxon) (Hanley & McNeil, 1982). AUC ROC represents the probability that the algorithm can differentiate between a student who is about to provide a correct answer, and a student who is about to provide an incorrect answer.

We built and compared three model variants within our data: Classic BKT, BKT-PL, and PFA, each discussed above.

We then determined whether the models are statistically significantly different in their ability to predict next-problem student correctness, using the method in (Baker, 2018), where the AUC ROC is computed for each student, compared statistically within each student using Hanley & McNeil's (1982) method – e.g. treating AUC ROC as Wilxocon – and then using Stouffer's method to integrate across students.

3. Results

Classic BKT performs relatively better than expected, achieving AUC ROC = 0.677, a level of performance that would typically be considered acceptable even if the learning system was closer to the system design BKT was originally envisioned for. It is theorized by Beck & Xiong (2013) that even in that context, performance should not be expected to be above AUC ROC =0.75.

BKT-PL performs better than Classic BKT, though only to a modest degree, achieving AUC ROC = 0.699. PFA performs slightly worse than Classic BKT and BKT-PL, achieving AUC ROC = 0.665.

Though the differences in AUC ROC value are relatively modest, BKT-PL performs statistically significantly better than Classic BKT, Stouffer's Z = 12.90, $p<0.001$, computed on 78 students (the test set). BKT-PL also performs statistically significantly better than PFA, Stouffer's Z = 25.72, $p<0.001$, computed on 78 students.

3.1 Use of BKT-PL

One advantage to using BKT-PL is that we can identify some of the most effective and least effective content. Learning rates, according to BKT-PL, were high overall, with an average P(T) = 0.182, indicating that students mastered the system's skills through accessing Alef NextGen's video/conceptual content 18.2% of the time. However, there is still room for improvement. The least effective content are candidates to be iteratively improved. The most effective content are cases to examine to understand why they are so effective, and learn lessons for the improvement of all the system content.

In identifying the most effective content, it is important to consider both P(T) and P(L0). Very high P(T) along with very high P(L0) may indicate very easy content. By contrast, lower P(L0) along with high P(T) indicates skills that are challenging to learners but are quickly learned with the content. In doing so, it is important to only consider content with sufficiently high frequency for estimates to be valid. Based on Slater et al. (under review)'s recommendations, we only consider cases where at least 25 students completed the material. In practice, every skill in the table below (see Table 1) has at least 150 cases.

Some of the skills with content that matches that profile – low P(L0), high P(T) – include:

Table 1
Content with the Highest Learning Rate among Skills with Low Initial Knowledge

Skill	Item-Set Preceding highly effective content	P(L0)	P(T)
Least Common Multiple	3rd set	0.031	0.989
Identifying Fractions	3rd set	0.026	0.989

Simplifying Fractions	3rd set	0.015	0.989
Estimate Products of Multi-Digit Numbers	3rd set	0.01	0.989
Multiply Decimals by Whole Numbers	3rd set	0.226	0.989
Percents Greater than 100% and Less than 1%	Initial set	0.002	0.451
Estimate Products of Fractions	Initial set	0.005	0.441
Dividing Fractions by Fractions	Initial set	0.001	0.431
Convert Customary Measurement Units (Capacity Length, and Weight)	Initial set	0.059	0.401
Least Common Multiple	Initial Set	0.031	0.382

Note that the highly-effective content comes both from the first set of remedial content (following the initial set of items) and the 3rd set of remedial content (following the 3rd set of items given to the student). This indicates that many students are still learning even after receiving multiple item sets and remedial materials, but that higher learning rates are not simply an artifact of the amounts of content received.

In identifying the least effective content, we again want to consider skills and concepts which are not initially known, and which are learned – e.g. low P(L0) and low P(T). Again, it is important to only consider content with sufficiently high frequency for estimates to be valid. Here we have an interesting surprise – some of the skills which have some of the most effective content also have some of the least effective content. These skills have at least one piece of content where the learning rate is at the minimum level of 0.001: Converting Customary Measurement, Percents Greater than 100% and Less than 1%, and Dividing Fractions by Fractions. In other words, these skills both have content associated with very high learning and content associated with very poor learning. While in some cases this may reflect a small number of students who struggle after all other students have mastered a skill – perhaps due to missing prerequisite knowledge or other factors – in other cases, this difference in learning rates may indicate differences in content effectiveness that are worth understanding and using to fix less effective content. Investigating the differences between the most effective and least effective content therefore represents an important area of future work.

4. Conclusions

In this paper, we propose a new model, BKT-PL (PeriodicLearning), which assumes that learning of content occurs periodically during the learning and assessment process, typically when video or other instructional materials are reached. This model is a closer match to the design of conceptual learning environments and Massive Online Open Courses than previous versions of Bayesian Knowledge Tracing (BKT) and other widely-used student knowledge modeling algorithms. We apply this model to data from a learning environment, Alef NextGen, where the design is for learning to be periodic, primarily driven by video content. We show that this model performs better on new data than Classic BKT or PFA, two popular learning algorithms. The differences are statistically significant, though relatively modest in size.

We also show that BKT-PL can be used to determine which content is most and least effective at promoting learning, and find that some skills have both highly effective and less effective content. One limitation to interpretation is that the order of content is not currently randomized. As such, a piece of content with a low learning rate seen after multiple pieces of highly effective content may simply be catching only a small number of students who are missing prerequisites for the current material. However, a piece of content with a low learning rate followed by content with a high learning rate can be more confidently inferred to be less effective and in need of revision.

