Evaluating The Relevance of Educational Videos using BKT and Big Data

Zachary MacHardy
UC Berkeley
354 Hearst Memorial Mining Building
Berkeley, CA 94720
zmmachar@cs.berkeley.edu

Zachary A. Pardos
UC Berkeley
4641 Tolman Hall
Berkeley, CA 94720
zp@berkeley.edu

ABSTRACT
Along with the advent of MOOCs and other online learning platforms such as Khan Academy, the role of online education has continued to grow in relation to that of traditional on-campus instruction. Rather than tackle the problem of evaluating large educational units such as entire online courses, this paper approaches a smaller problem: exploring a framework for evaluating more granular educational units, in this case, short educational videos. We have chosen to leverage an adaptation of traditional Bayesian Knowledge Tracing (BKT), intended to incorporate the usage of video content in addition to assessment activity. By exploring the change in predictive error when alternately including or omitting video activity, we suggest a metric for determining the relevance of videos to associated assessments. To validate our hypothesis and demonstrate the application of our proposed methods we use data obtained from both the popular Khan Academy website and two MOOCs offered by Stanford University in the summer of 2014.

Keywords
knowledge tracing, educational videos, instructional technology, bayesian inference, online education

1. INTRODUCTION
As the relative importance of MOOCs and other online learning platforms such as Khan Academy has increased, so has the importance of verifiably sound online pedagogy increased apace. While many of the lessons learned through a long history of research on the traditional classroom are applicable to the online environment, many indicators available during traditional instruction are not present for a designer of online material. In order to address the need for scalable and reproduceable evaluation, we hypothesize that by relating the use of materials and performance on subsequent assessment items, we can construct a metric to evaluate the relevance of those videos, without needing to resort to comparative studies.

To model student interactions with educational material and improvement over time, we have chosen to use an adaptation of Bayesian Knowledge Tracing (BKT), a technique developed and used with Intelligent Tutoring Systems (ITS) but which has been applied outside of that domain as well. We seek to incorporate behavior, such as video observation, which falls beyond the purview of attempting assessment items. We contrast this extended model with a simpler one excluding resource usage in order to discover whether videos contribute to model accuracy, and if some models benefit more than others.

Our ultimate goal is not to produce high predictive accuracy for the purposes of predicting students' latent knowledge, but rather to provide a quantitative framework for evaluating video resources. We set out first to prove that there is a reduction of predictive error when incorporating video resources into BKT analysis, in order to validate the inclusion of such observations. Second, we propose a metric based on a combination of both the delta in error between models using and eschewing video data and the learn rate associated with a particular video, in order to foreground both those which appear most relevant, as well as those which may need attention.

2. RELATED WORK
2.1 Bayesian Knowledge Tracing
Bayesian Knowledge Tracing [1] is used extensively in computer-assisted instruction environments, intended to approximate mastery learning. The model in its most basic form is defined by four parameters: $P(L_0)$, the prior probability that a student has mastered a particular KC, or knowledge component; $P(S)$, the probability a student who knows a concept will get an associated question wrong, or 'slip'; $P(G)$, the probability that a student who does not know a concept will correctly 'guess' the correct answer; and $P(T)$ the probability that a student who does not know a particular KC will learn it after a given observation. Through a process of Bayesian inference, an observed correct or incorrect response to an assessment item can be used to calculate a posterior probability that a student has mastered the KC. Using this posterior and $P(T)$ as described above, a new prior is calculated, accounting for the probability that the KC was learned between observations. This process is then repeated, using the updated estimate, for each subsequent observation.

We chose to use BKT as a modeling framework as it is well-studied and possesses relatively well understood properties, with parameters which are intuitively interpretable and therefore potentially actionable. Additional work has been done to extend this basic model of BKT to incorporate individualized parameters, based on factors depending both upon individual student properties (see e.g. [7], [2]), as well as properties of particular assessment items within a knowledge component [8].

Source	Total Events	Distinct KCs
Khan	353,202	176
Economics	689,709	94
Statistics	337,428	70

Table 1: Properties of the three sources

2.2 Online Course Resources

There has been a fair amount of research devoted to studying the efficacy of videos, forums, and other study aids offered in online educational contexts. Past work has typically focused on issues such as student attrition, student interaction, and building student-facing recommender systems. For example, Yang et al. described a framework for helping students sift through the the large volume of forum discussion posts in order to find content relevant to them [10]. Similar efforts have been made to provide recommendations for more general content, using methods such as social media analysis and reinforcement learning [5] [9].

