Applications of Bayesian Knowledge Tracing to the Curation of Educational Videos

Zachary MacHardy

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2015-98
http://www.eecs.berkeley.edu/Pubs/TechRpts/2015/EECS-2015-98.html

May 14, 2015

Acknowledgement

Thank you to my parents, Rebecca, my advisor Dan and everyone else who helped me stay sane and kept me on course

Applications of Bayesian Knowledge Tracing to the Curation of Educational Videos

by Zachary MacHardy

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Dr. Daniel D. Garcia
Research Advisor

(Date)

* * * * * * *

Dr. Zachary Pardos
Second Reader

(Date)

Abstract

With the popularity of MOOCs and other online learning platforms such as Khan Academy, the role of online education has continued to increase in relation to that of traditional on-campus instruction. At the same time, the need for analytical methods suited for the uniquely large and diverse populations that they serve has grown apace. In particular, as instructors and creators of online educational content grapple with these complex issues, the imperfect transfer of traditional informal, frequently affect-oriented methods of content iteration becomes clear. The need for additional quantitative tools for evaluating course content, taken alongside the opportunity presented by the scope and size of the data associated with such large enrollment courses, poses an interesting problem for analysis.

Rather than tackle the problem of evaluating large educational units such as entire online courses, our work 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 evaluate the usage of video content in addition to assessment activity. By exploring the change in performance when alternately including or omitting video activity, we suggest a metric for determining the relevance of videos to associated assessments.

This sort of evaluation is important for many reasons: struggling students can be pointed toward maximally efficacious resources, instructors can identify materials which may need adjustment, and courses as a whole can be better tuned to producing successful student outcomes. In order to provide an intuitive grounding for the validity of our results, we examine in detail the properties of videos that perform particularly well and those that do poorly, offering several case studies of the various data-sets included in this analysis. By proposing and demonstrating a new analytical approach to evaluating course content, we aim to move the promises offered by educational big data one step closer to practicable reality.

Contents

1 Introduction

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 learning [2]. As the number of online learners increases, so too does the importance of verifiably sound online pedagogy increase apace. Many of the lessons learned through a long history of research on the traditional classroom are applicable to the online environment; however, many of the indicators available to an instructor teaching classes to co-located students are not present for an instructor or a designer of online material. Teachers and designers are often unable to directly consult with students on what works and what does not, and lack, among other things, the affective in-class feedback that can often make such things apparent. This can be both a help and a hindrance; while the lack of affective feedback does hinder traditional techniques, research has shown that qualitative feedback collected from students doesn't always correlate well with learning outcomes [17].

Nonetheless, one part of the process of educational design that has been made particularly difficult by the move to a massive online format is the creation and curation of useful course resources. The design of curricular materials has been described as a process of iterative refinement [15] and, as with any design process, in order to refine curricular materials there must be metrics by which to evaluate them. Unfortunately, many of the strategies which have long been effective in the refinement of on-campus and in-person courses are less tractable in online environments. Because of a lack of affective information and severely differentiated levels of student knowledge and participation, many challenges which are less pronounced in traditional settings come to the fore when courses move to a massive, online format.

Research on how best to evaluate and improve online education is not new [1, 28] but there remain many distinct approaches to accomplishing this goal [18]. While the problems facing designers of instructional material intended for massive audiences are multifarious, the quantitative evaluation of course materials remains a particularly difficult, and as yet unsolved, problem. Though many data-driven metrics for examining assessments are available, there has been relatively little focus placed on assessing the course materials which aim to help students complete those assessments. This sort of evaluation is important for many reasons; struggling students can be pointed toward maximally efficacious resources, instructors can identify materials

which may need adjustment or removal, and courses as a whole can be better tuned to producing successful student outcomes.

Compounding the problem, simple analysis can often yield discouraging results. In many instances, usage of course resources can negatively correlate with assessment performance. It may be that this is reflective of reality; perhaps it is the case that particular course materials contribute to incorrect mental models, perhaps due to poor scaffolding, instructor error, or some other cause. But discounting this rather grim possibility, it seems more likely that there are a number of confounds which serve to obstruct simple analysis. Whether due to a diversity in student backgrounds, differentiated patterns of interaction with instructional materials, use of external resources, or some combination of these and other causes, it seems intuitive that accounting for such differences may improve our grasp of resource quality. Our proposed method accounts both for student growth over time, and the possibility of interventionary effects for students who first struggle, then succeed at certain problems. In so doing, We hypothesize that we might both be better able to predict future student performance, and as a result, measure resource efficacy.

In order 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 in conjunction with Intelligent Tutoring Systems (ITS) but which has been applied outside of that domain as well (e.g. [19]). Here we seek to incorporate video observation, which lies outside of the sort of student behavior, namely assessment activity, that is typically considered in BKT models. 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 in so doing is not to achieve high accuracy for the purposes of ITS-like prediction of students' latent knowledge. Rather, we intend to provide a quantitative framework to aid instructors in the evaluation of video resources.

We set out first to prove that there is a statistically significant improvement in performance when incorporating video resources into BKT analysis, in order to validate the inclusion of such observations. This step is a necessary one to validate any conclusions drawn from our analysis, in order to demonstrate that we are doing more than observing random noise. Second, we discuss a metric based on both the delta in predictive error when using

and eschewing video data, as well as the rate of learning associated with a particular video. To this end, we examine the application of our method across three sets of data, taken from Khan academy and edX, looking closer at models which perform particularly well and those which do poorly. By so doing, we hope to understand what qualities lend themselves to high performance, and reason about why certain videos are poorly associated with later assessments. Finally, we suggest a number of potential applications of such a metric, from student-facing recommender systems to instructor-oriented tools for improving course content, along with several algorithmic refinements that might further increase the power of our analytic approach.

2 Related Work

We are, of course, building upon a large quantity of work which has already been done, in the study of both Bayesian Knowledge Tracing and Computer Assisted Instruction, as well as in general instructional design. By understanding the theory and practice which underlie both of these larger concepts, we can better situate our work in the larger scheme of research.

2.1 Bayesian Knowledge Tracing

Bayesian Knowledge Tracing (BKT) is used extensively in computer-assisted instruction environments, intended to approximate the effects of one-on-one mastery learning in environments where such instruction is not economically feasible [7]. Originally designed for use with the ACT Programming Tutor, it has since emerged as a popular tool in the research and practice of intelligent tutoring system design.

2.1.1 Theoretical Foundation

The model's theoretical underpinnings are borrowed from the conceptual framework provided by cognitive theorists for understanding the way students learn. According to basic cognitive theory, introduced by Jean Piaget in the early twentieth century, and studied actively over the decades which followed, a child is born with a basic mental structure, the basic components of which are used to construct iteratively more complex models as the child learns[5].

Under the tenets of this theory of learning, all concepts mastered by a learner are assembled, or constructed, out of components already mastered

by the student. Implicitly, this gives knowledge a hierarchical structure; all knowledge is built from a series of prerequisite components which must be mastered before the learner can understand more advanced concepts. For instance, a cognitive theorist might hold that a student must master and understand addition before meaningful mastery of multiplication can be made possible. Each newly acquired concept is integrated, or assimilated, into a learner's mental framework, and can be used as a component in formulating an understanding of more complex topics.

The idea of mastery learning was essential in the work of Bloom et. al [4], who hypothesized that significant improvements in student performance could be observed if students could be brought to mastery of each concept they encountered before they moved onto the next. In a landmark study published in 1984, Bloom observed that the average student who was individually tutored to mastery in a number of successive concepts saw a performance increase of two standard deviations compared to peers who covered the material in the traditional fashion. This effect manifested not only in academic performance, but also in students' academic confidence and self-concept. Bloom claimed that the success of the intervention was due to two related factors: first, the students were treated individually in one-on-one environments with tutors; second, the students were brought to mastery (as determined by the tutor) of each subject before moving on to the next.

The so-called "Two Sigma Problem" posed by Bloom and researched actively in the decades following the publication of the study, is the search for methods of group instruction as effective as one-on-one mastery learning. Though subsequent studies on mastery learning have shown effect sizes smaller than those demonstrated by Bloom, the results have nonetheless overwhelmingly shown associated increases in academic performance. Bayesian Knowledge tracing, and the ACT-R tutor for which it was proposed as a component, are part of one attempt at replicating the two sigma effect. By leveraging the scalability of automatic tools for student instruction, Corbett and Anderson hoped to be able to provide a feasible mechanism for individualizing and automating mastery learning.

The essential pieces of the cognitive understanding of knowledge acquisition for the purposes of BKT are the existence of discrete knowledge components (KCs), as well as the concept of subject 'mastery'. More specifically, BKT is a means of predicting when a student has acquired a knowledge component associated with a set of assessment items, typically to ensure that a student has attained mastery before moving on to the next subject. Several

simplifying assumptions are usually made in order to facilitate the formulation of this model. First, subject mastery is modeled as a binary state: a student has either mastered a KC or has yet to grasp it. Second, this mastery, being itself unobservable, is assumed to be reflected in observed responses to assessment items concerning that KC. In order to account for the presence of lucky guesses or silly mistakes, the model conditions the probability of a correct response on the possibility of observing such 'noise.'

2.1.2 Knowledge Components

Though BKT provides a convenient framework for modeling the acquisition of skills over time, it does not provide a means of discovering exactly what knowledge components comprise a subject or set of subjects. Instead, it requires a manually-defined set of knowledge components to have been predetermined for use with the model.

The problem of defining knowledge components is more general than its application within Bayesian Knowledge Tracing. Sometimes referred to as 'Knowledge Structures,' or 'Knowledge, Skills, and Attitudes' [12] the concept of a set of discrete components which comprise a more complex subject or field has been a subject of active research for a number of years. Though there is not yet consensus on best practices for defining knowledge spaces, several approaches to discovering these structures have been described.

One relatively straightforward approach, and perhaps the most often utilized, is to defer the task of building a knowledge structure to domain experts. Typically drawing on a small pool of experts in order to establish consistency and ensure validity, such expert-defined structures are often used as first-pass attempts upon which further research can iterate. Additionally, there are a number of heuristic approaches to transforming expert input into well-defined knowledge structures. For example, Koppen and Doignon [11] describe a method for building 'quasi orders,' allowing a pool of experts to define structures implicitly by asserting dependencies between assessment items, rather than through explicit definition. Some critics of this process have noted that incorrect assertions by experts about such dependencies, not unlikely given the number of assertions that knowledge space construction can involve, can drastically change inferred knowledge structures [26], with deleterious implications for systems which utilize them.

Others approaches have considered the possibility of discovering knowledge structures by analyzing large quantities of data. Van Leeuwe [31] devel-

oped an algorithm referred to as classical Item Tree Analysis (ITA) in 1974, used to assemble test items into a hierarchical structure based on student response patterns. Schrepp [25] developed a similar method, inductive ITA, for performing the same function, though through a different process. Both methods, similarly to the methods of Koppen and Doignon describe above, construct a quasi-order of test items, which is used to define a knowledge structure underlying an assessment. Though not entirely robust to the effect of item difficulty, the concept that some assessments are more difficult than others while testing the same material, such automated assessment can give a reasonable, if not always intuitively interpretable, decomposition of knowledge components.

Because the discovery and interpretation of knowledge structures is an active subject of debate, some approaches to KC definition have been more restricted in their scope. Though the hierarchical structure of knowledge components is an essential part of the cognitive theory which underpins Bayesian Knowledge Tracing, the prerequisite relationships between KCs are not (necessarily) themselves part of the model. For the purposes of predictive analysis across large data-sets with many student participants, it has been shown to be sufficient to consider individual assessment items or groups of the same as knowledge components, while remaining agnostic to the relationships between them. Pardos et al. [19] have explored the quantitative differences in predictive accuracy and error when using different levels of problem granularity as KCs for the purposes of applying BKT. Though such approximations do not speak to an underlying structure in the content, it has been shown to be a reasonable approximation for the purposes of predicting student response patterns.