These findings can be used to understand the properties of more effective content. Past work, for instance, has identified attributes associated with differences in learning rates for different content (e.g. Slater et al., 2016). Even without doing so, however, it is now possible to identify which content is less effective and send it to a content development team for iterative revision. Following these paths has the potential to take a system which already has substantial amounts of effective content and make it even better.

Acknowledgements

We would like to thank Juan Miguel Andres for helpful suggestions on data pre-processing, Christine Whitlock for help in document preparation, and Dr. Saleh Al Hashemi, Dr. Tahir Khan, Fahim Kundi, and Sarath Chandran for their roles in making the research possible.

References

Anderson, J. R., Conrad, F. G., & Corbett, A. T. (1989). Skill acquisition and the LISP tutor. *Cognitive Science, 13*(4), 467-505.

Baker, R.S. (2018) *Big Data and Education: 4th Edition.* EdX/University of Pennsylvania.

Baker, R. S. J. d., Goldstein, A. B., & Heffernan, N. T. (2011). Detecting learning moment-by-moment. *International Journal of Artificial Intelligence in Education*, *21*(1-2), 5-25.

Beck, J., & Xiong, X. (2013). Limits to accuracy: how well can we do at student modeling?. In S.K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining*, 4-11. Memphis, TN: International Educational Data Mining Society.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *UserModeling and User-Adapted Interaction, 4*(4), 253-278.

Desmarais, M. C., & Baker, R. S. J. d. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction, 22*(1-2), 9-38.

FSG. (2015). Learning gets personal: How Idaho students and teachers are embracing personalized learning through Khan Academy. Retrieved from https://s3.amazonaws.com/KA-share/impact/learning-gets-personal.pdf

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36.

Heffernan, N. T., Ostrow, K. S., Kelly, K., Selent, D., Van Inwegen, E. G., Xiong, X., & Williams, J. J. (2016). The future of adaptive learning: Does the crowd hold the key? *International Journal of Artificial Intelligence in Education, 26*(2), 615-644.

Khajah, M., Lindsey, R. V., & Mozer, M. C. (2016). How deep is knowledge tracing?. In T. Barnes, M. Chi, & M. Feng (Eds.), *Proceedings of the Ninth International Conference on Educational Data Mining,* (pp. 94-101).

Koedinger, K. R., & Corbett, A. (2006). Cognitive tutors: Technology brining learning sciences to the classroom. In R. K. Sawyer (Ed.), *The Cambridge Handbook of: The Learning Sciences* (pp. 61-77). New York, NY, US: Cambridge University Press.

Koedinger, K. R., Kim, J., Jia, J.Z., McLaughlin, E. A., & Bier, N.L. (2015). Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In *Proceedings of the second (2015) ACM conference on Learning @ Scale,* (pp. 111-120).

Miller, W.L., Baker, R., Labrum, M., Petsche, K., Liu, Y-H., & Wagner, A. (2015) Automated Detection of Proactive Remediation by Teachers in Reasoning Mind Classrooms. In *Proceedings of the 5th International Learning Analytics and Knowledge Conference* (pp. 290-294).

Pardos, Z. A., Bergner, Y., Seaton, D. T., & Pritchard, D. E. (2013). Adapting Bayesian knowledge tracing to a massive open onlince course in edX. In D'Mello, S. K., Calvo, R. A., Olney, A. (Eds.) *Proceedings of the 6th International Conference on Educational Data Mining (EDM)*. 137-144. Memphis, TN: International Educational Data Mining Society.

Pardos, Z. A., & Heffernan, N. T. (2011). KT-IDEM: Introducing item difficulty to the knowledge tracing model. In Konstan, J. A., Conejo, R., Marzo, J. L., Oliver N. (Eds.) *User Modeling, Adaptation, and Personalization. UMAP 2011. Lecture Notes in Computer Science, 6787*. 243-254. Berlin, Heidelberg: Springer.

Pavlik, P. I., Cen, H., & Koedinger, K. R. (2009). Performance factors analysis—A new alternative to knowledge tracing. In *Proceedings of the 2009 Conference on Artificial Intelligence in Education:*

Building Learning Systems that Care: From Knowledge Representation to Affective Modeling (pp. 531-538). IOS Press.

San Pedro, M. O. Z., Baker, R. S. J. d., & Mercedes, M. M. T. (2014). Carelessness and affect in an intelligent tutoring system for mathematics. *International Journal of Artificial Intelligence in Education, 24*(2). 189-210.

Slater, S., & Baker, R. S. (under review). Degree of error in Bayesian knowledge tracing estimates from differences in sample sizes. Manuscript under review.

Slater. S., Ocumpaugh, J., Baker, R., Scupelli, P., Inventado, P. S., & Heffernan, N. (2016). Semantic features of math problems: Relationships to student learning and engagement. In T. Barnes, M. Chi, & M. Feng (Eds.), *Proceedings of the 9th International Conference on Educational Data Mining,* (pp 223-230).

Wilson, K. H., Karklin, Y., Han, B., & Ekanadham, C. (2016). Back to the basics: Bayesian extensions of IRT outperform neural networks for proficiency estimation. *Proceedings of Educational Data Mining 2016 (EDM 2016)*, (pp. 539-544).