Relative to the research on student perception and experience in the MOOC context, little attention has been paid to that of the instructor. That is not to say that such work has been absent. Guo et al. [3] and Kim et. al [4] offer guidance for the construction of videos used in MOOCs. Explorations of the application of Item Response theory in a MOOC environment [6] similarly offer instructors guidance in evaluating the efficacy of their assessments using traditional methods. Yousef et al. constructs an inventory of features, pedagogical and technological, which contribute to a sense of course quality. [11]. Yet there remains a relative paucity of research on the quantitative assessment of content outside of the scope of assessment items.

3. DATA

In order to demonstrate the generalizability of our results, we leveraged three sources of event log data. Two of our datasets were taken from Stanford Online courses run using the edX platform: 'Statistics and Medicine' and 'Principles of Economics.' The third was taken from the popular Khan Academy Website. See table 1 for details.

The data we obtained from Khan Academy contains observation events collected over about two years, from June 2012 to February 2014, while both edX courses were offered from June to September of 2014. Assessment items in Khan are categorized hierarchically as part of a larger 'exercise' representing a particular skill, and further as a member of a 'problem type,' describing the template used to generate a specific problem, while exercises from edX are categorized as individual problems. For the sake of simplicity we have chosen to consider each exercise as a separate knowledge component (KC) for the purposes of training BKT models.

For both the Khan and edX data, there was not an immediately available canonical mapping between videos and associated problems. By scanning the logs of learner activity and using a metric combining chronological proximity of use as well as frequency of associated observation, we produced a mapping between videos and their related KCs. Because our goal was not to produce a generative procedure for semantically associating log events, we chose our method to be sufficiently successful without introducing unnecessary complexity. However, this does introduce possible sources of error in terms of both overlooked and spuriously constructed mappings.

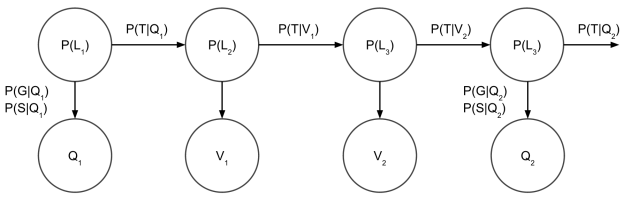

Figure 1: The Template-Videos Model

4. METHODS

Though the previous section describes the fundamentals of Bayesian Knowledge Tracing, we employ several extensions to the model. First, and for all models used in evaluation, we condition $P(G)$ and $P(S)$ for each observation on which specific problem template is observed, to model varying template difficulty. We will refer to this model as 'Standard BKT'.

Second, we similarly condition the transition probability $P(T)$ on the observed problem template, generating a second distinct but still video-free 'Template' model. We include this model for the Khan data for the sake of completeness, but note that there is only a single template for each edX problem in the data and thus the results of this extension are omitted for both the 'Statistics and Medicine' and 'Principles of Economics' cases

Third, we extend our model to incorporate video observations, conditioning $P(T)$ either on the specific template observed or the specific video, generating the 'Template Videos' model. The presence of a video observation functions similarly to that of a problem attempt, save that as there is no associated student response to be considered, a video is associated only with a unique $P(T)$. We simplify the 'Template Videos' into a fourth 'Template 1 Video' model, conditioning $P(T)$ only on the presence of either a video or a question, but not the specific identity of the resource observed.

All models were trained and evaluated using 5-fold cross validation. For each model above, one BKT model was trained for each of the knowledge components. For each model, for each fold, each of the KC models was randomly initialized and trained using Expectation Maximization (EM) algorithm to minimize the log likelihood of the observed events 25 times, with the maximally likely resulting model chosen for that model-fold-model tuple. The metric used to compare the four models is the root mean squared error (RMSE) taken across all five folds.

5. RESULTS AND DISCUSSION

Tables 2, 3, and 4 describe the results of running the data through the three analytical models. In each case, the 'Template Videos' and 'Template 1 Video' models tended to perform best, while the 'Template' model, using the Khan Academy

data, showed no significant difference from the baseline distribution. The significance test is performed across the distribution of RMSE across each of the KC models in each data-set.