2.1.3 The Bayesian Knowledge Tracing Model

Ultimately, the BKT model can be represented as a Bayesian network, with observed nodes representing responses to assessment items, and unobserved nodes representing the student's internal mastery of that concept at a given time.

The model in its most basic form is defined by four parameters: $P(L_0)$, the prior probability that a student has mastered a KC; $P(S)$, 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

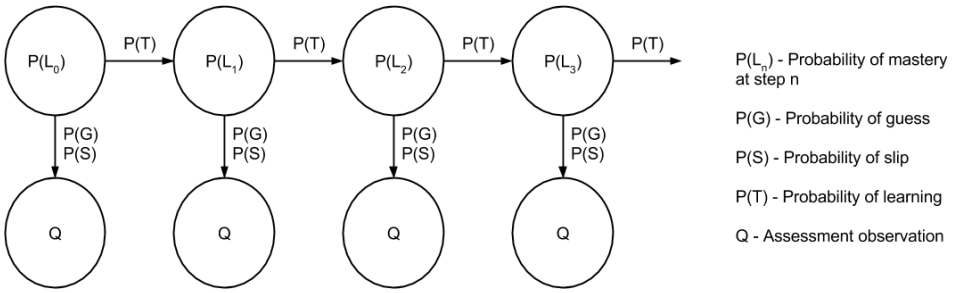

Figure 1: Bayesian Knowledge Tracing Model

who does not know a particular KC will learn it after a given observation.

The chance of a correct answer at a given point in time can be described simply as

$$P(correct) = P(L_n) * (1 - P(S)) + (1 - P(L_n)) * P(G))$$

where n represents the nth observation related to a particular knowledge component. Put simply, this equation represents the chance that the student either knew the answer and did not make a mistake (slip), or that they did not know the answer, but happened to guess correctly.

The process of inferring KC mastery based on observation is simply an application of the more general Bayes' theorem, which holds that, for some event A and some event B,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

In our case, we are measuring the probability of KC mastery, given the observed correctness of a student response. That is, $P(L_n)$ is calculated in an iterative process using Bayes' theorem, as follows. First, a $Posterior(L_{n-1})$ is calculated as the result of

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

if the observation was a correct problem attempt, or

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

if the observation was incorrect. Finally, after each observation, the probability of having learned the KC after a particular observation, $P(T)$, is incorporated into the model as

$$P(L_n) = Posterior(L_{n-1}) + (1 - Posterior(L_{n-1})) * P(T)$$

We chose to use BKT as a modeling framework as it is well-studied and has relatively well-understood properties, in addition to possessing parameters (guess, slip) 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 [20, 8], as well as properties of particular assessment items within a knowledge component [21].

Most typically, BKT is used in intelligent tutoring systems designed to track students as they work through a series of questions, which explains the reliance of the model on assessment-response observations. Much effort has been made to improve the performance of these systems, as well as to test the reliability of its application outside of traditional environments, but most focus remains on tracking students as they complete assessments. By using the predictions of student knowledge obtained through the BKT model, tutoring systems are able to optimize student activities to ensure an approximation of mastery learning, while not wasting time on redundant problem-solving.

2.2 Instructional Design in Online Education

An essential part of developing curricular materials for use in any educational context is a framework for reasoning about and judging the efficacy of curricular components. While broad pedagogical techniques are often developed and used based on a theoretical understanding of the process of learning or the nature of knowledge, the practical reality of employing these strategies often involves multiple iterations of curricular content. While the thrust of a pedagogical strategy may or may not differ between each iteration, the individual components of these curricula are frequently subject to change based on a number of different strategies, some formal and some informal.

2.2.1 Curriculum Evaluation

Though it is subject to a number of different constraints than courses developed and offered online, it will be useful to first understand the process

by which course materials are iterated upon in the traditional classroom. As Linn notes in [15], studies of learning in laboratory settings in the tradition of cognitive psychology can only be taken as partially informative as to the use of curricular interventions in real world classrooms. Quite a bit of recent research on curricular improvement has turned toward the use of design-based studies, employing and detailing a process of progressive refinement [6] as researchers encounter and account for emergent or unanticipated features of real-world usage. Several other researchers, toward the goal of effective course iteration, have described a number of different features which can be used to evaluate curricular components [23, 24, 27, 3, 6].

Common to many of these studies is a particular focus on adapting curricular intervention to the practiced reality of students in the classroom. Though specific methods differ, by observing students' interactions with one another and with educational materials, researchers are able to qualitatively identify and rectify problems with course content. In particular, researchers often watch students for affective details which might indicate a lack of engagement with material, as well as employ think aloud processes that allow insight into the ways in which students are forming mental models about relevant subject matter. Further, efforts are made to ground curricular material in the cultural realities of the communities in which they are employed, embracing the heterogeneity of the student population.

Of course, in both design-based research and more traditional laboratory-based studies, quantitative statistics reflecting student performance are used in roles of greater or lesser importance. However, these results are sometimes reflective of the purported effects of interventions as a whole unit, rather than of the utility of individual curricular components. While nomothetic considerations are important, particularly when trying to argue for the generalizability of results, it is very often the idiographic components of the classroom that are considered while iterating on curricular materials.

2.2.2 Developing Online Curriculum

This process of iteration can be challenging when grappling with the different sources of information available in online environments. Unlike traditional classrooms, where affective observation and culturally-relevant adaptation are a regular part of teaching practice, the anonymity and scale of online education can make such considerations difficult or impossible.

There has been a fair amount of research devoted to studying, both

qualitatively and quantitatively, 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 to foreground relevant course content to students enrolled in the course. 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 [33]. Similar efforts have been made to provide student-facing recommendations for more general content, using methods such as social media analysis and reinforcement learning [13, 22]. While useful, such efforts tend to focus on students as both consumers and curators of information available in the courses, agnostic to the quality of the content itself.

Relative to the research on student perception and experience in the MOOC context, somewhat less has been paid to instructor experience in constructing and maintaining online courses. That is not to say that such work has been absent. Guo et al. [9] and Kim et al. [10] offer guidance for the construction of videos used in MOOCs. Explorations of the application of Item Response theory in a MOOC environment [16] similarly offer instructors guidance in evaluating the efficacy of their assessments using traditional methods. Yousef et al. construct an inventory of features, pedagogical and technological, which contribute to a sense of course quality [34], while others have delved into sentiment analysis in MOOC forums [32]. There is, however, a relative paucity of research on the quantitative assessment of content outside of the scope of assessment items. In the absence of such quantitative information, instructors tend to look for traditional, affective feedback; prior work, such as that by Stephens-Martinez [29] has suggested that instructors frequently resort to observations of student forums and student surveys in order to draw conclusions about the quality and efficacy of course content.

Arguably, given the fundamentally different constraints placed upon the online and traditional environments, design considerations when developing online curricula should accordingly differ. While, for example, studies of student affect in forum populations may be useful, differential levels of participation in these social functions can make generalization difficult. Similarly, students in the same massive courses may have vastly different levels of ability or knowledge; making the process of designing appropriate assessments very difficult. Indeed, it is unclear that it is even possible to support all of these diverse students all of the time. Nonetheless, the development of a framework for making quantitatively driven decisions about the efficacy of

educational content used by many students, as reflected in student performance, may be an essential step in developing a design process which takes into account the affordances of online education.

3 Methods

In order to achieve our goal of incorporating video resources into traditional BKT analysis, we must first accomplish several goals. Below, we describe the processes we have designed for incorporating videos into the general BKT model, discovering knowledge components in data not structured for use with BKT, and constructing an evaluative metric to determing the relevance of video content.

3.1 Incorporating Course Resources

Our interest in leveraging BKT to incorporate course resources stems from two separate, though related, concerns. First, while BKT has traditionally served as a strong predictive model when considered within the framework of Intelligent Tutoring Systems (ITS), where student interaction is largely limited to responding to assessment items, interactions with MOOCs can be significantly more heterogeneous. Though strong predictions can nonetheless be made by only considering student response information, such analysis ignores a wealth of contextual information about student activity, from time spent, to interactions with other students, to the consumption of course resources. Though ideally, the BKT model might be adapted to consider much of this information, we have chosen first to investigate the interaction of course resources, in this case videos, with students' knowledge states. We hypothesize that the inclusion of such extra information in the BKT model can be used to reduce its predictive error, resulting in a more broadly informed and therefore more useful model. Further, by examining the properties of these more effective models, we will be able to discover the efficacy of the videos used to inform them, allowing us a broader view than that of traditional assessment-based BKT.

Second, one concern when reviewing and iterating upon course materials is the evaluation of the utility of course resources. Though one concern among many, it can be useful to understand how useful or unhelpful a particular piece of course content for students completing associated assessments.

However, an interesting and somewhat paradoxical trend has emerged in many sets of data obtained from massive offerings of online courses. That is, when considering the relationship of the consumption of course resources to success on subsequent assessments, a markedly negative trend is often observed.

A naive reading of this trend might explain this relationship as a process of negative learning; students come in with some notional understanding of a given concept, and as though subject to some sort of phantasmagorical, knowledge-sapping force, they leave less knowledgeable than they arrived. Or perhaps, less supernaturally, they are subject to some new set of misconceptions imparted by poorly designed course resources, hindering their progress toward true mastery. Either way the frequency with which this negative relationship is observed suggests that, excepting the possibility of uniformly bad design on the part of course curators, there is an alternate explanation. To this end, we hypothesize that this inverse relationship is reflective not of actively harmful learning effects imparted by course resources, but of modally different student interactions with massive online courses and their materials.

To motivate the hypothesis and give an intuitive example, imagine that two different students approach the same online course. The first student, Sage, is an expert in the domain that the course covers; she has approached the course with the intention of shoring up and self-assessing her own skills. The second, Joy, is entirely new to the domain, and excited to begin her studies; she is so excited, in fact, that she has neglected to study some of the pre-requisites listed on the course page.

As the two proceed through the course, they interact with the materials very differently. Since Sage has a background in the domain, she tends to skip the lecture videos and proceed directly to assessment. Joy, on the other hand, voraciously consumes course material, hoping to get an additional leg up on some of the more complex concepts being covered. As Sage moves on to assessment, she finds her faith in her own abilities borne out; after some effort, she succeeds with flying colors, typically needing no more than a single attempt to solve a given problem. Joy struggles at first, getting several problems wrong. Eventually, however, after returning to an earlier video, she returns to the problems that gave her trouble and manages to work out the correct answers.

Given context, the source of Joy's struggles is not the videos she is watching; in fact, she leverages them to improve her performance. Yet a naive

interpretation of the data would tell us that video consumption is associated with a lower rate of success. This is, of course, technically correct. However, to conclude from such an analysis that resource usage is harmful would be specious; while the correlation may be strong, the two variables are clearly dependent on a third, latent variable: the student's prior knowledge. Of course, this is a conveniently-constructed thought experiment. But if it were indeed the case that there are significant variations in student-resource interaction, and that these differences were informed by categorically distinct types of students, then such a distinction could be leveraged to better understand how the usage of course materials would affect a given student.

To this end, Bayesian Knowledge Tracing offers a useful frame for reasoning about such distinct student effects. Since BKT is designed with the concept of active student learning in mind, it is particularly adept at capturing and modeling interventionary effects. For example, Joy in our experiment above first struggles with a problem, references a resource, then returns and succeeds at the same problem. Modeled using Bayesian Knowledge Tracing, Joy's latent knowledge state would be, after the first mistake, computed lower than first assumed, incorporating her failed attempt. The utility of the resource she then consults is incorporated into the calculation of that latent value, raising the probability that she has attained concept mastery. Finally, our updated estimate indicates that she probably understands the concept, and, indeed, she succeeds on her next assessment attempt. In this way, since BKT is built to consider a temporally ordered series of events, with an updated tally of student knowledge, it is particularly well suited to modeling this sort of interaction with resources.

