Contextual Derivation of Stable BKT parameters for Analyzing Content Efficacy

Deepak Agarwal
Educational Initiatives
Ahmedabad, IN
+91 9731406003
deepak.agarwal@ei-india.com

Nishant Babel
Educational Initiatives
Ahmedabad, IN
+91 7022281740
nishant.babel@ei-india.com

Ryan S. Baker
3700 Walnut St.
Philadelphia, PA
+1 (412) 983-3619
ryanshaunbaker@gmail.com

ABSTRACT
One of the key benefits that Bayesian Knowledge Tracing (BKT) offers compared to many competing student modelling paradigms is that its parameters are meaningful and interpretable. These parameters have been used to answer basic research questions and identify content in need of iterative improvement (due to, for instance, low learning or high slip rates). However, a core challenge to the interpretation of BKT parameters is that several combinations of BKT parameters can often fit the same data comparably well. Even if, as some have argued, BKT is not truly non-identifiable, in practice highly different parameters with comparable goodness are often found using modern BKT fitting packages. These parameter sets can have highly divergent values for guess and slip. Several approaches have been proposed but none of those have yet led to fully stable and trustworthy parameter estimates. In this work, we propose a new iterative method based on contextual guess and slip estimation that converges to stable estimates for skill-level guess and slip parameters. This method alternates between calculating contextual estimates of guess and slip and estimating skill-level parameters, iterating until convergence. Thus, it produces a more stable set of parameters that can be more confidently used in analyzing content efficacy.

Keywords
Bayesian Knowledge Tracing, Contextual Guess Slip, Content Efficacy, Brute Force BKT Model, EDM

1. INTRODUCTION
The process of developing an intelligent tutoring system (ITS) is an iterative one, and content frequently needs revision to reach its desired effectiveness for students [1]. In addition, even as intelligent tutors have become more widespread, the quality of the content present in them has often become more varied, with the advent of approaches such as crowd-sourcing for generating large amounts of content quickly [2]. As such, improving the quality of content that the students are exposed to is one of the significant aspects of the development of ITS. One approach to achieving this is to put in place a framework for automatically reviewing content, identifying/flagging content that does not meet the desired objectives. In this paper, we discuss our efforts to create such a system within the context of Mindspark*, an ITS software being used by over 80,000 students in India.

The cornerstone of our efforts is discovering skills which have unexpected negative properties, specifically very low learning rates, or very high rates of guess and slip within the Bayesian Knowledge Tracing paradigm (Corbett & Anderson, 1995). Bayesian Knowledge Tracing is a highly-cited paradigm for student modeling and is used in a wide range of real-world adaptive learning systems. Although recent evidence suggests that extensions to BKT and competing paradigms may in some cases achieve better prediction of immediate correctness [3], BKT remains a high-quality, highly interpretable paradigm for modeling student latent knowledge as well as meaningful attributes of individual skills. However, one challenge to interpreting BKT parameters is that different sets of parameters can fit the data comparably well [4]. Although recent articles have argued that BKT is not truly non-identifiable [5], nonetheless contemporary packages for choosing BKT parameters regularly produce very different parameter values with comparable fits. Other researchers have noted the problem of unstable parameters; however, these approaches have tended to assume skills have similar parameters to each other [6], [7]. These assumptions may lead to more plausible parameters in general but may be unhelpful for identifying skills whose guess, slip, or learning rates are genuinely problematic.

Hence, in this paper we propose a new iterative approach for stabilizing the parameter values of BKT that leverages additional information about student performance. Previous work proposed contextually estimating guess and slip in all cases with situational information [8]; this approach produced unstable improvements in model goodness, however, with positive impacts in some data sets and negative impacts in other data sets. This paper instead uses iterative contextual estimation of guess and slip to help select guess and slip parameters for traditional Bayesian Knowledge Tracing – i.e. the final model is non-contextual. By using additional information to derive better estimates of guess and slip, we can be more confident about our model parameters, and more confident about our ability to use these parameters in driving quality improvement.