Model	Mean RMSE	Significance
Pct. Correct	.4930	.0000*
Standard BKT	.3824	—-
Template	.3824	.9448
Template Videos	.3810	.0253*
Template 1 Video	.3811	.0061*

Table 2: Khan Academy

Model	Mean RMSE	Significance
Pct. Correct	.6243	.0000*
Standard BKT	.3824	—-
Template Videos	.3715	.0000*
Template 1 Video	.3716	.0000*

Table 3: Principles of Economics

Model	Mean RMSE	Significance
Pct. Correct	.5551	.0000*
Standard BKT	.3711	—-
Template Videos	.3638	.0000*
Template 1 Video	.3642	.0000*

Table 4: Statistics and Medicine

Though the tables reflect changes in RMSE aggregated over all KC models, not all models benefited evenly from the inclusion of video resources. Among the Khan data 77 of 193 KCs saw more then a trivial amount of reduction in error, while in Statistics and Medicine and Economics, the bulk of the improvement could be seen in 57 of the 94 and 44 out of 70 models, respectively. This asymmetry of improvement is an expected behavior of the system. Intuitively, in the case that a particular video resource is either not helpful or actively harmful to a student in solving a particular problem or set of problems, this would be reflected in the trained model as additional noise, leaving the overall RMSE unaffected at best.

Rather, the presence of a statistically significant, though perhaps small, decrease in predictive error in some models is indicative of the soundness of the hypothesis that considering video usage can offer useful information.

5.1 Highest and Lowest Performing Models

In order to gain an intuition for why some models were better described by the inclusion of resources, we chose to consider a selection of the best and worst performers from each data set under the 'Template-Videos' condition. By examining what properties might explain the performance of each model, we seek insight into what sort of videos appear to offer the greatest benefits to student performance.

For the highest performing models in the Khan data, the videos appeared highly relevant to their associated exercises, often demonstrating solutions in the Khan interface. For example, 'The Fundamental Theorem of Arithmetic,' explains the manipulation of a bespoke tool created for that particular exercise, showing the completion of a practice problem using that tool.

For the low performing Khan models the possible sources of error mirror the effects seen in the high performing cases. 'Scalar Matrix Multiplication' and 'Linear Inequalities', for example, present video explanation very differently than their related videos and involve customized input fields, which may have been a source of trouble.

Though the Principles of Economics and Statistics in Medicine edX courses are formatted very differently than the lessons of Khan academy, the distinctions between the best and worst models are similar. In both cases, the best videos in the data-set are, while less compellingly visually similar than the Khan examples, pointedly related to the subsequent assessments. Additionally, most of the associated assessments allowed students only one attempt, explaining the particularly strong reduction in error when including video information.

Perhaps most interesting is that one of the best predicted models is the ninth question on the final exam of the 'Statistics and Medicine' course. The content of this question is nearly identical to content of the video from a couple of weeks previous, 'Practice Interpreting Linear Regression Results.' It is therefore unsurprising to find that the video, while not explicitly grouped with the exam, is associated with a very strong learn parameter; students who sought out the video succeed significantly more often on the assessment.

Two of the videos related to the worst models in the Economics set, 'The Spending Allocation Model', and 'The Fed and the Money Supply' are both relatively long, each over fifteen minutes. Despite their length, each video dwells only briefly on the subject concerned in the assessment, spending most of their running time on other topics, with the pertinent sections easy to skip or miss. Another worst performer is one of the first videos in the course, associated with a quiz with nearly a 90% correctness rate.

Intuitively, an unhelpful video does not contribute to a predictive model, simply adding additional complexity and noise. By measuring which videos do and do not contribute constructively to predictive accuracy, it may be possible to detect which videos might be most appropriately suggested as helpful for a learner, and which need revision. In particular, such results could be useful to an instructor or course manager in navigating what to improve and what to keep when iterating on a course between offerings.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have demonstrated that the inclusion of video observations in a KT model can offer information relevant to predicting student behavior, not only in one data-set, but generalizably across multiple domains. Though the effect size is small, the statistically significant decrease in error under the 'Template 1 Video' and 'Template Videos' conditions across the three data-sets considered is an encouraging sign. It is indicative that there is information to be gleaned from a learner's use of video resources. Further,