As a relatively simple graphical model, BKT is also easily modified to incorporate per-student parameters. By partitioning students into distinct groups based on some set of contextual features, it is relatively easy to condition both student priors and, potentially, other model parameters, on a student's membership to a certain group. By so doing, one might more accurately capture the utility of individual course materials, laying the groundwork for the construction of a quantitative measure of the same.

3.2 Generating and Associating Knowledge Components

Of course, in order for us to perform any analysis at all using the Knowledge Tracing Model, it is a necessity to identify both what knowledge components a course comprises and which videos and assessments are related to those

components. As discussed in the Related Work section above, the process of identifying KCs is onerous as well as controversial; the definition of a knowledge space is domain specific and can involve many iterations. Drawing upon previous work in the domain of MOOCs and KT, and in the interest of proving a generalizable framework for evaluation, we have chosen to identify KCs at the problem level, setting aside the issue of knowledge space construction in favor of relative simplicity.

While this handily avoids the issue of assessment-KC association, relating videos to related KCs remains problematic. Ideally, as with knowledge components, these associations would come provided, generated by course instructors or domain experts. Alternately, one simple solution would be to consider all videos that are a part of a particular section in an educational unit or course as related to assessments within that section.

A look at the data used for this report, however, reveals several issues with that approach when considering a generalizable framework. Besides issues of data-completeness, the variance in course format means that the meaning of a "unit" can vary broadly between courses, making the presumption of association more or less meaningful depending on the way the course was constructed. With a goal of preserving the generality of our approach and avoiding the ambiguity of instructor-defined units, we chose instead to design an algorithm for automatically tagging problem-video associations ourselves.

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. More specifically, we observed the KCs which appeared most frequently in student logs following the use of course resources, allowing for some limited distance between video and attempt, but excluding those activities which occurred more than an hour apart. 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. Restrictions placed on these associations were strict enough that, upon sampling and manually checking a number of generated associations, they appeared sound. Nonetheless, this does introduce possible sources of error in terms of both overlooked and spuriously constructed mappings.

An illustrative example of the association process can be seen in Figure 2. In this example, a short segment of 5 users' event logs is visible. Because quiz B occurs frequently after users view video A, and quiz N frequently appears after video L, A-B and L-N are suggested as candidate pairings for analysis.

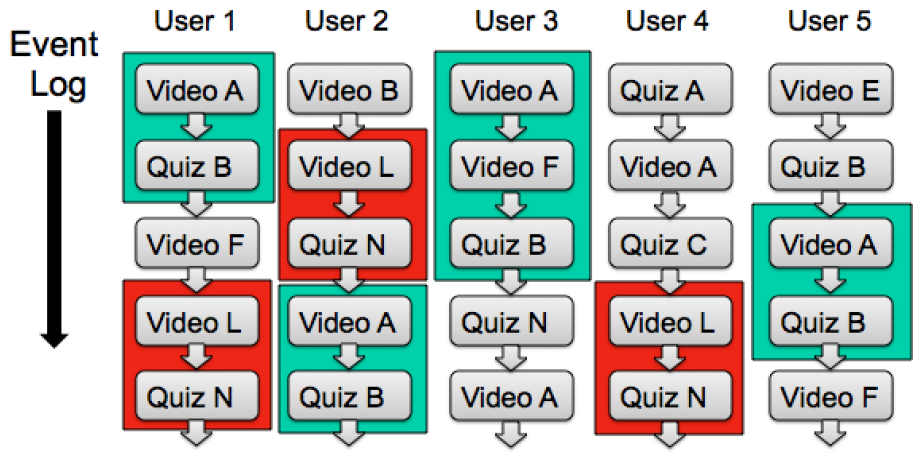

Figure 2: Candidate KC-video pairs are suggested by analyzing user event logs

Note that the video and quiz need not be consecutive, but rather need only to appear within the neighborhood of the exercise to be considered. For the purposes of this analysis, we chose to consider videos which appeared within 10 log events of a subsequent quiz, and which were observed in the hour prior to an attempt. Also worth noting is that for each video at most one exercise (or KC) is taken to be related, chosen based on which exercise appeared most frequently in relation.

In order to facilitate generalized analysis, all data was parsed and reformatted into an intermediate format, leaving the analysis agnostic to the source of the data analyzed. It is worth noting as a caveat, here as below, that only the Khan logs had information about the multiple templates used for each exercise. That is, rather than a single, identical problem, many Khan problems were composed of randomly generated numbers applied to a general problem template, generating different but structurally similar problems. Thus, though the data is ultimately in the same format for all three sources considered, there is some information available for the Khan data that is not present for either of the edX courses. This does not significantly affect the thrust of the analysis, but should be noted when considering the four models proposed below.

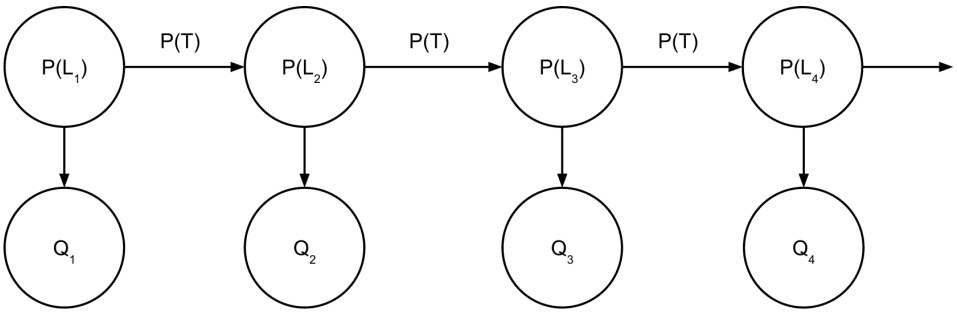

Figure 3: Standard Bayesian Knowledge Tracing Model

3.3 Extending the Bayesian Knowledge Tracing Model

In order to examine the effect of course resources on the learning process, we employ several extensions to the traditional Bayesian Knowledge Tracing model. First, and for each model we use in the evaluation, we condition $P(G)$ and $P(S)$ for each observation on which specific exercise within a KC is being observed. That is, given a number of KCs k, containing a number of sub-problems n, we generate $2nk$ total guess and slip parameters, a technique which has been shown in previous research on applying BKT to MOOC environments [19] to produce significantly better predictive accuracy. Intuitively, this extension allows the model to account for variations in problem difficulty among sets of problems related to the same knowledge component, allowing guess and slip to vary with the individual properties of an exercise. Hereafter we will refer to the traditional BKT with this extension as 'Standard BKT' (see figure 3), and it serves as the baseline to which other models are compared.

Our second extension mirrors our first, conditioning the transition probability $P(T)$ on the specific exercise within a KC that is observed. As before, this multiplies the space of transition parameters trained by the average number of problems that fall within each KC, accounting for differential learning effects which might be seen between different exercises. We include this model for the Khan data for the sake of completeness, to account for any change that might result specifically from conditioning $P(T)$ on individual exercises without including resource data. As we treated each individual problem within the 'Statistics for Medicine' and 'Principles of Economics'

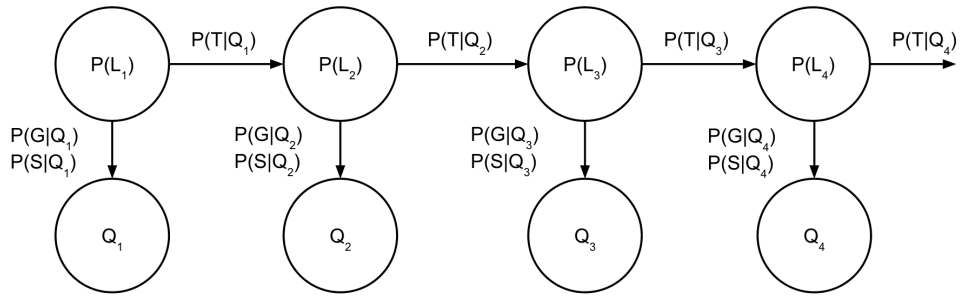

Figure 4: Template Model, conditioning $P(T)$ on which template is observed

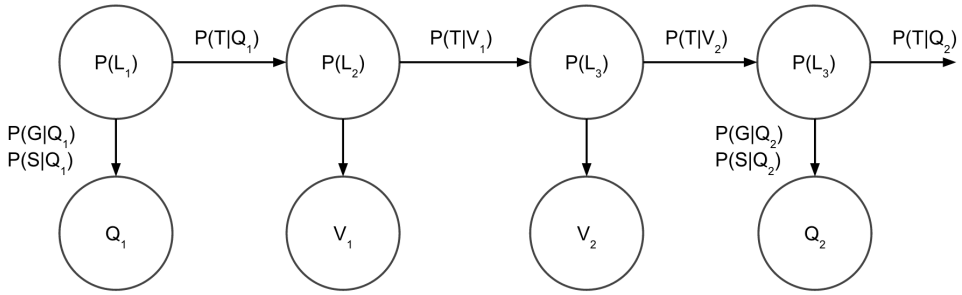

Figure 5: Template-Videos Model, including video observations

courses as separate KCs, this model is omitted for the edX courses. In our analysis, we refer to this extension as the 'Template' model (see figure 4).

Pursuant to our interest in incorporating course resources into our investigation, our third extension, and the first which considers resource-related data, adds video activity as additional observations to the BKT model. As these observations are not associated with notions of correctness, and there is consequentially no notion of 'guess' or 'slip', there is no inference performed as a result, unlike the incorporation of response data. Instead, video observations are associated only with a transition probably $P(T)$, taken to be unique to each video.

Conceptually, this third model includes the probability that a given course

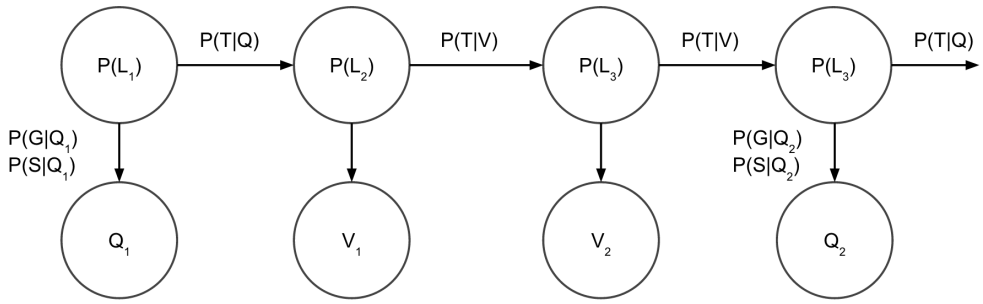

Figure 6: Template 1 Video model, not conditioning $P(T)$ on resource or exercises identity

resource will impart mastery upon a student, updating their calculated $P(L)$ accordingly. The mathematical implications of this inclusion are straightforward. Recall from the previous description of BKT that the calculated posterior probability of mastery is inferred from the observation of a correct or incorrect student response to an assessment item. This step is entirely omitted in the case of educational videos, as no such correctness information is associated with their use. Instead, we simply take the prior as the 'posterior' and use the same equation to update $P(L)$ by using the $P(T)$ associated with a particular video.

$$P(L_(n+1)) = P(L_n) + (1 - P(L_n)) * P(T))$$

While this is a simplifying assumption when considering student-resource iterations, it nonetheless fits well into the Bayesian Knowledge Tracing framework, allowing for the same simple calculations used to incorporate exercise data to be applied to the use of resources. We refer to this model as the 'Template Videos' model (see figure 5).