The subsequent sections explain the dataset used, the conceptualization of the iterative parameter estimation approach, the results obtained and how we validated the approach. We conclude with a discussion of how this approach will be leveraged, going forward, to analyze content efficacy in Mindspark.

2. DATA DESCRIPTION AND APPROACH

2.1 Data sets
In order to evaluate our approach to estimate BKT parameters, we considered a simulated data set and a genuine data set from the Mindspark platform.

*https://www.mindspark.in/

Simulated data:

Student responses were simulated for four different skills by assuming four different set of BKT model parameters values – L_0, G, S, T – and then estimating the probabilities of student responses using the BKT paradigm. Each skill had 2,000 users, with four attempts for every user. Each of the four skills came from a different [L_0, G, S, T] combination: [0.6, 0.3, 0.05, 0.25], [0.6, 0.3, 0.05, 0.02], [0.6, 0.05, 0.25, 0.25] and [0.6, 0.05, 0.25, 0.02] respectively. Data consisted of 0 and 1 for incorrect and correct response by users on each attempt. The data was simulated by calculating the likelihood of knowledge and a correct response based on the BKT model and each simulated user's response history.

Real student log data from Mindspark Math:

Actual student response data was also extracted from the Mindspark log data. Mindspark is an adaptive-learning program for Math and English, developed by Educational Initiatives (EI). Mindspark Math currently has 80,000 users, primarily from private schools, in grades 1 to 9, across India. For our purpose, data from the 'Revision Module' in Mindspark was taken. The Revision Module is a 30-minute session that gives students questions from topics selected by the teacher. It is intended to help students learn concepts for which that student had relatively low performance in regular modules. The reason for selecting the Revision Module for this exercise was that this module usually has multiple attempts per skill for each student. A ′attempt′ here means an opportunity provided to a student to apply the skill in order to solve a question. Hence when a student has multiple attempts on a particular skill, the student is presented a set of questions testing the same skill and each response is scored as correct/incorrect. Data of 1,032 users across a total of 5,200 attempts was extracted for six different skills. We limited the data set to students who had at least three attempts and capped the number of attempts at five per user in order to avoid focusing the data set on students who struggled to reach mastery.

Skill	Grade	Learning Objective
NTH001_8	5	Determining least multiple for a number out of a given set of numbers
WNC034_7	5	Estimating a number to the nearest hundred
NTH021_8	5	Writing factorizations of a number using the factor tree of the number
WNC059_12	5	Estimating a number to the nearest thousand
WNO033_10	5	Adding a 3-digit number to another 3-digit number vertically
WNO049_10	5	Writing quotient and remainder given dividend and divisor

Table 1. Skills used from the Revision Module of Mindspark

2.2 Approach

We derived contextually-inspired parameters for Bayesian Knowledge Tracing as follows: We start by obtaining initial parameter values for each skill using the common Brute Force grid search method [9] and classical BKT paradigm. This set of parameter values are used as input to the Contextual Guess Slip model [8] to estimate the contextual probability of guess and slip for each student attempt. The contextual probability of guess and slip is derived from the likelihood of a student knowing the skill at a specific attempt, which in turn is estimated based on the student's performance on the next two attempts on that skill. The formulas from the original contextual guess/slip model [8] were used to calculate $P(L_{n-1})$ which represents each student's knowledge state after (n-1) th attempt. The formulas take into account a student's subsequent two attempts (n and n+1 response data) to calculate $P(L_{n-1})$.

$$P(L_{n-1} \mid A_{n,n+1}) = P(A_{n,n+1} \mid L_{n-1}) * P(L_{n-1}) / P(A_{n,n+1}) \quad (1)$$

$$P(A_{n,n+1}) = P(L_{n-1}) * P(A_{n,n+1} \mid L_{n-1}) + (1 - P(L_{n-1})) * P(A_{n,n+1} \mid \sim L_{n-1}) \quad (2)$$

The probability of the actions at time n and n+1, in the case that the student knew the skill at time n (L_{n-1}), is a function of the probability that the student guessed or slipped at each opportunity to practice the skill. C denotes a correct action; ~C denotes an incorrect action.

