Understanding Hybrid-MOOC Effectiveness with a Collective Socio-Behavioral Model

Sabina Tomkins
Harvard University
sabina_tomkins@fas.harvard.edu

Lise Getoor
UC Santa Cruz
getoor@ucsc.edu

Online courses for high school students promise the opportunity to bring critical education to youth most at need, bridging gaps which may exist in brick-and-mortar institutions. In this work, we investigate a hybrid Massive Open Online Course for high schoolers which includes an in-person coaching component. We address the efficacy of these courses and the contribution of in-person coaching. We first analyze features of student behavior and their effect on post-test performance and then propose a novel probabilistic model for inferring student success on an AP exam post-test. Our proposed model exploits relationships between students to collectively infer student success. When these relationships are not directly observed, we formulate latent constructs to capture social dynamics of learning. By collectively inferring student success as a function of both unobserved individual characteristics and relational dynamics, we improve predictive performance by up to 6.8% over an SVM model with only observable features. We propose this general socio-behavioral modeling framework as a flexible approach for including unobserved aspects of learning in meaningful ways, in order to better understand and infer student success.

Keywords: latent-variable modeling, collective inference, probabilistic modeling, socio-behavioral modeling, high school, MOOC

1. Introduction

Online courses have become a popular and powerful mode of instruction, enabling access around the world to high-quality education. The deployment of massive open online courses (MOOCs) has primarily targeted college students and adult learners. Now open online education is being deployed to high schools, introducing students to vast amounts of content, and new methods of learning. Even as the popularity of high-school MOOCs increases, their efficacy is debated (Bock and O'Dea, 2013). The subject of this work is an assessment of a single computer science high-school MOOC, as a case study of the applicability of MOOC platforms for high-school students.

To understand the applicability of the MOOC model to high schoolers, we analyze student behavior in a year-long high-school MOOC on Advanced Placement (AP) Computer Science. This course is distinguished from traditional college-level MOOCs in several ways. First, it is a year-long course, while college MOOCs average 8-10 weeks in duration. This provides ample opportunity to observe student interactions for an extended period of time. Secondly, while traditional MOOCs have no physical student-instructor interaction, the high-school MOOC that we consider incorporates instructor intervention in the form of coaching *and* online forum instructor responses. Evaluating the effectiveness of this hybrid model allows us to investigate the effect of human instruction on high-school students, a group which may particularly benefit from supervision. Furthermore, we advance our analysis by introducing a collective socio-behavioral model of student performance.

We introduce a post-test as a comprehensive assessment occurring after the termination of the course. A valid post-test should assess students' knowledge on critical course concepts, such that students' course mastery is reflected in their post-test scores. We treat the Advanced Placement (AP) exam as a post-test and consider students' performance on this test as being indicative of long-term learning. Previous MOOC research evaluates students on course performance (Kennedy et al., 2015). While course performance can be a good metric for evaluating student learning in the short term, post-test performance is a more informative metric for evaluating long-term mastery. Throughout this paper the post-test score is the score a student achieves on the Computer Science AP exam.

Additionally, a post-test is valuable as it provides us with a single metric of student success. Though this should not be interpreted as the only metric, it is useful as an approximation of students' course knowledge. Here, we use the post-test score to evaluate the effectiveness of course content. We also build two predictive models to infer students' post-test scores. In the first model, we use recursive feature analysis to discover the aspects of student behavior most indicative of success. We then build a model which can use these informative features as well as latent student interactions and classroom strength to collectively predict post-test performance. Another highlight of this model is that it allows us to inspect potential discoveries of student collaborations. We propose this model as a general template for educational data mining settings with social interactions and where latent variables can capture unobserved student behavior.

In online environments, social behavior has been shown to affect attrition (Rosé et al., 2014), support knowledge construction (Kellogg et al., 2014; Aviv et al., 2003), and improve course enjoyment (Li et al., 2014). Forum interactions have proven invaluable as a source of information on how students form online relationships, share knowledge, and engage interactively. Critically, forum engagement has shown to be indicative of course success (Tomkins et al., 2016). However, not all interactions are automatically beneficial to student success and more work is necessary to understand how interactions influence course performance, especially in high school settings.

Here, we model two forms of interactions. First, we model students working together on the same assignment. This decision was made in consultation with course instructors who have witnessed students submitting the same assignment code. Second, we use classroom membership in inferring student's overall strength in the course, where we consider strong students to be those who perform well on the AP exam. Thus, we model the phenomenon that students with the same instructor might have similar knowledge and exhibit similar course success.

This paper extends Tomkins et al. (2016) and also introduces several novel contributions. In Tomkins et al., we identified two course success measures: course performance scores and post-test performance scores. We used these success measures to differentiate strong and weak students, to understand which parts of the course are most indicative of success, and to build a predictive model. We also assessed the effect of two important elements of this high-school MOOC: discussion forums and in-person instruction (which we refer to as coaching) on student performance.

Here, we significantly extend this prior work. In addition to the model introduced in our earlier work, we introduce a collective socio-behavioral model, SOCIO-BEHAVIORAL with several advantages for this domain: it can use student interactions to jointly infer success; it can discover useful latent quantities, such as classroom strength and student collaborations; and it encodes complex course dynamics with intuitive logical rules. By discovering latent student and classroom strength, as well as co-working dynamics, we are able to discover unobserved aspects of student behavior while achieving performance superior to our previous model. We propose this framework as a general approach for predicting aspects of student learning in the presence of meaningful latent constructs such as collaboration dynamics. Here, we demonstrate its effectiveness by modeling a potential indicator of collaboration, as well as classroom and student strength.

This paper is organized as follows. In Section 2, we provide an overview of related work. In the following section, Section 3, we discuss the data used in this analysis. In Sections 4-6 , we investigate the relationship between various course features and student performance. This analysis prompts the discovery of unexpected student types, which we introduce in Section 7. To utilize these relationships between students' course behavior and their post-test performance, we build a set of predictive features, which are shown in Section 8. We then develop two predictive models, which we introduce in Section 9. We present an empirical evaluation in Section 10, which we discuss in Section 11, before concluding in Section 12.

2. Related Work

We propose a collective socio-behavioral model for post-test performance prediction in a high-school MOOC. Our work is related to several categories of previous work, including performance prediction in MOOCs, latent factors in student success, peer interactions, and gaming behavior. We discuss each of these in turn below. First, we outline work on predicting performance in MOOC settings. Next, we discuss related work on high school MOOCs. Next, we provide a brief survey of connections between latent factors of student learning and how these relate to various measures of student success. As a critical part of our model is collaborative behavior and student interactions, we discuss related work on peer interactions. Finally, we briefly discuss gaming behavior in MOOCs.

2.1. Predicting performance

There are a variety of approaches for predicting student success in MOOCs. An overview of these approaches is provided in Gardner and Brooks (2018). While there is an abundance of work on predicting completion and dropout, less work has focused on academic performance. Li et al. (2017) predict student achievement from click-stream data. Kennedy et al. (2015) incorporate a variety of student activity features to predict final exam performance.

In our work, we make use of topics discussed in the student forum to predict success. Topic models have been useful in related work as well. Recently, Motz et al. (2018) demonstrate that topics constructed from course titles can reveal latent student interests. Klüsener and Fortenbacher (2015) build profiles of successfull and risky students using features from participation in student forums.

In this work, we predict performance on the AP exam, which differentiates our work in several ways. As this exam is administered by a third party, doing well on it requires that students have mastered relevant concepts, rather than the particularities of the course learning environment. Additionally, this is a high-stakes test which can directly influence students' preparation for and admittance to college. MOOCs offer the potential to improve access to AP courses and their ability to adequately prepare students is an important question in its own right. To the best of our knowledge, we are the first authors to both analyze the success of an AP-preparation MOOC and to build a predictive model for this unique setting.

2.2. High-school MOOCs

There is limited work on analyzing student behavior in high-school MOOCs. Kurhila and Vihavainen (2015) analyze Finnish high school students' behavior in a computer science MOOC to understand whether MOOCs can be used to supplement traditional classroom education. Najafi et al. (2014) perform a study on 29 participating students by splitting them into two groups: one group participating only in the MOOC and another group using a blended-MOOC that has some instructor interactions in addition to the MOOC. They report that students in the blended group showed more persistence in the course, but there was no significant differences between the groups' performance in a post-test.