Finally, we simplify the 'Template Videos' into a 'Template 1 Video' model (see figure 6), conditioning $P(T)$ only on the presence of either a video or a question, but not the specific identity of the resource observed. This reduces the number of parameters trained by the model, potentially allowing for better results when data is relatively sparse. We summarize each of the three models in Table 1 below.

Model	Properties
Standard BKT	Unique guess and slip trained for each assessment
Template	Unique $P(T)$ trained for each assessment
Template-Video	Videos included, each with a unique $P(T)$
Template-1-Video	Videos included, one $P(T)$ for each class of observation

Table 1: Properties of each BKT model

3.4 Constructing an Evaluative Metric

Constructing and testing the predictive validity of our extension to BKT is only a means to end. Ideally, we would like to use our model as a tool to help instructors reason about one dimension of the efficacy of their course resources. Toward that end, it is useful to understand what, exactly, our analytic methods produce.

There are two essential dimensions to the output of our framework. The first is simply the delta in model error when considering the use of course resources. That is, when we employ our extension to BKT, and compare its predictive error with that of a model disregarding resource usage, to what degree and with what significance does our error change. This is a fundamental consideration, particularly because not all models will necessarily be well informed by their associated videos. Indeed, in many cases, the addition of course resources may simply add noise, not affecting or possibly even increasing predictive error. This is not to say that such noise is entirely meaningless, but simply that in order to establish a notion of positive video efficacy, it is first important to establish the validity of the extended model. Put simply, a model which better fits the data will have relatively lower predictive error, and as such the delta in RMSE should increase positively as model fit improves compared to standard BKT.

Second, assuming that the model is validated by lower predictive error, the actual properties of that model can be examined. Since video observations are associated only with a $P(T)$, the transition probability that associates a resource with a chance of mastering material, this is a relatively straightforward process. Videos which have a high transition probability can be considered as tightly coupled with their related assessments, while low transition probabilities may be indicative of only loose relation. This measure is, of course, not a value judgment on the quality of the resource *per se*. For example, there may be a case where a conspicuously high observed

transition probability is undesirable, indicative that a video may be doing no more than simply 'teaching to the test'.

The specific aim of our proposed framework is to help instructors understand both which videos are associated most strongly with student success and also those which introduce the most noise. By using the delta in predictive error and the properties of the trained models, we are able to establish the validity of an established relationship and give a measure of the strength of that relationship. Were we only to use the notion of statistical significance, we run the risk of ignoring differential levels of benefit, while using only the properties of trained models risks using specious or unreliable information. Further, we hypothesize that particularly noisy models may be the result of particularly inapt resource-assessment pairings, an observation which may be of particular interest to instructors looking to improve or adjust course materials.

It is tempting to attempt to combine both of these measures into a single, easily-digestible summary statistic. Unfortunately, this reduction in dimensionality would come with a significant loss of information: one would not want to equate a statistically powerful model with a relatively low $P(T)$ with a statistically weak model trained with a spuriously high $P(T)$. For this reason, we consider and discuss these measures as two fundamentally different, but related metrics, both of which are useful for determining the properties of particular educational resources. After finding the models which seem most strongly correlated with their associated assessments as well those that were most deleteriously affected by including resource information, we can proceed to use the specific parameters of these models to draw conclusions about the resources themselves. By leveraging this information, we we hope to offer instructors an additional tool for understanding and improving subsequent iterations of educational material.

4 Analysis

We applied our methods, described above, to three different sets of data. One set comes from the Khan Academy platform, and consists of students working through a variety of problems, without the notion of an overall 'course' guiding their work. The second and third sets come from the Principles of Economics and Statistics in Medicine edX courses offered by Stanford during the summer of 2014. Below, we seek to verify the efficacy of our method,

and then proceed to a qualitative analysis of high and low performing models from each data-set.

4.1 Data

The data we obtained from Khan Academy contains 1,044,930 problem attempts and 3,797,676 video observation events collected over about two years, from June 2012 to February 2014. Assessment items 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. Though more complex approaches to discovering the concepts which underlie educational content have been described [14], 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. After filtering out unassociated videos and exercises that were associated with fewer than 500 events, 353,202 events remained, representing work within 187 distinct exercises and 353,202 distinct student-exercise pairings. Of the 187 exercises, 176 (91%) were associated with video observations, with around 10% of all events being video viewings. Each exercise was associated on average with 1,803 events.

In order to demonstrate the generalizability of our results, we also leveraged event log data taken from two Stanford Online courses run using the edX platform: 'Statistics and Medicine' and 'Principles of Economics.' Both were offered from June to September of 2014. After filtering the data provided down to problem and associated video activity, we were left with, respectively, 215,716 and 122,077 problem attempts as well as 473,993 and 215,351 video viewings. Based on past research [19], we chose to consider each individual problem as a knowledge component, leaving us with observations spread among a set of 95 and 71 KCs. Each individual KC was associated, on average, with 6,250 events. Unlike the Khan data-set, the preponderance of observations were video events, comprising around 67% of all recorded events.

All models used in our analysis were trained and evaluated using 5-fold cross validation. For each model above, one BKT model was trained for each of the knowledge components identified in each of the data-sets. 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. The error used to compute the RMSE was calculated by predicting the probability of correctness at each problem attempt, the finding the difference between the computed probability of success [0, 1] and the observed result {0, 1}.

4.2 Results

Tables 2, 3, and 4 describe the results of running the data through the three analytical models. Figure 7 shows the performance of each model in each data-set, under both the 'Template Videos' and 'Template 1 Video' conditions, with each KC represented by one bar in the graph, and the y-axis showing the delta in performance (higher is better). In order to make the distribution more visible, the KCs are ordered by the delta in performance observed when employing video data.

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. The mean RMSE across all KC models is provided only as a guide for understanding how each analytical model performed compared very generally to the others, and is not the focus of the analysis.

Model	**Mean RMSE**	**p**
Pct. Correct	.4924	.00
Standard BKT	.3837	—
Template	.3837	.94
Template Videos	.3825	.02
Template 1 Video	.3826	.01

Table 2: Khan Academy

Though the tables reflect changes in RMSE aggregated over all KC models, not all models in each data-set benefited evenly from the inclusion of video resources. Among the Khan data 72 of 187 KCs saw more then a trivial amount of reduction in error between the 'Standard BKT' and 'Template Videos' conditions. In the case of the Statistics and Medicine class, the bulk

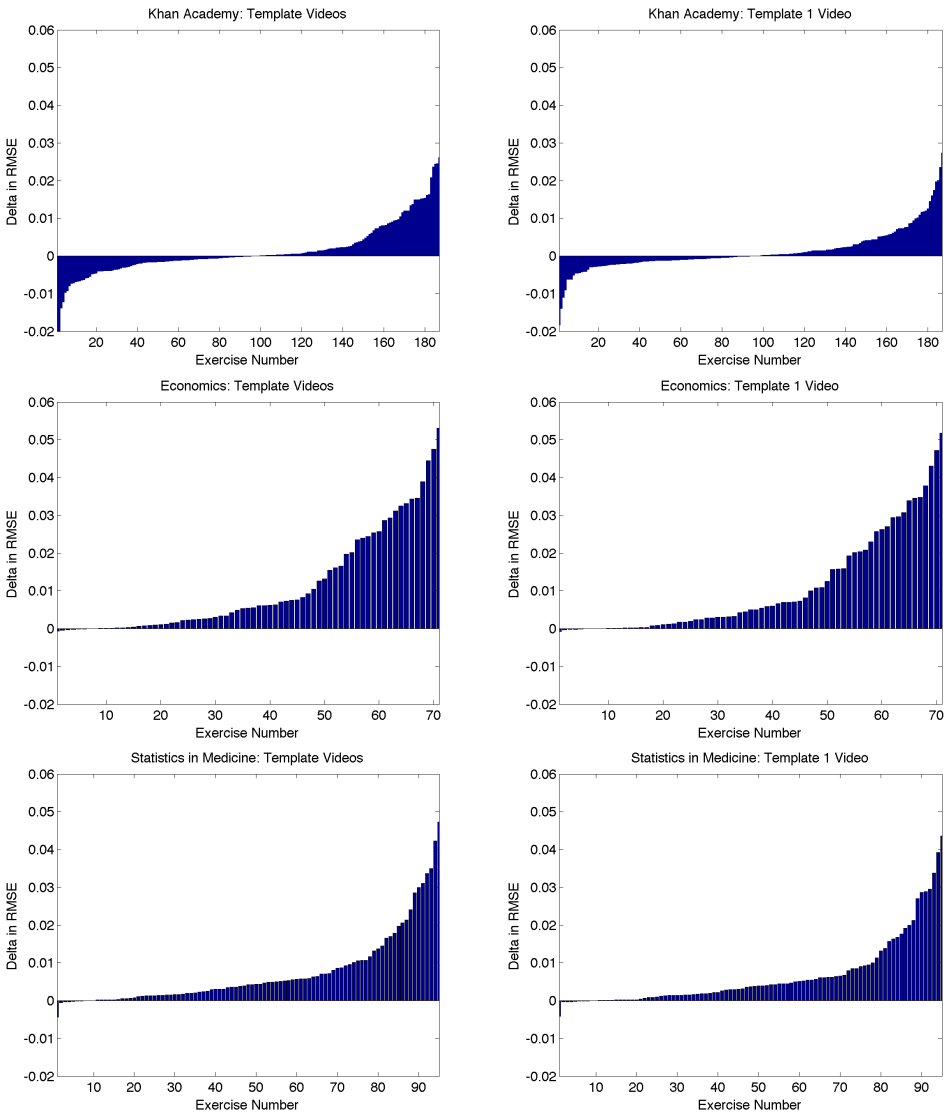

Figure 7: Delta RMSE by KC (Higher is better)

Model	Mean RMSE	p
Pct. Correct	.6229	.00
Standard BKT	.3824	—
Template Videos	.3715	<.001
Template 1 Video	.3718	<.001

Table 3: Principles of Economics

Model	Mean RMSE	p
Pct. Correct	.5144	.00
Standard BKT	.3711	—
Template Videos	.3638	<.001
Template 1 Video	.3646	<.001

Table 4: Statistics and Medicine

of the improvement could be seen in 73 of the 95 models, with the remaining models performing the same or slightly worse than before. For Principles of Economics, the numbers were similar, with 51 out of 71 models showing more than very minor improvements.

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 and possibly even worse. 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. Further, by examining those videos which offered the best improvements and those that affected their associated model most deleteriously, it may be possible to discover and highlight both the most and least useful.

4.3 Analysis Properties

Before moving on to qualitative analysis, however, other properties of the data are worthy of consideration. Tables 5 and 6 give some general properties of each data-set included in the analysis. The differences between the distributions of resource usage and assessment attempts are notable, with

Data set	Average Attempt Count	Standard Deviation
Khan Academy	6.21	2.29
Principles of Economics	1.19	0.37
Statistics in Medicine	1.23	0.35

Table 5: Average and standard deviation for number of attempts made by each student on each exercise for each data set

Data set	Average Resource Count	Standard Deviation
Khan Academy	0.64	0.45
Principles of Economics	1.87	0.64
Statistics in Medicine	2.45	1.23

Table 6: Average and standard deviation for number of resources viewed by each student on each exercise for each data set

individuals in the Khan data generally registering more assessment attempts but fewer video observations than the data drawn from edX. This is consistent with the features of each platform, but is important to consider as we move on to examine some of the other properties of the analysis.

Figure 10 shows the measured delta in RMSE when considering student traces involving different numbers of question attempts. Interestingly, though the details differ for each data-set, several properties are shared. First, predictive error across students who made only a single attempt improved for all three data sets. This seems to follow relatively logically, as we typically have more information before making our first prediction, if that student has used a resource one or more times. Second, though the delta is not generally positive for students who make 3 attempts on exercises in Khan Academy, there is a general trend of improvement as student traces grow to 4 attempts in length.