$$P(A_{n,n+1} = C,C \mid L_{n-1}) = P(\sim S)^2 \quad (3)$$

$$P(A_{n,n+1} = C,\sim C \mid L_{n-1}) = P(\sim S)*P(S) \quad (4)$$

$$P(A_{n,n+1} = \sim C,C \mid L_{n-1}) = P(S)*P(\sim S) \quad (5)$$

$$P(A_{n,n+1} = \sim C,\sim C \mid L_{n-1}) = P(S)^2 \quad (6)$$

The probability of the actions at time n and n+1, in the case that the student did not know the skill at time n ($\sim L_{n-1}$), is as below:

$$P(A_{n,n+1} = C,C \mid \sim L_{n-1}) = P(G)*P(\sim T)*P(G) + P(G)*P(T)*P(\sim S) \quad (7)$$

$$P(A_{n,n+1} = C,\sim C \mid \sim L_{n-1}) = P(G)*P(\sim T)*P(\sim G) + P(G)*P(T)*P(S) \quad (8)$$

$$P(A_{n,n+1} = \sim C,C \mid \sim L_{n-1}) = P(\sim G)*P(T)*P(\sim S) + P(\sim G)*P(\sim T)*P(G) \quad (9)$$

$$P(A_{n,n+1} = \sim C,\sim C \mid \sim L_{n-1}) = P(\sim G)*P(T)*P(S) + P(\sim G)*P(\sim T)*P(\sim G) \quad (10)$$

After calculating $P(L_{n-1})$, the contextual probabilities of guess and slip at nth attempt was assigned as:

$$P(G'_n) = 1 - P(L_{n-1}) \quad (11)$$

$$P(S'_n) = P(L_{n-1}) \quad (12)$$

The probabilities obtained are at a student attempt level. To obtain skill level parameters, we aggregate these values across all the attempts for a given skill as below:

$$G = \Sigma P(G'_n \mid C) / \Sigma P(G'_n) \quad (13)$$

$$S = \Sigma P(S'_n \mid \sim C) / \Sigma P(S'_n) \quad (14)$$

where $P(G'_n)$ and $P(S'_n)$ are taken from equations 11 and 12 respectively.

In other words, to obtain the guess parameter, we take the ratio of the sum of the $P(G'_n)$ values for the attempts where the response by the student was correct and the sum of the $P(G'_n)$ values across all the attempts for the skill. Similarly, to obtain the slip parameter, we take the ratio of sum of the $P(S'_n)$ values for attempts where the response was incorrect and the sum of $P(S'_n)$ values across all the attempts for the skill.

Subsequently, if the skill level guess and slip estimates from the contextual model do not agree with the guess and slip estimates obtained from the Brute Force grid search method originally, then it means that the BKT parameters are not stable. These parameters can be refined by using contextual estimates of guess and slip as input to Brute Force grid search algorithm and iterating this process until the skill level G and S values from the two approaches match with each other. Here note that each iteration is performed on the entire dataset (all the attempts of the students) which means the dataset does not change from one iteration to another. The flowchart below summarizes the whole process:

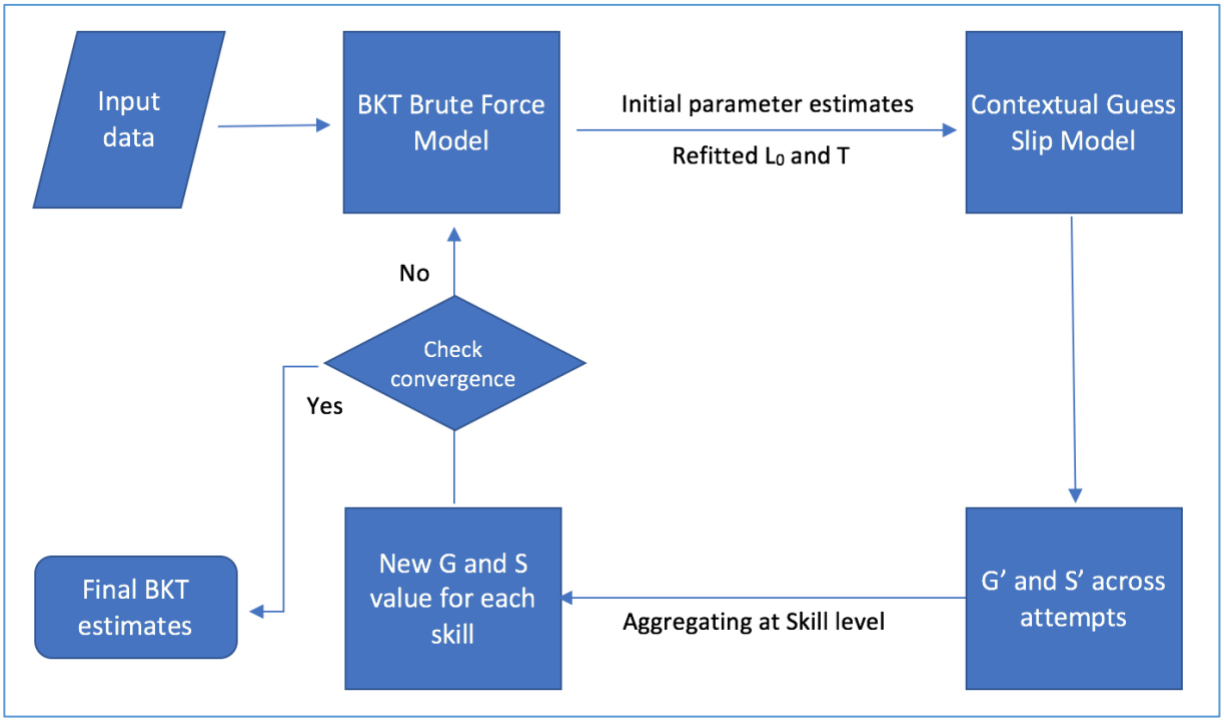

Fig. 1. BKT with Contextual Guess Slip Flow

3. VALIDATION AND RESULTS

3.1 Validations carried out on simulated data

The purpose of using simulated data was to confirm that the skill level G and S values calculated through the contextual model match the original G and S values used to generate the simulated data. The calculated values from the proposed approach achieved a match within 1% error margin to the parameter values used for simulating the data for all four cases.

Table 2. Original vs Calculated G and S values for each skill

Data set	Original G	Original S	Calculated G	Calculated S
Skill 1	0.3000	0.0500	0.2998	0.0499
Skill 2	0.3000	0.0500	0.3001	0.0500
Skill 3	0.0500	0.2500	0.0500	0.2501
Skill 4	0..0500	0.2500	0.0500	0.2501

We also used the simulated data to check if the iterative model achieves convergence over time and results in a stable set of parameter values. For this purpose, we used arbitrary parameter values [L_0=50%, G=15%, S=15%, T=10%] for first iteration for all four simulated skills instead of estimating the parameters from the Brute Force BKT model. We observed that L_0, G, S, T values started converging after a reasonable number of iterations for all four cases and the output matched the original parameter set used to simulate the data. Fig. 2 shows the convergence for all four simulated skills. We have also shown the trend in RMSE values over the iterations for all four skills in Fig 4. Since we had started with arbitrary parameter values, RMSE is quite high in the first iteration but decreases continuously to achieve the minimum value over multiple iterations.

3.2 Validation carried out on Mindspark data

It is not possible to determine whether the proposed approach reaches "true" parameter values for real-world data, because it is unknown what those true parameter values would be (and, indeed, we know that BKT is an imperfect model of the real world). However, we are still able to validate how well the proposed approach converges when applied to real data, where the noise may be different in kind than the noise generated by BKT for the simulated data. As Fig. 3 shows below, the curve for all four skills starts flattening out after a tractable number of iterations, exhibiting convergence in the data.

We also show here that the model achieves convergence without significantly increasing the original RMSE achieved through the Brute Force BKT model which indicates that the model fit does not worsen over the iterations. The change in the RMSE values was observed to be under 0.002 across all six skills and the trend has been shown in Fig 4.