2.3. Latent student factors and learning

Another aspect of our work is modeling latent factors of student success. The influence of latent psychological factors (e.g., motivation, conscientiousness) on learning outcomes has been explored in several settings. For example, motivation has been related to student performance and learning (Mega et al., 2014). Additionally, many have investigated latent factors of engagement in the context of student enrollment and dropout.

Kizilcec et al. (2013), Anderson et al. (2014), and Ramesh et al. (2014) have developed models for understanding student engagement in online courses. Others have looked at more fine-grained psychological states. Sun et al. (2018) propose a model relating the psychological factors of autonomy, competence, and relatedness to intrinsic motivation. Furthermore, they demonstrate a strong relationship between psychological engagement and behavioral engagement.

Loya et al. (2015) investigate the relationship between conscientiousness and course completion in a computer programming MOOC. They find that behavioral traces indicative of conscientiousness are positively correlated with course completion. This work suggests the potential benefits of incorporating conscientiousness into predictive models of student outcomes.

There has been some work on relating psychological factors to evaluation performance in MOOCs. Hanzaki and Epp (2018) analyze the effect of personality on MOOC grades. They find that including features related to personality can improve the ability of machine learning algorithms to predict course performance. However, Chen et al. (2016) found that personality was correlated with some features of course performance. In line with Loya et al., one of the personality traits with consistent correlations to course behavior was conscientiousness.

We propose a model which includes latent student and section strength. Student strength is a coarse variable which can encapsulate several related psychological concepts, such as motivation and conscientiousness. Here, we demonstrate that even the inclusion of this coarse variable can improve predictions of final exam performance. However, our model is quite flexible, and allows for the inclusion of additional constructs, such as conscientiousness, as relevant data becomes available.

2.4. STUDENT INTERACTIONS AND LEARNING

One component of our work is the inclusion of student interactions in a predictive model of course success. Understanding the social processes of student learning is an active area of research (Anderson, 2003). Recently, Gitinabard et al. (2018) demonstrated that social graph features could improve prediction of MOOC dropout. Andrews-Todd et al. (2018) propose a system for identifying collaboration patterns from course logs, finding four unique groups of collaborators. Importantly, they find that the collaboration profiles correlate with course success. In online courses, student forums are essential in lending insight to how peer behavior impacts motivation, engagement, and other performance metrics (Rosé et al., 2014; Kellogg et al., 2014; Aviv et al., 2003; Huang et al., 2014). Other work has inspected the impacts of physical colocation (Li et al., 2014; Blom et al., 2013) and social search (Su et al., 2016) on learning, the meaningfulness of social networks outside of course forums (Veletsianos et al., 2015), and ritual formation on forums (Longstaff, 2017). Simon et al. (2013) analyze the impact of peer instruction on student learning.

As social interactions are found to have large positive impacts on student experiences, much work has focused on building tools to improve collaborative learning. In their work, Rosé et al. (2008; 2011; 2015) have proposed supportive technologies to enhance forum participation and course collaboration and to automatically analyze both online networks and forum corpora. Additionally, some studies have focused on the effects of peer interactions on learning. Wang et al. (2015) investigated the relationships between types of MOOC forum behavior and learning and found interactions to be indicative of learning gains in some cases. For example, the extent to which a student actively recounts course material on the forum correlates to post-test performance. Wen et al. (2015) analyzed various team aspects in an online course to discover which correlated with performance. Their work demonstrates the potential positive benefits of learning teams, especially when team leaders are active in team building.

We incorporate peer interactions in two ways. First, we model latent section or classroom strength. This allows us to better predict student performance, especially when data is scarce, as we observe some relationship between certain sections and overall post-test performance. Secondly, we infer working-together ties between students. This allows us to explicitly model how peer interactions might influence post-test performance.

Recent work has focused on designing productive team compositions. Staubitz and Meinel (2019) show that mindfully designing teams can reduce MOOC dropout and propose several criteria to assess and design teams. Er et al. (2019) propose a method for forming productive groups by predicting whether students will post in a given group discussion. We hope to incorporate such insights in our setting, to not only understand but also arrive at successful student collaborations.

2.5. GAMING BEHAVIOR

Particularly relevant to our findings is the impact of gaming the system on long-term learning. Baker et al. (2004) investigate the effect of students gaming an intelligent tutor system on post-test performance. In the high-school MOOC setting, we observe similar behavior in some students achieving high course performance, but low post-test performance. We identify plausible ways in which these students might be collaborating in unhealthy ways to achieve high course performance and present analysis that is potentially useful for MOOC designers to prevent this behavior. Additionally, we model how interactions might impact student strength and performance.

In our work, we focus on empirically analyzing different elements of a high-school MOOC that contribute to student learning in an online setting. We use post-test scores to capture student learning in the course and examine the interaction of different modes of course participation with post-test performance. In our work, we infer unobserved offline collaboration through online interactions. These inferred offline interactions are then used in modeling the influence of collaboration on performance. In addition, we model latent student strength. Thus, we can model how interactions between students of varying strength levels can impact performance. Our analysis reveals course design insights that may be helpful to MOOC educators.

3. DATA

The data used in our study is from a two-semester high-school Computer Science MOOC, offered by a for-profit education company. The course prepares students for College Board's Advanced Placement Computer Science A exam and is equivalent to a semester-long college introductory course on computer science. In this work, we consider data from the 2014-2015 school year for which 5692 students were enrolled.

The course is structured by terms, units, and lessons. Lessons provide instruction on a single topic and consist of video lectures and activities. The lessons progress in difficulty beginning with printing output in Java and ending with designing algorithms. Each lesson is accompanied with activities. These activities are not graded; instead students receive credit for attempting them. Students take *assessments* in three forms: assignments, quizzes, and exams. Assignments are generally released every other week, while on alternating weeks a quiz or an exam is released. However, this frequency can vary, for example students might be given an extra week to complete a particularly difficult assignment.

At the end of the year, students take an Advanced Placement (AP) exam. Depending on the criterion of the institution, students can use their AP exam performance as a substitution for a single introductory college course. Simultaneously, students might earn credit from their high school for completing an AP course. The AP exam score ranges from 1 to 5, where a 3.0 is widely considered a passing score. In all, we have data for 1610 students who take the AP exam. This number is a lower limit on the total number of students who may have taken the course and the AP. The course provides a forum service for students, which is staffed with paid course instructors. Approximately 30% of all students who created course accounts also created forum accounts, 1728 students in all.

This course is unique in that it provides a coach service which high schools can purchase. This option requires that the school appoint a coach, who is responsible for overseeing the students at their school. The coach is provided with additional offline resources and has access to a forum exclusive to coaches and course instructors. The average classroom size is approximately 9 students with a standard deviation of approximately 12 students. The largest classroom size coached by a single coach is 72, while some coaches supervise a single student. Of all students who have enrolled in the course, approximately 23% (1290) are coached and 77% (4402) are independent. From here on we refer to the students enrolled with a coach as *coached students*.

We summarize the class statistics in Figure 1 below. The majority of coached students sign up for the student forum, and many persist with the course to take the final AP exam at the end of the year.

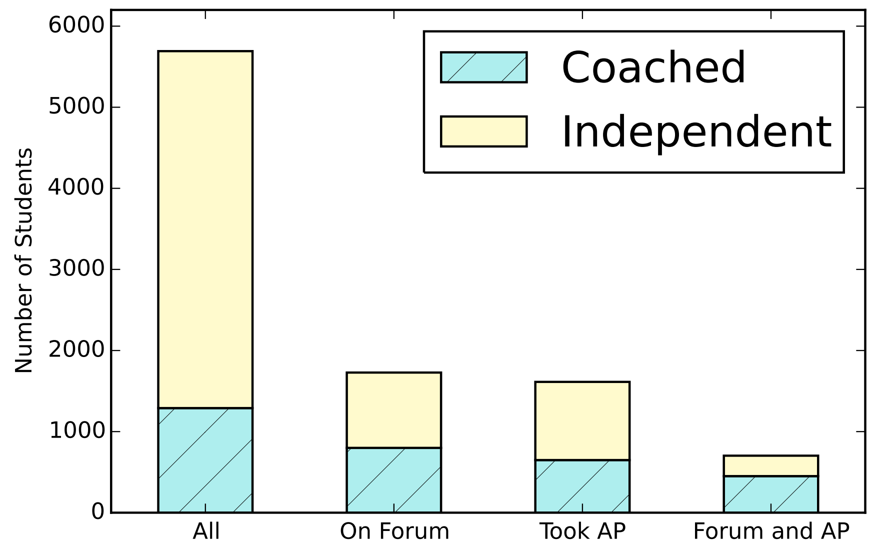

Figure 1: Course participation behaviors for coached and independent students. Participation behavior varies between coached and independent students.