Oddly, in the cases of both Statistics in Medicine and Principles of Economics, student traces which had five or more attempts showed the worst performance, while Khan Academy saw relatively poor performance with three. It appears that incorporating resources helps us do particularly well when predicting single-response sequences, but seems to generate some confusion with moderate-length sequences. It may be that the resource-inclusive model deals relatively poorly with sequences involving an initial slip or two and then a subsequent correct response, as the P(T) for attempt-related ob-

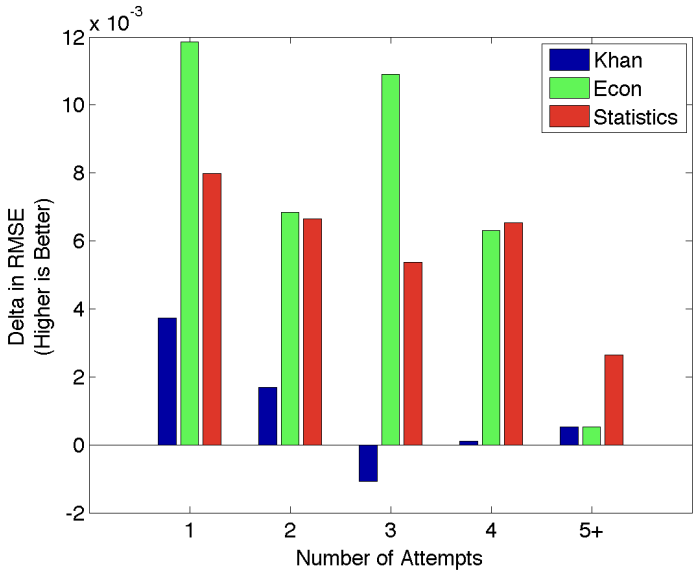

Figure 8: Delta RMSE by total number of attempts in student trace

servations tended to be lower than in resource-exclusive models. Regardless, while the reasons for this relationship are not immediately clear, it may be worthy of further investigation.

Looking at the performance across models on the Nth student attempt, some interesting model properties emerge (see Figure 9). While these relationships are not as strong as those seen with number of resources viewed, particularly in the case of the Khan Academy data, we can still make some observations. As matches the relatively poor performance on longer student traces detailed above, our predictions of students' fifth attempt and beyond is somewhat shaky: this is commensurate with the relatively few attempts seen in the edX data, but somewhat more troubling for Khan Academy, where longer attempt traces were more common. This steady decrease in predictive delta as a student proceeds with the exercise seem to indicate that our method is best when considering short student traces, which may explain the relatively stronger results when considering edX data. This is not entirely unexpected, as learning effects from resources might intuitively be ultimately outweighed by practical experience or influence from external stimuli as a student continues to struggle with a particular exercise.

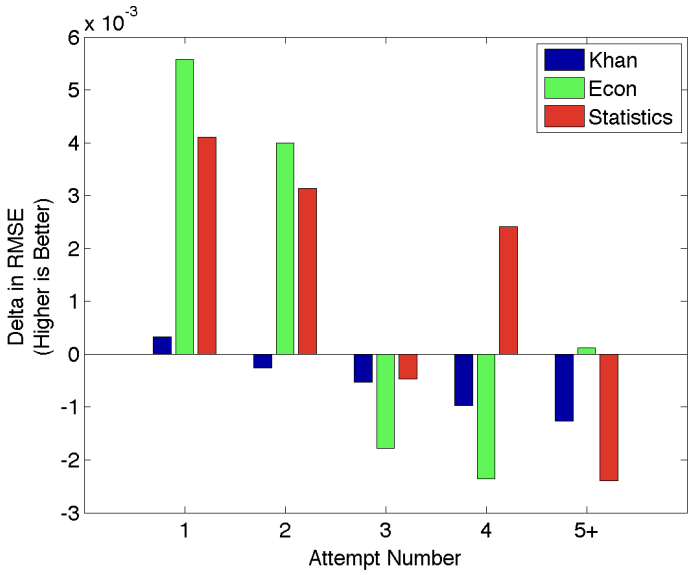

Figure 9: Delta RMSE by Nth individual student attempt

Somewhat unsurprisingly, our analysis showed relatively poor predictive performance in the cases of students who used no or very few course resources. This is likely due to several causes, but one major factor may simply be that, because most students tended to use course resources, the expectation maximization process tended to bias the model toward better fitting the bulk of students. Whatever the reason for the relative dearth of improvement for students that used no resources, the addition of more resources to the trace seems to steadily improve our predictive capacity in the cases of both edX courses (see Figure 10). This is a heartening result, as it does seem to indicate that the presence of resources gave us meaningful information about student behavior.

More inscrutably, students registering the use of precisely one course resource saw the most improvement among models predicting Khan exercises. The reasons for this are not entirely clear; it may be that most of the concepts in Khan academy were simple enough that a single resource access was typically sufficient to grasp the necessary information. Alternatively, it may be the result of a difference in the structure of each set of data: As can be seen in Table 6, users of Khan academy used far fewer resources on average

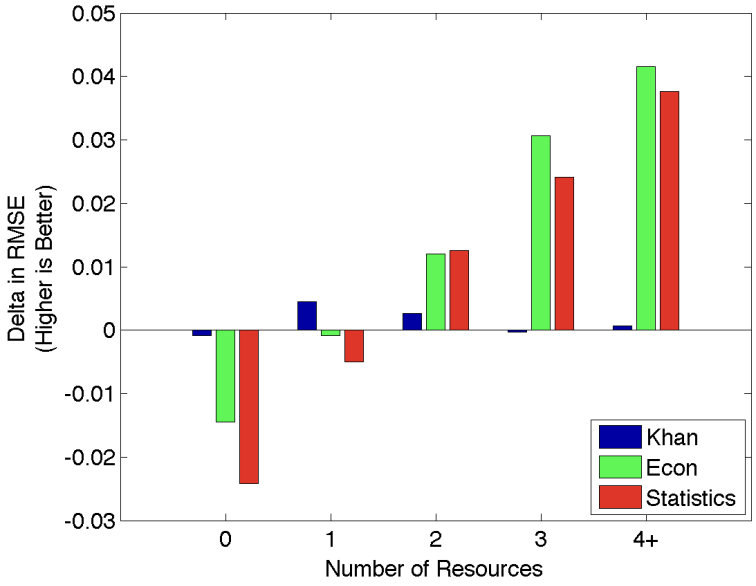

Figure 10: Delta RMSE by number of resources

(0.64) than either of the edX courses, and so the models may have been optimized for students who used fewer resources. The variation in error seen in the edX courses as number of resources observed change is nearly an order of magnitude higher than that seen in the Khan Academy data-set, in either direction, as can be seen in Figure 10.

4.4 Case Studies

In order to gain an intuition why some models were better described by the inclusion of resources and others by excluding resources altogether, we chose to consider the three models in each data-set that performed best under the 'Template-Videos' condition as compared to the baseline, and the three that performed worst. By examining what properties might qualitatively explain the performance of each model, we additionally seek insight into what sorts of videos appear to offer the greatest benefits to student performance.

4.4.1 Khan Academy

Unlike the other two data sets involved in this investigation, the scope of Khan videos is relatively broad. Rather than being a corpus representative of study in a single subject, the Khan data used for this analysis represents work on subjects ranging from basic subtraction to art history to galactic collision. Though work in many subjects was ultimately dropped due to insufficient data, the analysis of the remaining portions of the data covered a broad array of often unrelated subjects. While this scope would pose a problem for a model dependent on accurate knowledge map construction and manual tagging, our reliance on automatic association of videos and assessments meant that our analysis functioned without serious alteration for both the edX and Khan data.

Khan Academy's videos tend to be characteristic of its relatively unique approach to educational video design. Typically the videos affect a relatively informal attitude, with an unseen narrator talking through the theory behind or application of one concept, while illustrating their thought process. Unlike more lecture-oriented videos, Khan's videos tend to be more akin to screencasts, particularly for mathematically oriented concepts, stepping through a problem-solving process while paying relatively less attention to context or historical information.

One feature in particular which sets the Khan data-set apart from the other data included here is the relative abundance of video content. Likely due to the fact that Khan academy is intended as a broad learning resource rather than a single coherent course, relatively little focus is placed on assessment, though quizzes are far from absent. But particularly as concerns subjects which do not lend themselves to the composition of multiple choice questions, there is not a guarantee that any given video or set of videos exists alongside a complementary assessment. Because of this asymmetry, there are typically many candidate pairings between video and assessment; in fact, in many cases an assessment directly references information not found in one, but two or more preceding videos (an example is show in Figure 11).

On the other hand, because the automatic association algorithm used in this analysis uses only notions of log and chronological distance, and does not employ any sort of semantic analysis, the possibility of making spurious associations between content that is used in sequence but ultimately unrelated grows accordingly larger. Though, as can be seen in Table 2, the delta in performance seen across the KCs in the Khan Academy dataset is somewhat

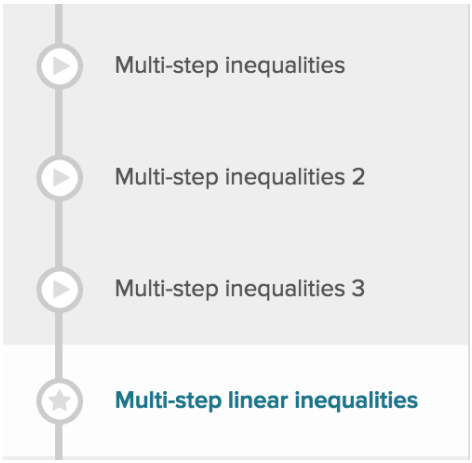

Figure 11: Khan Academy: Some assessments were accompanied by several videos

weaker and more variable than that seen in the edX data, our analysis still yielded usable results.

In fact, despite the possibility of spurious associations, the incidence of such appears to be relatively low. During a manual verification of a random sampling of 100 of the 1096 generated associations, only two spurious associations were found. One related a video on Communism to an assessment testing elementary division, while another related a video about the Bay of Pigs invasion to an assessment testing elementary subtraction. It is not entirely clear why these two associations were formed, other than that users were seen moving from those videos to the corresponding assessments in relatively short order. Possibly this is the result of some educational activity that utilized the Khan platform, but it is unclear; any attempt to explain the association would be speculative. Regardless, the 'KC's generated using these spurious associations did not create significant problems for the analysis. Unsurprisingly, neither the Bay of Pigs-subtraction or Communism-division associations showed any statistically significant change in predictive accuracy when considering or ignoring video observations, indicating a verifiable lack of relationship between video and assessment.

Among those associations which did appear reasonable, there was a wide variation in model performance. In order to better understand this relationship, we chose to manually examine the videos and assessments which

KC Name	Delta RMSE	Resources	P(T|Q)	P(T|V)
Measuring Segments	.024	1	.35	1.00
One Step Equation Intuition	.025	2	.85	0.21
Fundamental Theorem of Arithmetic	.025	2	.74	0.66

Table 7: Khan Academy: Best-performing models

performed best as well as those for which the addition of video observation was actively deleterious.

For all three of the highest performing models in the Khan data, seen in Table 7, the videos showed a striking resemblance to those videos with which they were associated. One particularly compelling example, a video concerning the fundamental theorem of arithmetic, can be seen in Figure 12. Immediately obvious is an aesthetic similarity between the video and the subsequent assessment. In fact, the video explicitly details the manipulation of a bespoke tool designed by Khan academy for that particular exercise, with a narrator stepping through the completion of an example problem nearly identical to the one actually presented to the student. The video does not actually tackle the explanation of the fundamental theorem of arithmetic, merely demonstrates solving a problem posed and resolved by use of a tool exhaustively detailed in the video.