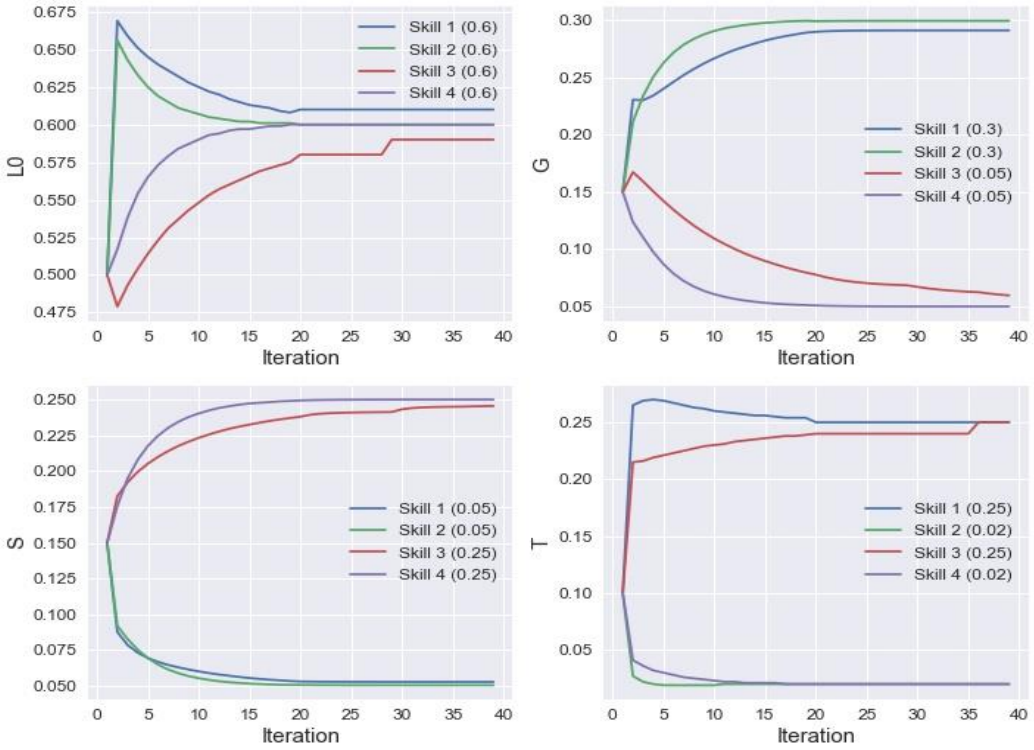

Fig. 2. Contextual BKT approach on the simulated data with four skills across 40 iterations. The values in the parentheses represent the actual values used for simulation