4. Empirically Characterizing Success of a High-School MOOC

In this section, we use post-test performance and course performance to study the effectiveness of MOOCs for high school students. Using an empirical analysis, we provide insights on how to adapt high-school MOOCs to benefit different groups of students. To investigate this question, we focus on the subset of students for whom we have post-test data. To evaluate student success in the course, we identify three measures of course participation in MOOCs that are relevant to the high school population: *overall score*, *course completion*, and *post-test score*.

Overall Score: The overall score captures the combined score across course assignments, quizzes, exams, and activities, each of which contributes to the final score with some weight. We maintain the same weighting as used in the course: exams are weighted most heavily and activities the least. In this grading, we only consider the final score for a given assessment and not the number of attempts taken to complete the assessment.

Overall Score = .3*(Assignment Score + Quiz Score) + .6*Exam Score + .1*Activity Score.

Course Completion: Course completion measures the total number of course activities and assessments completed by the student.

$$\textit{Course Completion} = \frac{\text{Total Activities and Assessments Attempted}}{\text{Total Number of Activities and Assessments}}$$

Post-Test Score: This score captures student scores in the post-test that is conducted two weeks after the end of the course. The score ranges from 1 to 5. This is a scale score based on the composite scoring of multiple-choice and free-response questions. This score captures the advance placement (AP) score; hence we also refer to it as the AP score.

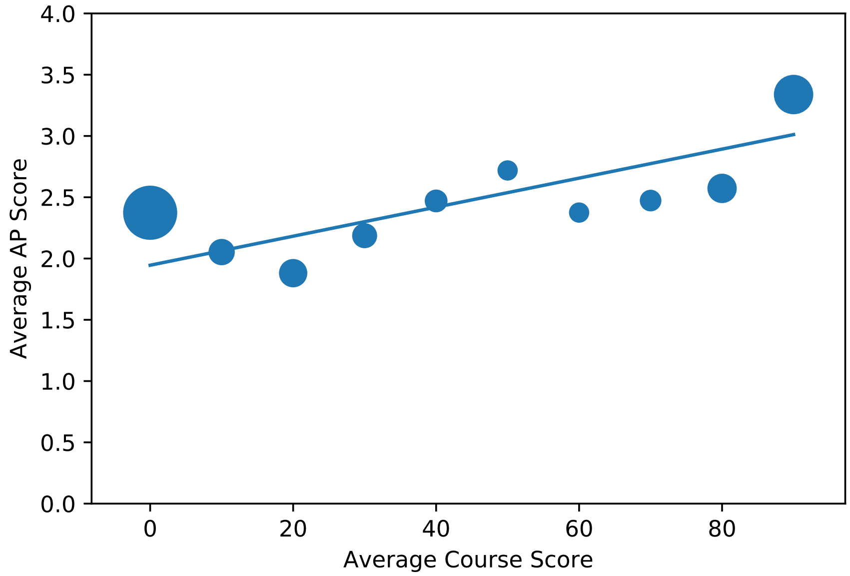

Figure 2: Correlation between average course scores and AP scores, including a regression line summarizing the trend. The dot sizes are proportional to the number of students achieving the overall score.

To evaluate the effectiveness of the high-school MOOC on student performance, we first examine the relationship between course completion and course performance. We hypothesize that as students complete a higher percentage of the course, they should do better in the course assessments leading to higher course performance scores and post-test scores. Examining the correlation of course completion to post-test performance, we find that they are positively correlated. This suggests that the course indeed helps students in achieving good performance in the assessments. However, we find that of the students that achieve an overall score of 90 or greater, only 70% pass the post-test. Similarly, of the students who complete 90% of the course, only 63% pass the post-test. These initial observations indicate the need to perform a more detailed study in order to understand the different student populations in the course.

Next, we examine the relationship between overall score and post-test score, captured in Figure 2. From this plot, we see a positive relationship between course performance and post-test score. To illuminate this we include a line fit to show the relationship between overall score and AP score, Spearman score, $r(1608) = 0.20, p < .001$. Notably, we observe that the average post-test score of the students who achieve a 90% or higher in the course is above a 3.0, and well above a passing score.

The relationship shown in Figure 2 may be distorted by students who only lightly participate in the course. For example, students with high prior knowledge may use the course as a reference/review resource rather than as a primary learning resource. To inspect the relationship between overall course score and AP score for students who are more likely to be using the course as a primary resource, we plot only those students who completed more than 30% of the course in Figure 3. In this setting, we see an even stronger relationship between overall course score and AP score, as illustrated by the steeper line, Spearman score, $r(865) = 0.30, p < .001$.

In both Figure 2 and Figure 3 we see a dip in AP exam performance in the region of 60% to 80%, overall course scores. There may be several reasons for this. One is that the course

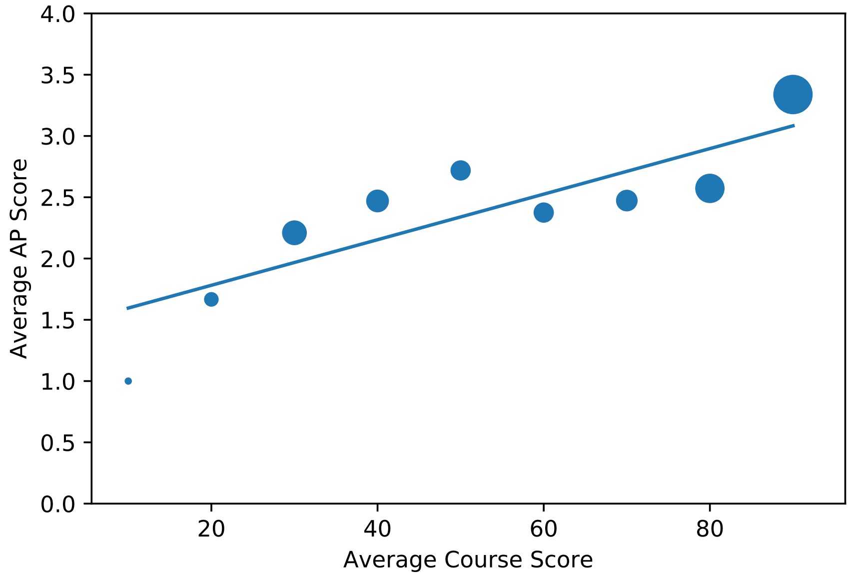

Figure 3: Correlation between average course scores and AP scores, only for those students who complete at least 30% of the course. Again, we include a regression line summarizing the trend and the dot sizes are proportional to the number of students achieving the overall score.

score is not the only indicator of success, as we discuss in the paper, forum participation and participating in ungraded activities are also predictive of success and are not included in the overall course score. It might be that in the region of 60% to 80%, it is possible for students to achieve passing scores on course assessments without mastering the material to an extent sufficient to pass the post-test. This is in contrast to the 90% region, where students who achieve a high score generally pass the post-test. To address this, the course might consider expanding the overall course score to cover more material.

Students regularly complete three kinds of assessments: assignments, quizzes, and exams. Assignments are programming exercises, testing students' coding abilities. Programming assignments are submitted online through an interface capable of compiling programs and displaying error messages. Quizzes are multiple-choice assessments on course material, with an emphasis on recently covered topics. Exams have a similar format to quizzes but are slightly longer. Both quizzes and exams are timed, and students cannot change their answers once they submit them. In all, there are 15 assignments, 8 quizzes and 6 exams in the course. We refer to them as $A_{1:15}$, $Q_{1:8}$, and $E_{1:6}$, in the discussion below.

Figures 4, 5, and 6 present average student assignment, quiz, and exam scores for students who passed/failed the post-test, respectively. We find that students who pass the post-test do better on assessments. We also observe that the scores across all assessments show a decreasing trend as the course progresses. This signals that the assessments get harder for both groups of students as the course progresses. Another important observation is the increase in scores for both groups at assignment 8, quiz 5, and exam 4; these assessments are at the start of the second term in the course, indicating that students may have higher motivation at the start of a term.