In almost the same fashion, the 'One Step Equation Intuition' exercise involves the use of a unique tool designed for the users of Khan Academy. In this case, users manipulate an animated set of scales, adding and subtracting blocks until the scales balance, with an animation visually indicating the point at which a solution is found. While both videos associated with this exercise ('One Step Equation Intuition' and 'One Step Equation Intuition Introduction') deal directly with the matter at hand, the second of the two, 'Introduction', once more involves the narrator explicitly manipulating the exercise tool which appears in the subsequent exercise.

Likewise the 'Measuring Segments' video, associated with the exercise of the same name, very closely visually mirrors the assessment with which it is associated. Though not detailing the use of a particular tool, the video details the process of solving the following problem, in an environment almost

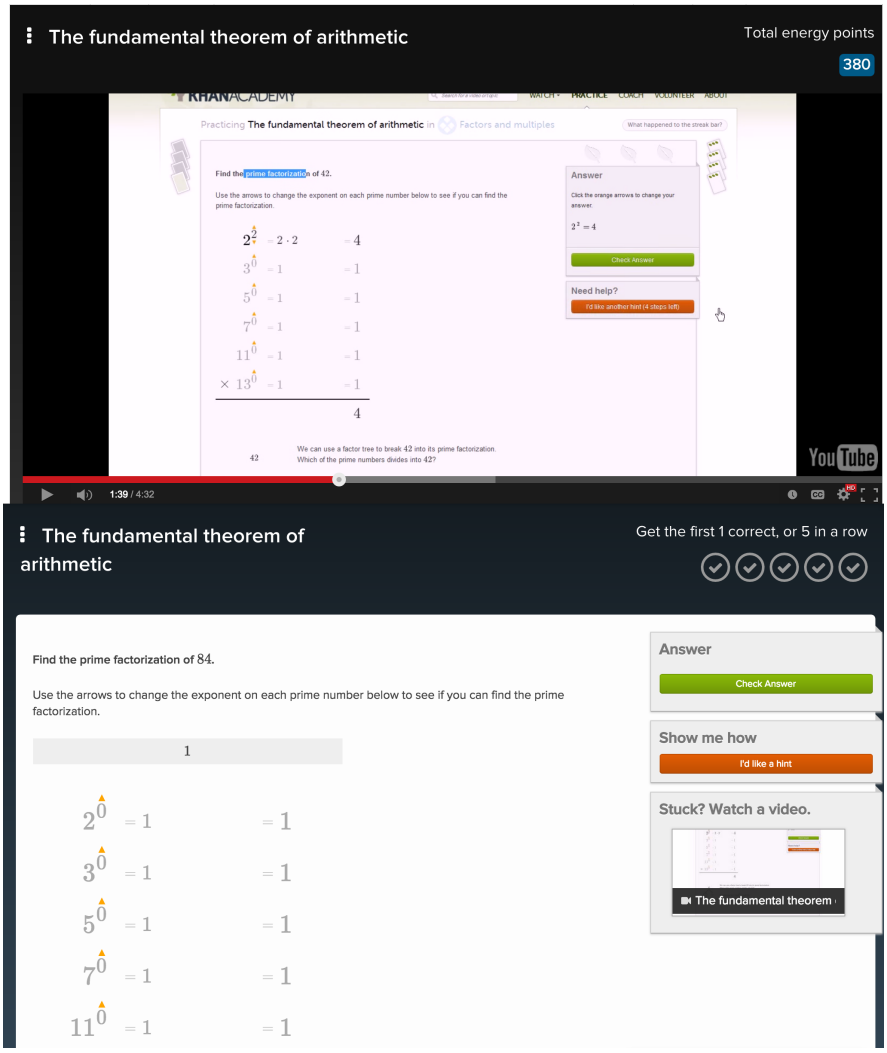

Figure 12: Khan Academy: Above, the 'Fundamental Theorem of Arithmetic' video. Below, the subsequent assessment.

KC Name	Delta RMSE	Resources	P(T|Q)	P(T|V)
Scalar Matrix Multiplication	-.014	2	.62	.17
Direct and Inverse Variation	-.021	5	.02	.10
Balancing Chemical Equations	-.023	2	.20	.17

Table 8: Khan Academy: Worst performing models

identical to the one presented to the student. Perhaps most interestingly, the video is associated with a P(T) of literally 1, which means that the model expects a student who watches the video to solve the next problem with near certainty. While this is a very strong assumption on the part of the model, given the content of the exercise and the associated video, it is not entirely unbelievable.

The commonalities among best performers are relatively obvious. In particular, there is often a strong aesthetic similarity between videos and strongly related assessments. Further, the videos not only convey a concept which students are expected to apply themselves, but walk students through a visually similar process to the one that they will soon be asked to complete. This is actually good evidence that strong associations can be a double-edged sword. That is, though student success is a desirable outcome, knowledge transferable to another domain is ideal. The more directly an assessment mirrors the instructional vehicle, the less demonstrative of transferable knowledge an assessment can be. That said, such obvious links between video and assessment do seem to support the tractability of our hypothesis.

For two of the three lowest performers, the possible sources of model error somewhat mirror the characteristics seen in the highest performing cases. In the case of 'Scalar Matrix Multiplication', for example, the assessment is presented aesthetically differently than the associated video (see Figure 13). In particular, the assessment makes use of custom input fields, which may introduce an additional obstacle to performance to students already struggling to grasp the concept of scalar matrix multiplication itself. Similarly, in the case of 'Balancing Chemical Equations,' (Figure 14) while the most strongly

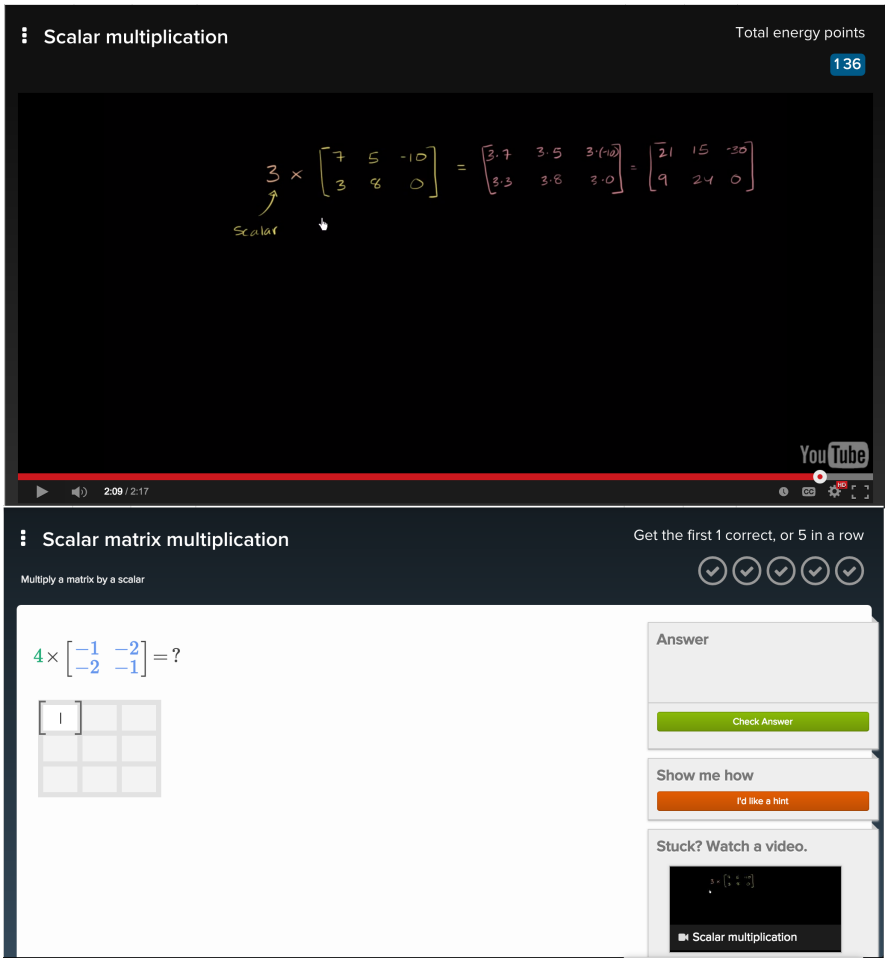

Figure 13: Khan Academy: Above, the 'Scalar Matrix Multiplication' video. Below, the subsequent assessment.

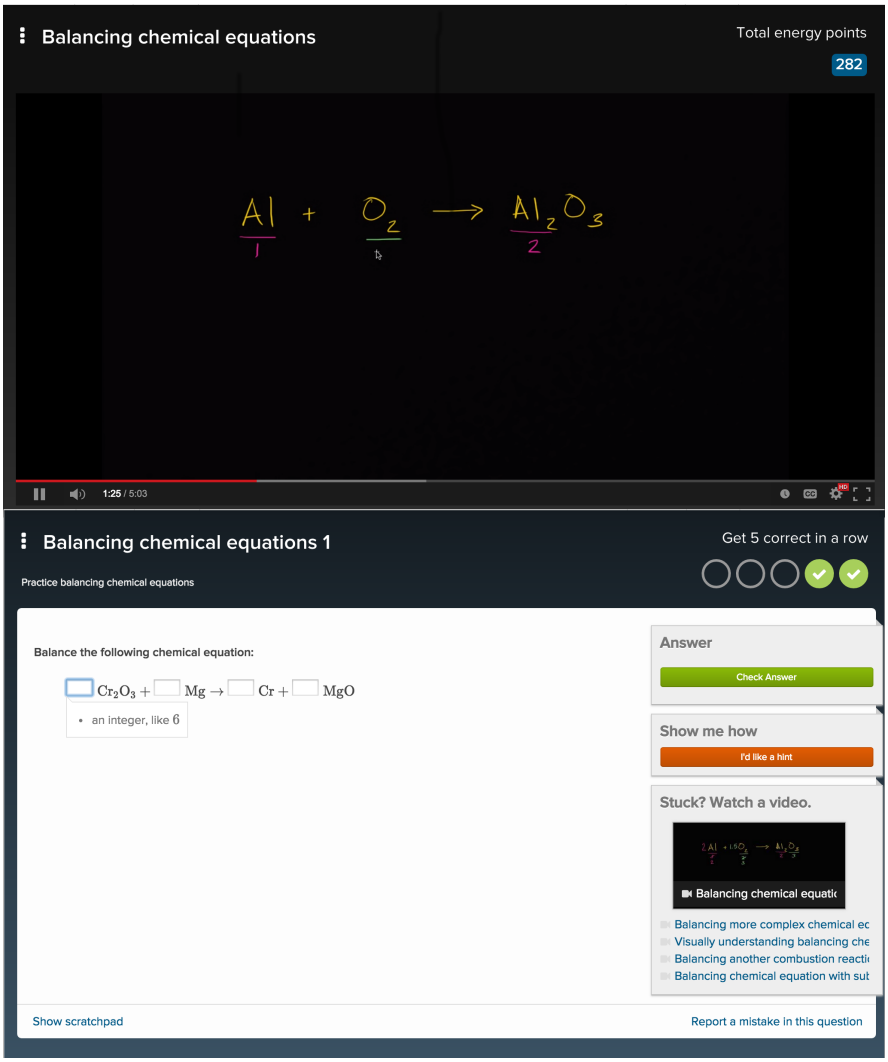

Figure 14: Khan Academy: Above, the 'Balancing Chemical Equations' video. Below, the subsequent assessment.

associated video does walk through the process of balancing chemical equations, the quiz environment may appear somewhat unfamiliar to students, making solving the problems harder than might otherwise have been the case. But more importantly, perhaps, the video details the solution to a relatively simple problem, the combination of single element molecules into a molecule consisting of both elements. While this process should readily be transferable to more complex balancing equations, the process of decomposition and combination of more complex molecules may throw students who have watched a simpler solution for something of a loop. There are actually several videos which each describe an essential part of this process, and the failure of any one particular video to capture the entire learning process may have hindered the model somewhat.