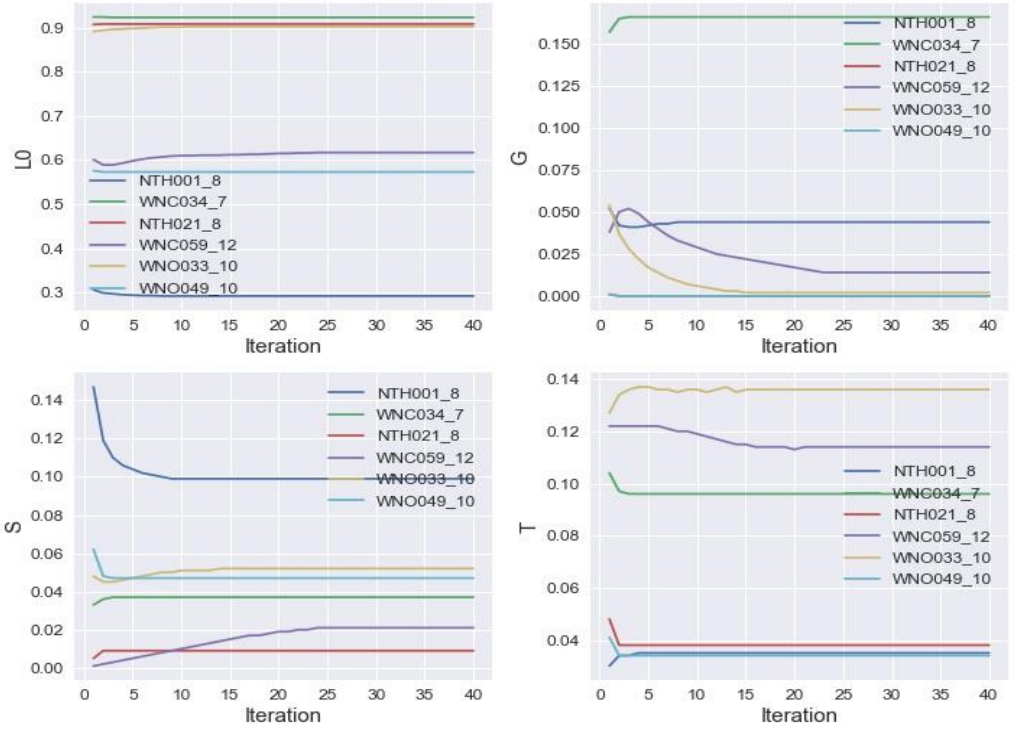

Fig. 3. Contextual BKT approach on Mindspark data with six skills across 40 iterations

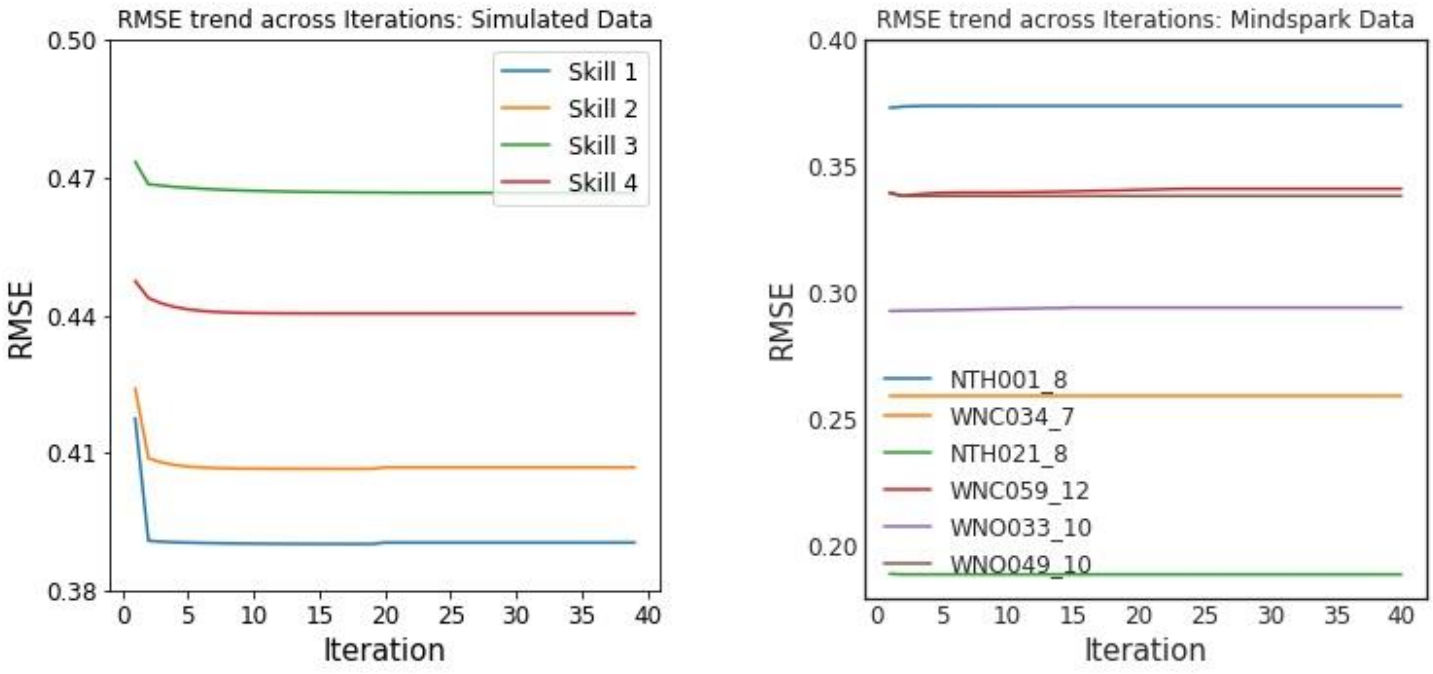

Fig. 4. RMSE values of the Contextual BKT model for simulated and real data across 40 iterations. As can be observed, the error metric changes very minimally across iterations.