Additionally, some assessments show a greater difference between the two groups of students, and performance on these assessments are more informative of student learning. In Figure 6, we observe that for both passed and failed students, we see the greatest dip in performance

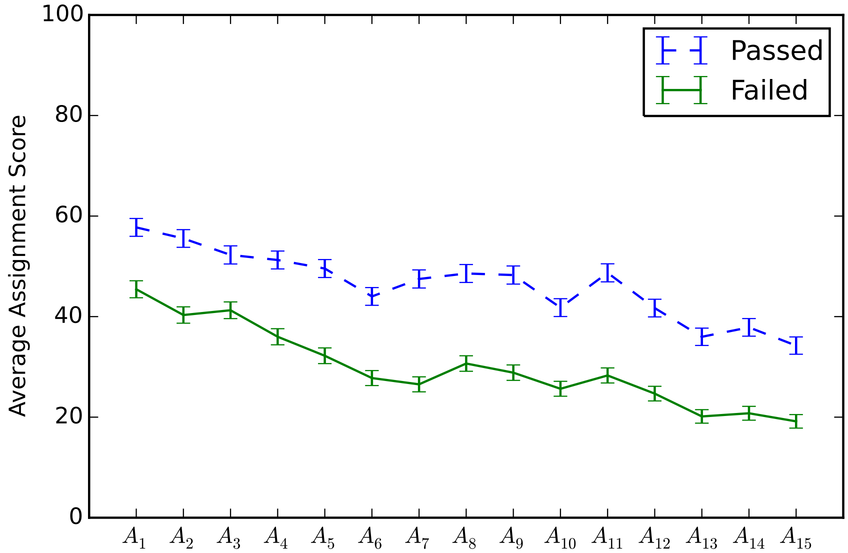

Figure 4: Average assignment scores of passed and failed students. Error bars indicate standard error.

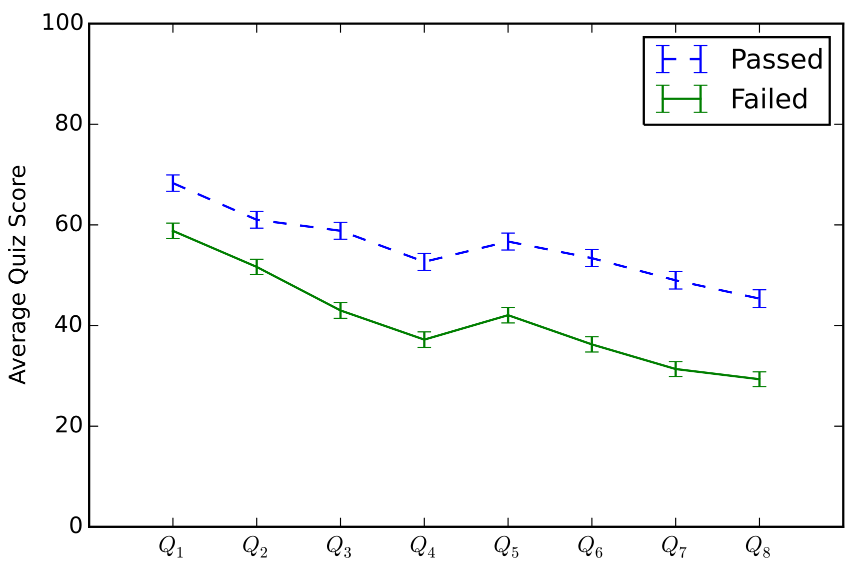

Figure 5: Average quiz scores of passed and failed students. Error bars indicate standard error.

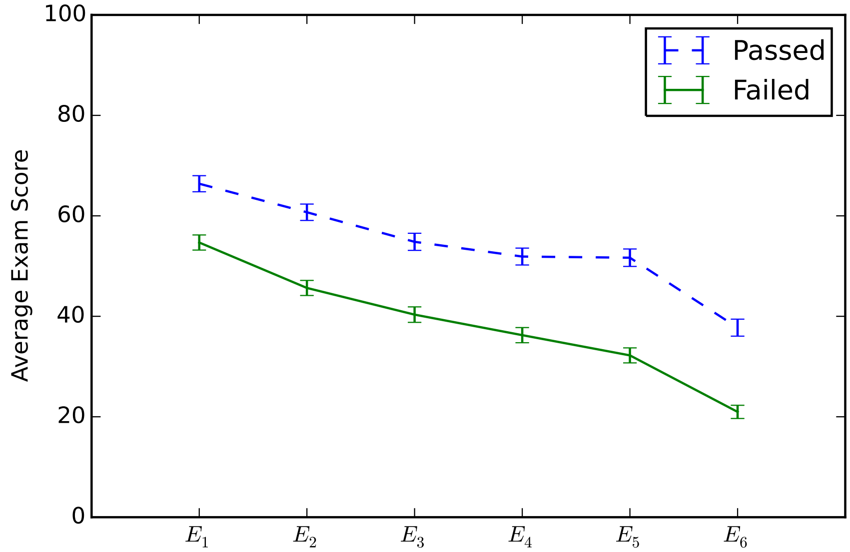

Figure 6: Average exam scores of passed and failed students. Error bars indicate standard error.

in the final exam. As the final exam is the most comprehensive exam, and possibly most related to the post-test, analyzing why students do so poorly on this exam is a worthwhile direction of study in its own right.

Another important dimension is considering assignment completion rate of these two groups of students. In Figure 7, we examine the relationship between attempting assignments and course performance and find that students passing the post-test also attempt more assignments. This suggests that the high scores of these students are not only the product of strong prior knowledge but are also the result of learning from the course.

5. Forum Participation and Post-Test Performance

In this section, we analyze forum participation of students and examine its effect on course success. To do so, we consider the following questions:

- Does participation in forums impact post-test performance and learning?
- What are the key differences between participation styles of students who pass the post-test and students who do not?

The forum is a key feature of this MOOC. Unlike many MOOCs, it is fully staffed by trained instructors who respond to all questions within 24 hours. These instructors are independent of the coaches. Students can interact with coaches in a variety of ways ranging from in-person interactions to email. On the forum, they interact with other students and with the paid instructors who are experts on the course content.

We then look at the average score of students who use the forum compared to the average score of students who do not use the forum. Students who use the forum have a statistically higher post-test performance score of **2.77**, whereas students who do not use the forum obtain a score of **2.34**, $t(1608) = -5.32, p < .001$. It is not clear if the forum impacts learning, or if instead, students with a high desire to learn are more likely to use the forum.

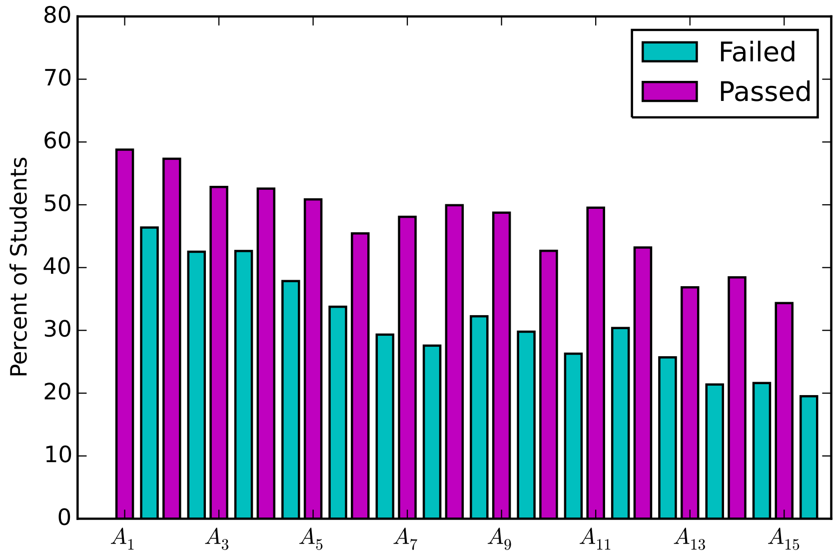

Figure 7: Shows percent of students who attempt assignments. Students who pass are more likely to attempt assignments than students who fail.