The relationship between the video concerning 'Direct and Inverse Variation' is somewhat less clear, though several potentially complicating observations can be made. First, the problem seems to be a difficult one for students to solve. Not only are the P(T) values associated with both question templates and videos very low (see Table 8), but the prior is also a mere .0719. Typically students only answered the question correctly 53.5% of the time, registering an average of 9.075 attempts, which is on the high end for even the Khan data. Further complicating matters may be that Khan switches freely between inverse and direct variation and inverse and direct proportionality, which may confuse some students who seize on one or the other.

Ultimately, in a way complementary to the most related example seen above, significant dissimilarities between videos and their accompanying assessments seems to contribute to poor model fit. It could be the case that these videos need additional work, or it could simply be that it is difficult to convey such concepts through merely didactic methods. Either way, it does appear that the poor model fit may indeed be indicative of a relatively weak relationship between a video and subsequent content.

4.4.2 edX - Principles of Economics

Both edX courses, Principles of Economics and Statistics in Medicine, differ from the Khan Academy corpus in several key ways. First, the scope of the content included in each course is much more limited: as the edX courses are intended as consistent educational units, their materials and assessments concern student performance in a much more constrained domain. Though the material is still divided into distinct units and subsections, the content is

generally related in some way, either in the sense of pre-requisite and post-requisite relationships, or simply by being under the umbrella of a particular domain.

Second, the format of edX is significantly different from Khan Academy. Since the course, as with most on edX, is intended for consumption as a unit, the quizzes are tracked and aggregated into a grade which accompanies the student. If a student hopes to achieve certification in the course, they must complete a certain number of the quizzes offered, with a passing grade. Further, the two courses were run on a schedule. Students who took each course were required to complete each piece of content by a particular deadline, rather than proceeding at an individual pace.

Finally, unlike Khan academy, which allows for unlimited attempts on questions that are typically templated to allow for repeatability, most edX assessments limit the number of student attempts. This limitation has implications not only for our analysis, but for the application of Bayesian Knowledge Tracing to MOOC data in general; the restricted number of allowed attempts can introduce difficulties when attempting to reason about student growth over time, particularly when identifying KCs at a problem, rather than section or unit level. Despite these differences, our analysis, as well as BKT in general, is still applicable to edX data; it is merely important to note as a caveat that while our analysis remains the same, the characteristics of the two sets of edX data are significantly different than their Khan counterpart.

The Principles of Economics course, true to its name, covers some of the most basic principles of Macroeconomics. The topics covered range from the basic competitive equilibrium model to macro policy issues and international trade, but all concern understanding and applying basic economic theory. Unlike the videos in the Khan data, the Principles of Economics videos much more closely reflect the atmosphere of a traditional undergraduate classroom. Each video is relatively long, with each 'lecture' typically running around twenty minutes, often accompanied by supplemental video material. Visually, the lectures consist of a set of lecture slides, often with Professor Taylor, the course instructor, superimposed in front of them, lecturing.

Course content is generally arranged into sections, each featuring a collection of video, text, and assessment content. Typically composed with a less skewed ratio than comparable topics in Khan Academy, videos available for the class outnumber related assessments by a relatively small margin, making the process of video-assessment association easier and less error prone than

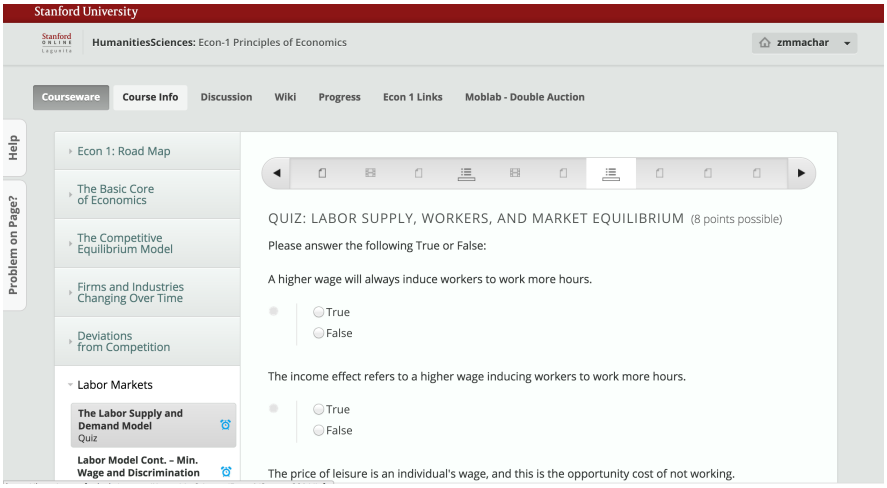

Figure 15: Principles of Economics: An example of the edX interface.

with the Khan data. Usually one or two videos are followed by a text-based summary covering key concepts, which is itself followed by a small quiz to test student comprehension. Assessments consist of multi-part quizzes, considered as knowledge components for the purpose of this analysis, each quiz composed of a number of sub-parts. Individual sub-parts are either multiple choice or value-based free answer. Worth noting here is that we did not consider student use of text-based content. Though there is no obstacle to including such observations, for the sake of simplicity we have chosen to restrict our analysis to video-based content. Future work might explore further the differential quality and usefulness of text-based resources.

As can be seen in Table 3, our analysis of the Principles of Economics data generated stronger results than the data gathered from Khan Academy ($p < .0001$). There are many reasons that might explain why this is the case: it could be that the restricted number of attempts meant that the relative effect was larger, or it could be that the domain lent itself better to video-based resources. Regardless, as with the Khan data, we look to the best and worst performing models to gain an intuition as to what distinguishes them.

Though the Principles of Economics edX course is formatted very differently than the lessons of Khan academy, the distinctions between the best and worst models are similar. The three best performing models (see Table 9) from the Principles of Economics course all concerned KCs drawn from

KC Name	Delta RMSE	Resources	P(T|Q)	P(T|V)
Change in Monetary Policy	.048	2	.35	.17
Monetary Policy	.044	2	.30	.65
Monetary Policy Analysis	.053	2	.03	.50

Table 9: Principles of Economics: Best-performing models

KC Name	Delta RMSE	Resources	P(T|Q)	P(T|V)
Production Possibilities	-.0003	2	.42	.40
Oligopoly	-.0007	1	.49	.05
Productivity and Growth	-.0004	2	.43	.00

Table 10: Principles of Economics: Worst-performing models

one unit of the course, on Macro Economic Policy. All three of these best models are, while less compellingly visually similar than the Khan examples, pointedly related to the subsequent assessments.

Two of the lowest performers (Table 10) told very similar stories. The videos concerning 'Oligopoly', and 'Productivity and Economic Growth' are relatively long, with the video on each topic totaling 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. The other worst performer, 'Production Possibilities and Economic Growth', is one of the first videos in the course, associated with a quiz with nearly a 90% correctness rate. It may be the case that the video offered little additional help, not as a function of video quality, but rather as a result of low assessment difficulty.

4.4.3 edX - Statistics and Medicine

The Statistics in Medicine edX course shares many characteristics with Principles of Economics. Among other similarities, the statistics course features more formal videos, tending to longer durations and featuring undergraduate-style lectures. Further, the course contents were organized such that there were typically one or two videos for each assessment, rendering the association process relatively simple.

KC Name	Delta RMSE	Resources	P(T|Q)	P(T|V)
P-value Pitfalls	.047	2	.59	.54
Comparing Means	.035	3	.64	.36
Exam Question 9	.030	1	.82	.64

Table 11: Statistics in Medicine: Best-performing models

There was, however one significant difference between the two courses. Unlike the Principles of Economics course, students taking Statistics in Medicine were required to complete a final exam, if they hoped to achieve certification in the course. While there were still quizzes associated with videos throughout the course, this capstone assessment stood alone, and included information drawn from most of the previous units in the course. While this did not offer any particular trouble to the association algorithm or the analysis, it did offer an example of one useful feature of the automated algorithm: not just as a tool for judging the relationship between quizzes and videos which are obviously related to them, but also for discovering what resources students sought and found most useful. In fact, as discussed below, one each of the best and worst performing models were videos actually found to be associated with the final exam.

The most effective videos in the Statistics and Medicine course, seen in Table 11, once again nearly directly concern the associated assessment item, though in a way somewhat less visually compelling than their counterparts in the Khan Academy data. Most interesting is that one of the best predicted models is the ninth question on the final exam of the 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 perhaps unsurprising to find that the video is associated with a very strong learn parameter; students who sought out the video tended to do significantly better on the assessment. It is not entirely surprising to see improvement on the final exam: since only one attempt is allowed on each question, one would expect that students who have taken the time to study related material would have a better shot at succeeding on each problem.

The worst models in the Statistics course (Table 12) suffer from problems similar to those seen in the Economics course. The first deals with the second quiz in the course, which involves reading a value from a table. While the

KC Name	Delta RMSE	Resources	P(T|Q)	P(T|V)
Intro to Datasets	-.0005	2	.350	.17
Liner Regression	-.0044	3	.004	.10
Exam Question 21	-.0001	3	.482	.15

Table 12: Statistics in Medicine: Worst-performing models

video does depict that table, it is questionable how well that particular skill might be taught by a video at all. The second model, concerning a quiz on simple linear regression, offered insight into the interpretation and use of simple linear regression, and walked through the interpretation of a certain set of computer-generated regressions results. The associated assessment did indeed concern the interpretation of such a table, but may have confused some students who misunderstood the difference between the table rows. While the video dwells on the interpretation of the intercept of linear regression, all of the distractors ask about the slope: a concept which is not overly difficult, but to which the video may have contributed little understanding.

The last model in the group was the twenty-first question of the final exam. As with the other exam-related model, the data was presented in a way very similar to the most strongly associated video, 'Comparing Proportions Between Two Groups,' but with one key difference. This time, while the data superficially resembled the first example in the video, it actually required a strategy from the second half. Further, the instructor reveals very late in the video the calculation for a two-sided p-value; students who watched only the beginnings of the derivation and went back to the test may have only calculated a one-sided p-value and fallen for a distractor.

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 indicated to an instructor as in need of further attention, and which might be highlighted to students as particularly useful. Though such a metric is by no means a silver bullet solution for managing course content, it does provide a potentially useful and currently lacking metric for understanding the import of various videos on student performance later in the course. It is ultimately up to the instructor how to best use this information, but the more information that

is at their disposal, the more informed the decisions they can make.

5 Future Work

Though we have demonstrated the applicability of video data to BKT analysis, and suggested the utility of examining the model properties of videos thus applied, much work remains to be done. In order to move forward from a proof-of-concept, there are several avenues of development which might be pursued.

5.1 Applications

Essential to any analytical method is moving from theory to actual practice. In particular, there are two applications to which our method might be most appropriate.

5.1.1 Content Recommendation

One opportunity afforded to producers of online educational content, but absent in a traditional context, is the possibility of a dynamically-curated set of recommended materials. That is, by leveraging the massive amount of information available in the context of MOOCs or websites like Khan Academy, it is possible to provide students who are struggling with or simply approaching an assessment for the first time, a data-driven suggestion of content in which to seek additional aid. This would closely mirror the behavior of Bloom's ideal one-on-one tutor, pointing troubled students to materials most appropriate for helping them master material. Data-driven recommendation is hardly a new idea, and present in a number of domains, from online videos, to advertisements, to suggested social contacts and beyond. Even in the field of MOOCs, the idea of a recommender system is not new; however most work has dealt with recommending particularly useful forums posts or entire courses, based on a students' past behavior. Recommendations at the granularity of resources relevant to a particular assessment has been the focus of relatively little research.