4. CONCLUSION AND FUTURE WORK

In this paper, we discuss a new, iterative approach to fitting BKT parameters, involving iteration between fitting skill-level parameters for guess and slip, and contextually estimating guess and slip within each problem attempt. This approach converges across iterations to a stable and single set of parameters which have a principled justification for their selection. As such, we can then have higher confidence about interpreting and using these parameters for commenting on content efficacy. The skill level BKT parameters enable us to evaluate content on multiple dimensions: (a) grade appropriateness, (b) learning rate, and (c) quality of the content as indicated by low or high guess and slip values. For content to be effective in a given context, it should have BKT parameters within a desired range. In future work, the desired range of values will be determined through multiple approaches including analyzing parameter distributions to set up heuristic rules, anomaly detection, and discussions with our pedagogical experts.

If the parameter values for a skill is outside of those permissible ranges, it would indicate that the content does not meet the quality/effectiveness standard. For example, for a piece of content to be grade appropriate, L_0 should likely be between 25% to 85%. Any content which has L_0 above 85% may not lead to substantial improvement in student learning, as most of the students already know it. By contrast, any content which has L_0 below 25% is also not appropriate as students may not know the pre-requisite skills to learn the content.

Our next step is therefore to establish thresholds (the desired range of values) for each parameter to develop filters which will automatically identify ineffective content and bring it to the attention of the content developers.

An added advantage of screening content using BKT parameters is that we can also provide auto-generated guidance on what the issue might be with the content rather than just highlighting that the content needs improvement. Apart from being a tool for ITS / content developers in identifying lower quality content for revision, this approach has wider application in the domain of Bayesian Knowledge Tracing as it provides a means to capture a single set of skill parameters and have a justification for preferring this set of values to others with comparable fit.

5. REFERENCES

[1] Koedinger, K. R., Stamper, J. C., McLaughlin, E. A., & Nixon, T. (2013). Using data-driven discovery of better student models to improve student learning. *Proc. International Conference on Artificial Intelligence in Education*, 421-430.

[2] 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

[3] Khajah M., Lindsey, R.V., Mozer, M. (2016) How deep is knowledge tracing? *Proc. Int'l. Conf. on Educational Data Mining*, 94-101

[4] Beck, J. E., & Chang, K. M. (2007, July). Identifiability: A fundamental problem of student modeling. In *International Conference on User Modeling* (pp. 137-146). Springer, Berlin, Heidelberg.

[5] Doroudi S., Brunskill E. (2017). The Misidentified Identifiability Problem of Bayesian Knowledge Tracing. *Proc. Int'l. Conf. on Educational Data Mining.*

[6] Rai, D., Gong, Y., & Beck, J. E. (2009) Using Dirichlet priors to improve model parameter plausibility. *Proc. Int'l. Conf. on Educational Data Mining*, 141-150.

[7] Ritter, S., Harris, T. K., Nixon, T., Dickison, D., Murray, R. C., & Towle, B. (2009). Reducing the Knowledge Tracing Space. *Proc. Int'l. Conf. on Educational Data Mining.*

[8] Baker R.S.J.d., Corbett A. T., Aleven, V. (2008) More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. *Proc. Int'l. Conf. on Intelligent Tutoring Systems.*

[9] Baker, R.S.J.d., Corbett, A.T., Gowda, S.M., Wagner, A.Z., et al. (2010) Contextual Slip and Prediction of Student Performance After Use of an Intelligent Tutor. *Proceedings of the 18th Annual Conference on User Modeling, Adaptation, and Personalization*, 52-63.