To accurately evaluate forum participation of the passing/failing students, we analyze different types of forum participation. Forum participation is comprised of various student interactions: asking questions, answering other student questions, viewing posts, and contributing to conversation threads. The average length of a post is 203 words. The average length of a thread is 2.3. Table 1 compares students who pass the post-test to student who do not across the various forum participation types. The different types of forum participation types are referred to as Questions, Answers, Post Views, and Contributions. We also consider the number of days that a student was logged into the forum, which is denoted by Days Online.

On average, students who pass the post-test make more contributions than students who fail. They also answer more questions. Both groups seem to spend roughly the same amount of time online, to view the same number of posts, and to ask the same number of questions. What most distinguishes a student who passes, from one who fails is whether they are answering questions and contributing to conversations.

Table 1: Forum participation statistics. The average forum participation tends to be significantly more for students that pass the post-test. Behaviors for which there was a significance difference ($p < .05$) between the groups are highlighted in bold.

Forum Behavior	Failed Mean	Passed Mean	Failed Median	Passed Median
Questions	2.95	3.75	0 .00	1.00
Answers	1.32	4.32	0.00	0.00
Post Views	147.54	140.92	73.00	62.00
Contributions	8.66	15.60	1.00	2.00
Days Online	19.46	21.50	11.00	13.00

This analysis further supports the importance of forums to MOOCs. Answering questions and contributing to conversations are two behaviors indicative of strong post-test performance. However, it is unclear if answering questions helps students learn, or if students with high prior knowledge are more likely to answer questions and contribute to conversations. We hope that MOOC designers can use this information to create appropriate intervention and incentive strategies for students.

6. Coaching

In this section, we evaluate the effect of coaching on student learning. We compare coached students to independent students using their participation in course assessments and forums. We further analyze the subset of students who have only one coach in order to isolate the effect of coaching from other classroom effects.

6.1. Course Behavior

We inspect the average assessment scores of coached and independent students in Figures 8-10. Observing scores across assignments, quizzes, and exams in Figures 8, 9, and 10, respectively, we find that coached students perform better than independent students across all assessments.

Such differentially high performance in the course should indicate higher performance in the AP exam for coached students. However, this is not reflected in the post-test scores. The average post-test score for coached students is 2.43, while it is 2.59 for independent students, where the differences between these average scores is significant, $t(1608) = 2.01, p < .05$.

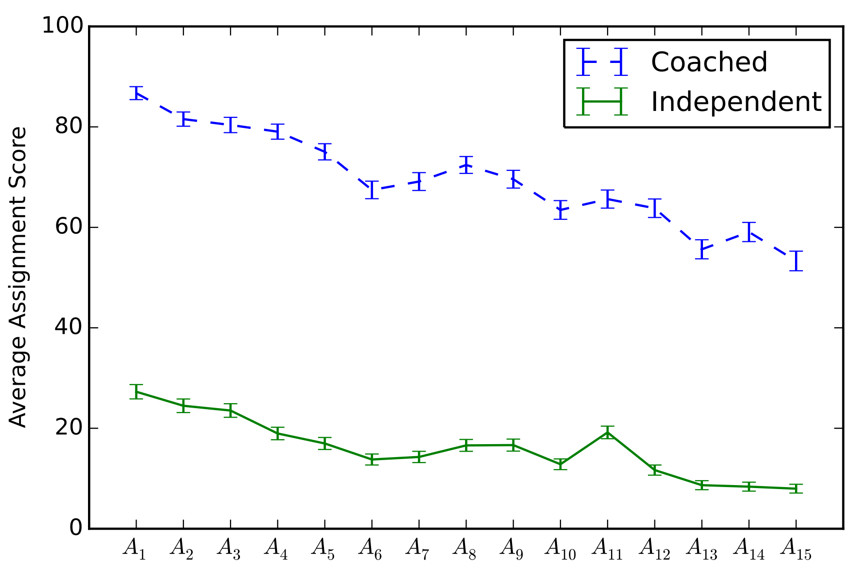

Figure 8: Average assignment scores of coached and independent students.

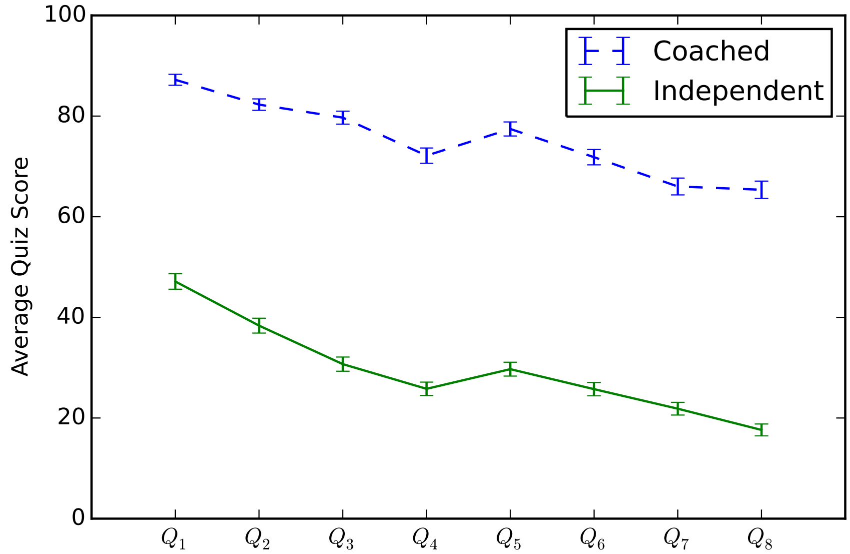

Figure 9: Average quiz scores of coached and independent students.

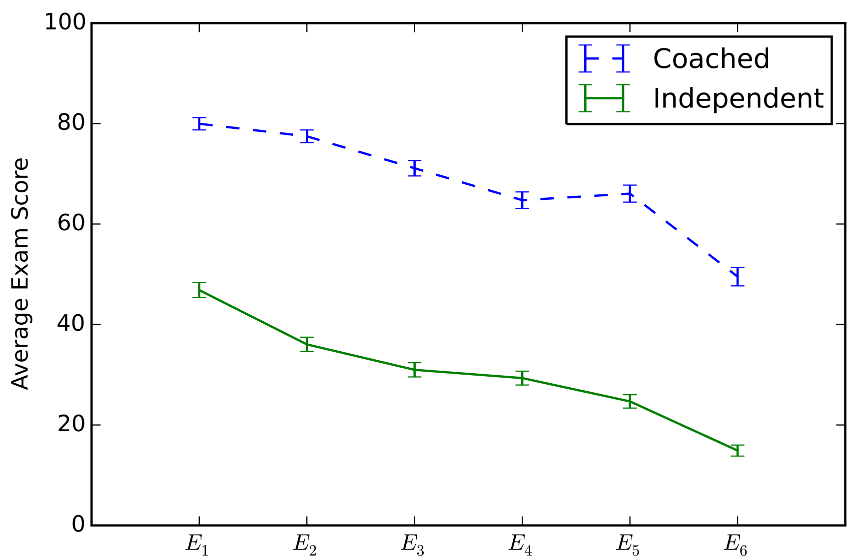

Figure 10: Average exam scores of coached and independent students.

Table 2: Forum participation statistics for coached and independent students. Coached students view more posts and ask more questions. The behavior for which there was a significant difference between the groups are highlighted in bold ($p < .05$).

Forum Behavior	Coached Mean	Independent Mean	Global Mean
Questions	**2.81**	**1.90**	2.32
Answers	1.45	1.72	1.59
Post Views	**145.49**	**81.50**	111.11
Contributions	8.10	7.33	7.69
Days Online	**20.64**	**12.55**	16.29

6.2. Forum Participation of Coached and Independent Students

In order to understand this difference in scores, we analyze forum participation of coached and independent students. This analysis reveals a significant difference in forum participation between coached and independent students. Table 2 gives the comparison between coached and independent students in forum participation. On average, coached students ask more questions and answer fewer questions on the forums when compared to independent students. Coached students exhibit more passive behavior by predominantly viewing posts rather than writing posts compared to independent students. This can be particularly dangerous if the posts which are viewed contain assignment code.

In Table 3, we compare coached students who pass to coached students who fail and see the same differences as those observed between all students who pass, and all students who fail. Students who pass are more likely to answer questions and contribute to conversations.