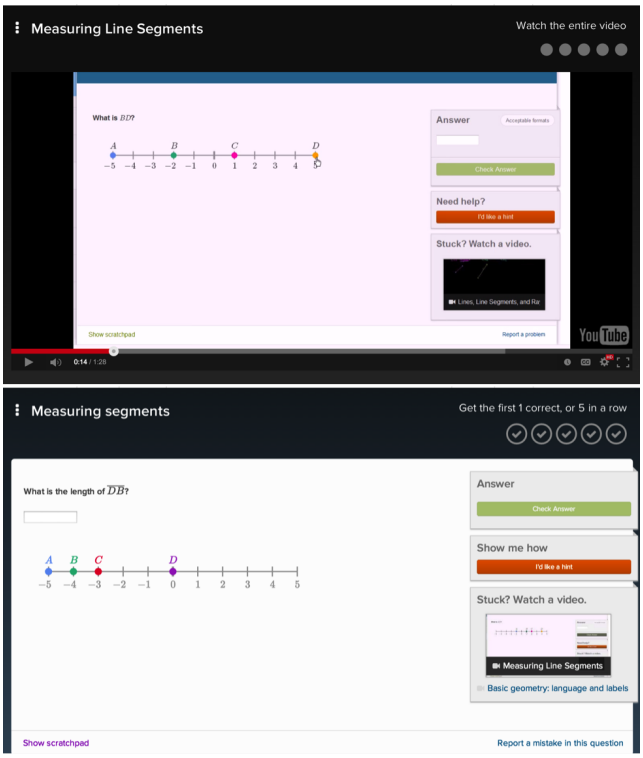

Figure 2: Videos from Khan Academy contributing maximally to model accuracy tended to closely mirror subsequent assessments

as suggested by our investigation of some of the superlative models, it is possible that the delta in error generated by a given model, coupled with the associated P(T) for a video within that model, could be a useful metric for evaluating video relevance.

One piece missing from this analysis is a canonical association of videos to exercises. Though we generated and used a set of associations, we may have lost information in the process. Another avenue worth pursuing is the possibility that some users would benefit strongly from video resources while others may not. To that end, it would be useful to examine potential reductions in error that might be made by individualizing parameters to each KC-Student pair.

An important caveat of this analysis is to note that our results do not speak to a general 'quality' of a video, and indeed that is perhaps beyond the scope of a quantitative analysis. A video rated poorly by our metrics need not necessarily be a bad video, merely unrelated or unhelpful for a subsequent assessment task. The importance of this particular property is a matter of educational policy, and thus beyond the scope of this paper. Our goal is not to supplant the role of instructor decisions in course management, only to support them.

7. REFERENCES

[1] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, Dec. 1994.

[2] R. S. d Baker, A. T. Corbett, and V. Aleven. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In *Intelligent Tutoring Systems*, pages 406–415. Springer, 2008.

[3] P. J. Guo, J. Kim, and R. Rubin. How video production affects student engagement: An empirical study of MOOC videos. In *Proceedings of the First ACM Conference on Learning @ Scale Conference*, L@S '14, pages 41–50, New York, NY, USA, 2014. ACM.

[4] J. Kim, P. J. Guo, D. T. Seaton, P. Mitros, K. Z. Gajos, and R. C. Miller. Understanding in-video dropouts and interaction peaks inonline lecture videos. In *Proceedings of the First ACM Conference on Learning @ Scale Conference*, L@S '14, pages 31–40, New York, NY, USA, 2014. ACM.

[5] D. Kravvaris, G. Ntanis, and K. L. Kermanidis. Studying massive open online courses: recommendation in social media. In *Proceedings of the 17th Panhellenic Conference on Informatics*, pages 272–278. ACM, 2013.

[6] J. P. Meyer and S. Zhu. Fair and equitable measurement of student learning in moocs: An introduction to item response theory, scale linking, and score equating. *Research & Practice in Assessment*, 8(1):26–39, 2013.

[7] Z. A. Pardos and N. T. Heffernan. Modeling individualization in a bayesian networks implementation of knowledge tracing. In *User Modeling, Adaptation, and Personalization*, pages 255–266. Springer, 2010.

[8] Z. A. Pardos and N. T. Heffernan. Kt-idem: Introducing item difficulty to the knowledge tracing model. In *User Modeling, Adaption and Personalization*, pages 243–254. Springer, 2011.

[9] V. Raghuveer, B. Tripathy, T. Singh, and S. Khanna. Reinforcement learning approach towards effective content recommendation in mooc environments. In *MOOC, Innovation and Technology in Education (MITE), 2014 IEEE International Conference on*, pages 285–289. IEEE, 2014.

[10] D. Yang, M. Piergallini, I. Howley, and C. Rose. Forum thread recommendation for massive open online courses. In *Proceedings of 7th International Conference on Educational Data Mining*, 2014.

[11] A. M. F. Yousef, M. A. Chatti, U. Schroeder, and M. Wosnitza. What drives a successful mooc? an empirical examination of criteria to assure design quality of moocs. In *Advanced Learning Technologies (ICALT), 2014 IEEE 14th International Conference On*, pages 44–48. IEEE, 2014.