The need for such a recommender is not immediately obvious. Most MOOC assessments immediately follow ostensibly related videos, and typically involve applications of very recently learned materials. But this is not

always the case; As we observed in the case of the final exam questions in the Medical Statistics data-set or the asymmetry of assessments and videos in the Khan data-set, finding useful references is not always so direct. Often, there are multiple recent candidate videos or other resources which may be related to a given assessment. Similarly, in the case of capstone assessments, it may not be obvious which portion or portions of a course an assessment is intended to test. To that end, a metric for understanding which resources have been demonstrably the most useful in the completion of a given assessment would provide exactly such a link to students who might need it.

5.1.2 Instructional Design

Another affordance unique to online educational material is the potential for statistically significant evaluation of course materials, even as a course is in progress. While a traditional classroom instructor typically must rely on affective feedback and intuition to iterate on course materials, the breadth of data available to designers of online content offers opportunity to make well-informed decisions about content quality.

As previously discussed, much of the process of iteration for many modern MOOC instructors depends on comparable but potentially misleading input from course forums and student surveys. Ideally, rather than resorting only to sources of affective feedback, instructors would also be privy to some notion of the effect that their course material has had. While it should by no means serve as a replacement for the judgment of the instructors or even for the consideration of affective feedback, a quantitative measure of resource efficacy would be useful in supporting instructors as they support their students.

Of course, one obvious first step to providing instructors this feedback is bundling our analysis in such a way as to render it usable without detailed knowledge of the scripts and platforms on which it depends. Whether this tool be provided to instructors offline, for their own use with data procured on their own machines, or offered directly on the relevant platforms, a usable solution for understanding student use of resources is a sorely-needed feature.

5.2 Extensions to BKT

Though we have taken preliminary steps toward including video information in Bayesian Knowledge Tracing based analysis, there are a number of possible

extensions to work in this domain.

5.2.1 Broadening Scope

An obvious and relatively easy extension to work we have done so far would be to consider resources beyond educational videos. Though we have constrained ourselves to videos for the purposes of keeping our analysis tractable across both the Khan Academy and edX data-sets, the inclusion of other types of resources, like text and interactive content, is a relatively low hurdle. Such an extended analysis would be useful not only as a tool for extending analysis to those resources, but also as a lens through which to compare differential educational impacts of content in different forms. One could imagine comparing the measured relevance of a textual resource to a particular assignment to that of a related video, for example. Such an approach to analysis would not substitute for instructor discretion, but it would ideally be an aid to instructors seeking more information about the success of different types of content, and the value of investing in producing one sort of content over another.

Another extension which would be useful would be to increase the granularity of the analysis of videos. That is, rather than considering video resources on the video level, use overlapping or consecutive segments of the video in the analysis, when data about video usage of that granularity is included. By so doing, one would be able to judge not just which videos were particularly constructive to student success, but which segments of which videos are particularly useful. While most instructors would likely prefer their students to consume course material wholesale, it would be a useful tool to students who are studying for a test, or struggling with a particular question, to have more information about which parts of a video are the most helpful. Further, by highlighting those portions of video that particularly contribute to success, instructors may be able to get a better idea of what kind of content is most useful for their students in their courses. Such an analysis would also account for the amount of time a student has spent on a particular resource, allowing for differentiated learning effects when watching a video in its entirety or just watching a portion. The main hurdle to such an extension would be the granularity and reliability of the log data which informs the analysis, since discrete video interaction events are not always made available by different platforms.

5.2.2 Resource Ordering

One interesting result in recent work by Tang et. al is the effect of item ordering on traditional BKT analysis [30]. It is possible that, in the common case where several resources exist to support student work in a single knowledge component, consuming resources in one particular order may be more useful than another. While this is by no means necessarily the case, and such effects may vary from student to student, it is certainly worth investigating whether our analytic approach might be useful for discovering an optimal path through course material. Dealing more intelligently with repetitive viewings of the same materials may also be a useful refinement. As our analysis stands, we treat each subsequent viewing of the same resource the same way in our model, simply adding an additional observed node to the trace of student activity. It is hardly outside the realm of possibility, however, that there would be differential learning effects related to viewing a video the first time and watching it again. Such an extension would be relatively easy to incorporate into the model, but may suffer from a proliferation of parameters necessary to be trained.

5.2.3 Incorporating Knowledge Structures

As we discussed previously, the theoretical foundations of Bayesian Knowledge Tracing involve a hierarchy of concepts, with previously learned concepts being leveraged to construct understandings of new ones. Though our analysis disregards relationships between knowledge components, both for the sake of simplicity and due to a lack of a canonical knowledge structure underlying each course, it may be useful to incorporate such a structure into future analysis. Either by incorporating performance on prerequisite KCs into the generation of a prior for a post-requisite, or by designing a more complex Bayesian network to account for inter-KC relationships, it may be that considering underlying knowledge structures may improve the performance of our methods.

5.2.4 Incorporating Student Characteristics

Finally, it may be useful to better incorporate individual student characteristics into the analysis. We have done some preliminary testing of conditioning student priors based on student characteristics, but we have, for the sake of simplicity, avoided including such distinctions in our analysis. Intuitively,

if we see that a student has in the past consumed few resources and done well on most exercises, or that a student has voraciously consumed resources and typically still struggles, we would adapt our analytic model to account for those differences. This sort of per-student across-exercise conditioning of student priors could be useful in a real-time application of our methods, but would be particularly appropriate in a post-facto analysis, when clairvoyance about student properties discovered in the running of the course might be useful for better understanding their interactions with materials early on. Such an approach might also involve linking student performance across KCs, rather than treating each student-KC pair as entirely independent.

6 Conclusion

In this paper, we have demonstrated the effect of including video observations in a traditional KT model when applied to large-scale educational data. In so doing we have found our model to give improved results over models that do not include resource information, and helped ameliorate the sometimes negatively correlated effects of resource usage on student performance. Qualitatively, we have found that our results correlate with intuitive expectations of resource performance, giving some evidence that our results are not just statistically meaningful, but may indicate properties of educational content that are interpretable and useful to the humans who design and refine it.

Though the effect size is small, the statistically significant decrease in error under the âĂŹTemplate-1-VideoâĂŹ and Template-Videos conditions is an encouraging sign. It is indicative that, though relatively few resource observations were recorded and many potential video-problem associations were missed or incorrectly made, there is information to be gleaned from a learnerâĂŹs use of educational resources. Further, as suggested by our qualitative investigation of the best and worst performing âĂŹTemplate-1-VideoâĂŹ models when compared to the baseline, it is possible that the delta in accuracy, coupled with the associated P(T) when including resource observations could itself be an interesting metric for evaluating video relevance.

Much work remains to be done to make our methods applicable to educational practice, however. As we have discussed, a number of theoretical extensions to our work may increase the power of our analysis. But regardless of what future analytical work is performed, bridging the gap between a post-facto analysis of large data-sets to a tool useful to instructors currently

designing educational content will require significant engineering. As platforms for the analysis of MOOC data proliferate in the wake of their surging popularity, the design and hosting of such an analytical tool has become significantly less difficult, but as with any research, bridging the gap between theory and practice is an essential step toward the relevance of our analytic approach.

Our methods are not intended to be a substitute for individual instructors' judgment, or for more traditional affective methods of determining the efficacy of course content. Rather, we hope to supplement the information available to instructors struggling with the creation of courses or content intended for an audience whose diversity and scale can make the application of such methods difficult to pursue. The relevance of a particular piece of content to an assessment may or may not be an indication of the quality of that content, often depending on the requirements of the particular course, students, and material. The design of educational content is ultimately a human endeavor and involves decisions best left to the discretion of the instructors themselves. To this end, we hope primarily to support the decisions of such instructors by providing them with more complete information about the performance of their students and the properties of their educational content.

References

[1] S. D. Achtemeier, L. V. Morris, and C. L. Finnegan. Considerations for developing evaluations of online courses. *Journal of Asynchronous Learning Networks*, 7(1):1–13, 2003.

[2] I. E. Allen and J. Seaman. *Changing Course: Ten Years of Tracking Online Education in the United States.* ERIC, 2013.

[3] P. Bell. On the theoretical breadth of design-based research in education. *Educational Psychologist*, 39(4):243–253, 2004.

[4] B. S. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, pages 4–16, 1984.

[5] A. Collins, J. Greeno, L. Resnick, B. Berliner, and R. Calfee. Cognition and learning. *B. Berliner & R. Calfee, Handbook of Educational Psychology, New York: Simon & Shuster MacMillan*, 1992.

[6] A. Collins, D. Joseph, and K. Bielaczyc. Design research: Theoretical and methodological issues. *The Journal of the learning sciences*, 13(1):15–42, 2004.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] M. Koppen and J.-P. Doignon. How to build a knowledge space by querying an expert. *Journal of Mathematical Psychology*, 34(3):311–331, 1990.

[12] K. Kraiger, J. K. Ford, and E. Salas. Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of applied psychology*, 78(2):311, 1993.

[13] 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.

[14] R. V. Lindsey, M. Khajah, and M. C. Mozer. Automatic Discovery of Cognitive Skills to Improve the Prediction of Student Learning.

[15] M. C. Linn, E. A. Davis, P. Bell, and A. P. D. o. M. C. P. Bell. *Internet Environments for Science Education*. Routledge, July 2013.

[16] 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.

[17] D. A. Muller, J. Bewes, M. D. Sharma, and P. Reimann. Saying the wrong thing: Improving learning with multimedia by including misconceptions. *Journal of Computer Assisted Learning*, 24(2):144–155, 2008.

[18] S. Oncu and H. Cakir. Research in online learning environments: Priorities and methodologies. *Computers & Education*, 57(1):1098–1108, Aug. 2011.

[19] Z. A. Pardos, Y. Bergner, D. T. Seaton, and D. E. Pritchard. Adapting Bayesian Knowledge Tracing to a Massive Open Online Course in edX.

[20] 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.

[21] 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.

[22] 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.

[23] A. S. Rosebery, M. Ogonowski, M. DiSchino, and B. Warren. âĂIJThe Coat Traps All Your Body HeatâĂİ: Heterogeneity as Fundamental to Learning. *Journal of the Learning Sciences*, 19(3):322–357, July 2010.

[24] R. S. Russ, V. R. Lee, and B. L. Sherin. Framing in cognitive clinical interviews about intuitive science knowledge: Dynamic student understandings of the discourse interaction. *Science Education*, 96(4):573–599, 2012.

[25] M. Schrepp. A method for the analysis of hierarchical dependencies between items of a questionnaire. *Methods of Psychological Research Online*, 19:43–79, 2003.

[26] M. Schrepp and T. Held. A simulation study concerning the effect of errors on the establishment of knowledge spaces by querying experts. *Journal of Mathematical Psychology*, 39(4):376–382, 1995.

[27] R. J. Shavelson, D. C. Phillips, L. Towne, and M. J. Feuer. On the science of education design studies. *Educational researcher*, 32(1):25–28, 2003.

[28] L. Song, E. S. Singleton, J. R. Hill, and M. H. Koh. Improving online learning: Student perceptions of useful and challenging characteristics. *The internet and higher education*, 7(1):59–70, 2004.

[29] K. Stephens-Martinez, M. A. Hearst, and A. Fox. Monitoring moocs: which information sources do instructors value? In *Proceedings of the first ACM conference on Learning@ scale conference*, pages 79–88. ACM, 2014.

[30] S. Tang, E. McBride, H. Gogel, and Z. A. Pardos. Item ordering effects with qualitative explanations using online adaptive tutoring data. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, pages 313–316. ACM, 2015.

[31] J. F. Van Leeuwe. Item tree analysis. *Nederlands Tijdschrift voor de Psychologie en haar Grensgebieden*, 1974.

[32] M. Wen, D. Yang, and C. P. Rosé. Sentiment analysis in mooc discussion forums: What does it tell us. *Proceedings of Educational Data Mining*, 2014.

[33] 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.

[34] 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.