6.3. Coaches with Only One Student

To examine the effect of coaching class size on coached students' post-test performance, we examine coached students with no classmates, that is, when there are coaches with whom only one student registers[1]. Comparing average post-test scores of coached students who are singly advised by their coaches (classroom size of one) with independent students, we find that the average post-test score for the coached students is 3.6, while it is 3.2 for independent students.

Table 3: Forum participation statistics for coached students. The differences in forum behavior between coached students who pass and who fail follow the same trends in forum behavior exhibited by the general population, and shown in Section 5. The behavioral features for which there was a significant difference between the groups are highlighted in bold ($p < .05$).

Forum Behavior	Passed Mean Coached	Failed Mean Coached
Questions	3.97	2.87
Answers	**3.04**	**0.56**
Post Views	141.56	164.14
Contributions	**14.19**	**5.93**
Days Online	22.71	21.53

[1]Other students might register independently at the school and pursuing the course on their own have no record of being coached.

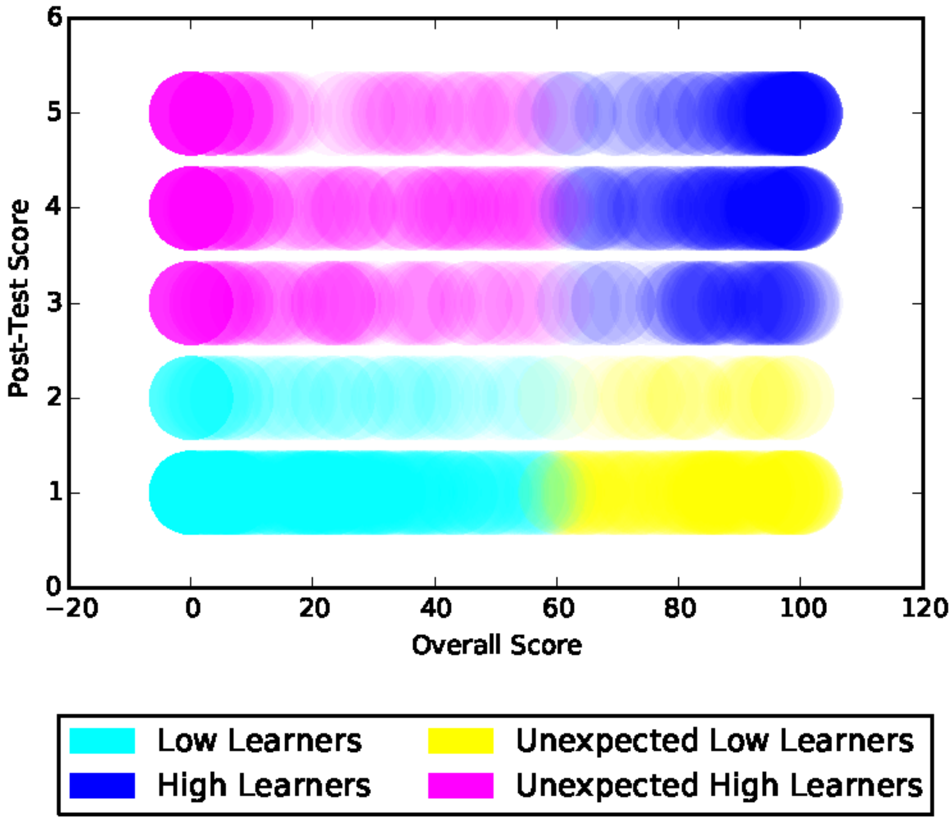

Figure 11: Grouping of students into four groups: low learners, high learners, unexpected low and high learners.

This analysis suggests that the effect of coaching is confounded by the effects of learning in a classroom with peers. To fully understand the effect of a coach guiding a student through the learning process, the peer-effects of classmates should be better understood and isolated. In Section 7, we take first steps in this direction by proposing student types.

7. Inspecting Unexpected Student Types

In this section, we identify and analyze various types of students in the course based on their performance in the assessments. We classify students into two broad types based on whether the overall scores and post-test scores are correlated. Overall scores were partitioned into low (0-60%) and high groups (61%-100%), and post-test scores were likewise partitioned into low (1-2) and high groups (3-5). Figure 11 gives the relationship between the overall score and post test score for all students. Two groups of students emerge, students who exhibit a correlation between overall scores and post test scores, and students who do not. These two groups can be further broken down based on whether they obtain a high score on the post test, yielding four groups of students.

- *Low learners*: These students have low values for both overall scores and post test scores.
- *High learners*: These students obtain high values for both overall scores and post test scores.

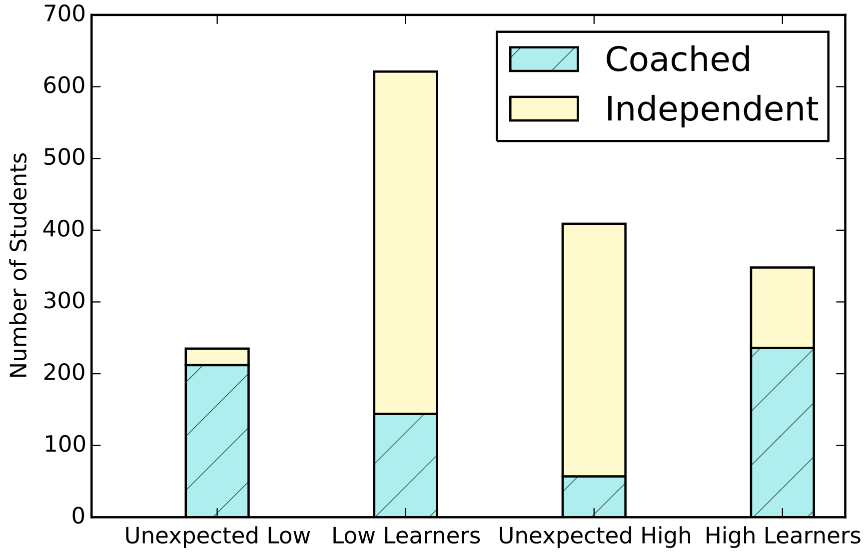

Figure 12: Number of students in each group. The majority of unexpected low learners are coached, while the majority of unexpected high learners are independent.

- *Unexpected low learners*: These students obtain high overall scores, but low post test scores.
- *Unexpected high learners*: These students obtain high post test scores, but low overall scores.

Among these, the unexpected low learners and unexpected high learners deviate from the rest of the students. To analyze these two groups, we delve deeper into other aspects of the course such as forum participation and coaching.

7.1. Unexpected Low Learners

Unexpected low learners are those students who perform well on the course assessments (with an overall score of over 60%) but who do not earn a passing post-test score. We hypothesize that this might be due to their not retaining information from the course, or not arriving at high overall course scores on their own. To understand their low post-test performance, we examine their forum behavior and coaching environment.

As can be seen in Figure 12, approximately 91% of unexpected low learners are coached students. Most of these students are part of large classrooms coached by the same coach. While it is unclear why the majority of this group would be coached students, one possible explanation is that classroom dynamics influence learning in potentially negative ways. For example, working together might result in unforeseen negative consequences. If students who do not fully understand the material achieve high course scores due to the help of their peers this help may not prepare them to pass the post-test.

Further, analyzing forum performance, we find that approximately 76% of unexpected low learners use the forum. Of those who use the forum, 91% are coached. Table 4 gives the forum participation of coached and independent unexpected low learners. The forum participation of these students has a strong similarity to failing students in Table 1, participating passively in

Table 4: Forum participation statistics. Forum behaviors for which there is a significant difference ($p < .05$) between groups are highlighted in bold.

Forum Behavior	Coached Mean	Independent Mean
Questions	3.46	9.19
Answers	**0.54**	**15.31**
Post views	195.52	293.69
Contributions	**7.11**	**67.31**
Days Online	25.59	35.19

the course by viewing forum posts and contributing to fewer answers. The coached students are less active than the independent students on the forum in every way, even in post views. While it has been posited before that active forum participation is indicative of learning and high AP exam performance, this may not be the case in all groups. For example, the small number of independent students may be using the forum for social, rather than learning purposes.

7.2. Unexpected High Learners

Unexpected high learners earn an overall course score of less than 60% but still pass the AP exam. Approximately 86% (357 out of 409) of unexpected high leaners are independent, and approximately 80% of the unexpected high learners (323 out of 409) are not on the forums, so descriptive statistics of forum participation are omitted. That this group can do so well on the post test, without either a high amount of course or forum participation strongly suggests that either these students have prior knowledge in computer science or that they are not being primarily exposed to computer science through this course but are instead using it to supplement another mode of instruction. A pre-test of students' prior computer science knowledge would provide further clarity.

8. Modeling Student Behavior

In this section, we build models of student behavior which can inform predictions of their performance. For example, we saw that students' post-test performance is associated with their course and forum behavior in Section 4 and Section 5. Here, we investigate *which* aspects of student behavior are most indicative of post-test performance. By discovering the relative rank of the student model features, we draw insights about student behavior relevant to learning, and to course design.

8.1. Student Model Features

We group the course features from student interactions into four broad categories: 1) course behavior, 2) forum behavior, 3) coaching environment, and 4) topic analysis of forum posts. We extract features from student course behavior and forum behavior; two other feature categories are described below.

Table 5: Coaching related features

Feature	Explanation
Coached	Boolean feature capturing whether a student is coached or independent
Coach Views	# posts viewed by the coach
Coach Questions	# questions posted by the coach
Coach Answers	# answers posted by the coach
Coach Contributions	# contributions in the forum

8.1.1. Coaching Environment

Here, rather than modeling student features, we model features related to coaches. Coaches are provided a separate discussion forum, apart from the student forum, where they can interact with other coaches and instructors of the course. We define features that may capture coaches' prior knowledge and their involvement in guiding students. Table 5 gives the list of coaching related features derived from the discussion forum for coaches.

8.1.2. Posts Topic Distribution

In order to better understand online forum behavior, we make use of topic models. Topic models are popular approaches to using the distribution of words appearing across posts to identify common themes. For example, in an online course, one might uncover themes of course-related concepts. Here, we use Latent Dirichlet Allocation (LDA; Blei et al., 2003) implemented in the Machine Learning for Language Toolkit (MALLET; McCallum, 2002). Before using LDA we clean the text data by removing stop words, stemming certain words, and removing all common course words, such as the word *code*. We use the following parameters for the topic model: number of topics = 150 and optimize-interval = 100, where the hyper-parameters required by LDA, α and β, are set to the default values. All discovered topics are included in an initial feature set which is later reduced using feature elimination as described in the next section. That is, for a given student we can model the extent to which they posted about a given topic as their total topic score for each topic across all posts. We choose a large number of initial topics as those which are uninformative will be lost in the feature elimination stage.

9. Predictive Models

Here we discuss the goal of predicting student post-test performance. To do so, we introduce two types of predictive models. The first employs the features detailed in Section 8. The next is able to reason collectively, using relations between students to better infer performance.

In both types of models, we use only the most predictive features. A subset of features that are predictive of post-test performance was selected using recursive feature elimination in scikit-learn (Pedregosa et al., 2011). Recursive feature elimination works by training a classifier which weighs features and then trims all features with the lowest weights; this trimming allowed us to obtain the best predictions, and to understand which features are most predictive of student success.

9.1. STANDARD PREDICTIVE MODELS

We incorporate extracted features in a linear kernel Support Vector Machines (SVM) model, using the Python package scikit-learn (Pedregosa et al., 2011). Comparing this model with other machine learning algorithms such as logistic regression, decision trees, and Naive Bayes, we found the results to be comparable. However, a failing of this model is its inability to reason collectively about student behavior.

9.2. COLLECTIVE SOCIO-BEHAVIORAL MODELS

In order to model the intricate dependencies between students in this online course we employ a collective probabilistic approach. An advantage of a collective approach is that while predicting performance, we can jointly infer the values to descriptive latent variables. For example, here we model student and classroom section strength as latent variables, in addition to unobserved collaborations between students. To do so, we use Probabilistic Soft Logic (PSL; Bach et al., 2017), a probabilistic programming framework[2] which allows us to encode student interactions and their potential effects with intuitive logical rules.

In PSL, domain knowledge is encoded with weighted logical rules. The values to these weights are learned in a training process, and thus the relative importance of each rule is informed by data. Incorporating multiple rules of varying strengths allows us to fuse multiple sources of information to varying degrees. These rules can describe complex relationships and, crucially, capture dependencies not only from observed features to target variables but *between* target variables. This expressivity allows us to encode relationships between students. Finally, PSL provides an intuitive framework for representing latent abstractions and an efficient procedure for inferring their values.

To illustrate PSL in the online course context, consider a rule which says that if two students exhibit similar course performance and one passes the post-test, the other will be likely to as well. To express this rule we introduce the predicate SIMILARCOURSEPERFORMANCE, which takes two students as arguments and which expresses their course performance similarity as a value between 0 and 1. Additionally, we introduce the predicate PASS, which takes a student as an argument and whose truth value indicates whether this student passes the post-test. With these predicates, we define our rule in PSL as follows:

$$w_{sim} : \textsc{SimilarCoursePerformance}(S_a, S_b) \wedge \textsc{Pass}(S_a) \Rightarrow \textsc{Pass}(S_b)$$

9.2.1. Student strength

Next, we demonstrate how PSL can be used to template a predictive model for post-test performance. For each student, S_i, we would like to predict if this student will pass or not; that is, we would like to infer the truth value of $\textsc{Pass}(S_i)$. We would also like to assign each student a strength value, such that strong students are likely to pass. To do so, we introduce $\textsc{Type}(S_i, T)$, where T is either *Strong* or *Weak*. To learn the latent strength of each student we use the student models developed in Section 8. To express that a student's strengths are related, i.e., a strong student cannot simultaneously be a weak student, we impose a functional constraint. This constraint ensures truth values of all potential types for a given student sum to 1 (in Ruleset 1).

[2]Open-source software available at: http://psl.linqs.org

$$\infty : \sum_{t \in Types} \textsc{StudentStrength}(S_i, t) = 1$$

Ruleset 1: Constraint on student strength.

$$\begin{aligned}
w_{svms} &: \textsc{SVMPredicts}(S) \Rightarrow \textsc{Type}(S, \textit{Strong}) \\
w_{svmw} &: \textsc{SVMPredicts}(S) \Rightarrow \textsc{Type}(S, \textit{Weak}) \\
w_{t_j} &: \textsc{PostTopic}(P, t_k) \wedge \textsc{Author}(S, P) \Rightarrow \textsc{Type}(S, \textit{Strong}) \\
w_{a_h} &: \textsc{Answers}(S, high) \Rightarrow \textsc{Type}(S, \textit{Strong}) \\
w_{a_l} &: \textsc{Answers}(S, low) \Rightarrow \textsc{Type}(S, \textit{Weak}) \\
w_{c_h} &: \textsc{Contributions}(S, high) \Rightarrow \textsc{Type}(S, \textit{Strong}) \\
w_{c_l} &: \textsc{Contributions}(S, low) \Rightarrow \textsc{Type}(S, \textit{Weak}) \\
w_{v_h} &: \textsc{Views}(S, low) \Rightarrow \textsc{Type}(S, \textit{Strong})
\end{aligned}$$

Ruleset 2: Inferring Student Strength

We make use of the predictions from the model explained in the SVM model to infer students' latent strengths. In the first two rules in Table 2, we learn weights to express the dependence between the SVM predictions and students' types. We also use the topics of the posts each student contributes to the student forum. For each topic which is considered by feature selection to be predictive, we learn a weight for the rule which states that mentioning this topic is indicative of a student being *Strong*. This is shown with the rule template shown in the third line of Table 2, where the value of $\textsc{PostTopic}(P, t_k)$ is the topic distribution assignment of topic t_k and $\textsc{Author}(S_i, P)$ is 1 if S_i authored post P. Similarly, we would like to express the relationship between forum contributions and student strength. To do so, we introduce the predicates $\textsc{Answers}(S_i, A)$, $\textsc{Contributions}(S_i, A)$ and $\textsc{Views}(S_i, A)$, where A can be either low or $high$. A variable representing the number of questions a student asks was deemed unimportant by the RFE procedure. We then express the relationships between these behaviors and strength with the remaining rules in Ruleset 2.

9.2.2. Section strength

We are also interested in how membership in particular classrooms might impact performance. To express the strengths of various learning environments, we introduce $\textsc{SectionStrength}(C, T)$, where C refers to a class or section ID, and T is either *Strong* or *Weak*. As for students, we constrain sections' types with the rule in Ruleset 3, which expresses that a section cannot be simultaneously *Strong* and *Weak*.

$$\infty : \sum_{t \in Types} \textsc{SectionStrength}(C, t) = 1$$

Ruleset 3: Constraint on section strength

$$
\begin{aligned}
w_{cah} &: \textsc{CoachAnswers}(C, high) \Rightarrow \textsc{SectionStrength}(C, \textit{Strong}) \\
w_{cao} &: \textsc{CoachOnline}(C, low) \Rightarrow \textsc{SectionStrength}(C, \textit{Weak}) \\
w_{ccs} &: \textsc{CoachClassSize}(C, low) \Rightarrow \textsc{SectionStrength}(C, \textit{Strong})
\end{aligned}
$$

Ruleset 4: Inferring Section Strength

To infer section strength, we use features discovered in the RFE procedure. In addition to the student forum, the course includes a forum where coaches can interact with each other and with course instructors. To model the behavior of coaches on this forum, we introduce the predicates $\textsc{CoachAnswers}(C, A)$ and $\textsc{CoachOnline}(C, A)$, where A can be either low or $high$. $\textsc{CoachAnswers}$ refers to the number of questions a coach answers on the coach forum, and $\textsc{CoachOnline}$ refers to the number of days the coach logs into the coach forum. Additionally, we consider the number of students in a classroom with $\textsc{CoachClassSize}(C, A)$. The rules in Ruleset 4 relate coach forum behavior to section strength.

$$
\begin{aligned}
w_{secs} &: \textsc{SectionStrength}(C, \textit{Strong}) \wedge \textsc{InSection}(S, C) \Rightarrow \textsc{Type}(S, \textit{Strong}) \\
w_{col} &: \textsc{SameSection}(S_i, S_j) \wedge \textsc{Type}(S_i, T) \Rightarrow \textsc{Type}(S_j, T) \\
w_{col_s} &: \textsc{InSection}(S_i, C) \wedge \textsc{InSection}(S_j, C) \wedge \textsc{SectionStrength}(C, T) \wedge \textsc{Type}(S_i, T) \\
&\qquad \Rightarrow \textsc{Type}(S_j, T)
\end{aligned}
$$

Ruleset 5: Section Strength and Student Strength

We then use section strength to inform the predictions of students' strengths. This is shown in Ruleset 5. This rule expresses that if a student is enrolled in a strong section, they may also be a strong student. The next rule in this table expresses that students in the same section will have similar strengths. The final rule restates this but modulates this dynamic by section strength.

For each rule in Ruleset 2, we also have a rule which propagates these student behaviors into section strengths. These rules express that if a student is predicted to be strong and is enrolled in a given section, then that section should also be strong. We make this clear with the following rule template, where an example behavior correlated with strength is answering questions on the student forum:

$$
\textsc{Behavior}(S) \wedge \textsc{BehaviorCorrelated}(T) \wedge \textsc{InSection}(S, C) \Rightarrow \textsc{SectionStrength}(C, T)
$$

Thus, each rule in Ruleset 2 that infers student strength has a version following this template which instead infers section strength.

9.2.3. Classroom style and student strength

Additionally, we observed that coached and independent students exhibited different behaviors on the forum. Moreover, these behaviors are differently correlated with success. For example, when coached students view a high number of posts, this behavior is associated with low performance. However, when independent students view a high number of posts, it is correlated with high performance. In order to capture these differences, we introduce two predicates $\textsc{Coached}(S)$ and $\textsc{Independent}(S)$ where each return binary values according to whether a

$$
\begin{array}{|l|}
\hline
w_{ec_x} : \textsc{AssessmentScore}(S, A_x) \wedge \textsc{Coached}(S) \Rightarrow \textsc{Type}(S, \textit{Strong}) \\
w_{ei_x} : \textsc{AssessmentScore}(S, A_x) \wedge \textsc{Independent}(S) \Rightarrow \textsc{Type}(S, \textit{Strong}) \\
w_{v_hc} : \textsc{Views}(S, high) \wedge \textsc{Coached}(S) \Rightarrow \textsc{Type}(S, \textit{Weak}) \\
w_{v_hi} : \textsc{Views}(S, high) \wedge \textsc{Independent}(S) \Rightarrow \textsc{Type}(S, \textit{Strong}) \\
w_{d_i} : \textsc{DaysOnline}(S, high) \wedge \textsc{Independent}(S) \Rightarrow \textsc{Type}(S, \textit{Strong}) \\
w_{ct_c} : \textsc{Contributions}(S, high) \wedge \textsc{Coached}(S) \Rightarrow \textsc{Type}(S, \textit{Strong}) \\
w_{ct_i} : \textsc{Contributions}(S, high) \wedge \textsc{Independent}(S) \Rightarrow \textsc{Type}(S, \textit{Weak}) \\
\hline
\end{array}
$$

Ruleset 6: Student Types, Behavior, and Performance

student, S, is coached or independent.

Furthermore, we introduce these type-specific rules in Ruleset 6. In the first rule, we learn the association between student strength and performing well on course assessments. The $\textsc{AssessmentScore}(S, A_x)$ of student S on exam A_x is their true score scaled to be between 0 and 1. In the following rules, we model the relationship between types of forum behavior and student strength, when these relationships differ by student instruction type. As in Ruleset 2, the rules which concern coached students also have a duplicate for inferring section strength, following the template above. The final rule in this table is derived from inspecting the data and observing that for independent students forum contributions were negatively correlated with success. A contribution is more general than an answer in that it can include informal conversation. One possible explanation for this is that independent students who do not have a physical cohort of peers might be more inclined to use the forum to interact socially.

9.2.4. Collaboration dynamics

Next, we would like to uncover collaborative behavior and utilize it to improve predictions. Here, we are interested in one form of collaborative behavior, when students work closely together on the same assignment. We model this behavior with the predicate $\textsc{WorkingTogether}(S_i, S_j)$, which takes two students as arguments and expresses the extent to which these students might be inferred to be working together. As we do not know if two students are working together or not, we model working together behavior as a latent variable and infer the values to $\textsc{WorkingTogether}(S_i, S_j)$ for all potential S_i, S_j pairs. We express the prior belief that most pairs of students do not work together with the rule in Ruleset 7.

$$
\boxed{w_{nwt} : \neg\textsc{WorkingTogether}(S_i, S_j)}
$$

Ruleset 7: Collaborative Prior

To infer that students may be working together, we inspect forum post content. We define a post similarity function, $\textsc{PostSim}(P_1, P_2)$, which takes two posts as arguments and returns their similarity as a value between 0 and 1, where 1 indicates equality. If two students have similar posts, we predict that they are working together, as shown in the first rule in Ruleset 8.

$$w_{wt} : \textsc{Author}(S_i, P_1) \wedge \textsc{Author}(S_j, P_2) \wedge \textsc{SimPost}(P_1, P_2) \wedge \textsc{SameSection}(S_i, (S_j) \Rightarrow \textsc{WorkingTogether}(S_i, S_j)$$
$$w_{wta} : \textsc{WorkingTogether}(S_j, S_i) \Rightarrow \textsc{WorkingTogether}(S_i, S_j)$$

Ruleset 8: Working Together

Now, we express how working together might relate to student and section strength in the rules in Ruleset 9. We model that a student is in a given section, C, with $\textsc{InSection}(S_i, C)$. The first rule in Table 9 expresses that if two students are in a strong section and they are working together, if one of them is a strong student, the other is likely to be as well. Similarly, if two students are in the same weak section and they are working together, if one is weak, the other one likely is as well. The last rule in Ruleset 9 expresses a different dynamic. Here, if two students are in a weak section and one of them is strong, they may be working with a weaker student. This rule captures the potential dynamic of answer sharing. Coaches of strong sections might be better able to prevent such behavior; consequently, we model this dynamic occurring in weaker sections.

$$w_{cs} : \textsc{InSection}(S_i, C) \wedge \textsc{InSection}(S_j, C) \wedge \textsc{SectionStrength}(C, \textit{Strong}) \wedge \textsc{WorkingTogether}(S_i, S_j) \wedge \textsc{Type}(S_i, \textit{Strong}) \Rightarrow \textsc{Type}(S_j, \textit{Strong})$$
$$w_{cw} : \textsc{InSection}(S_i, C) \wedge \textsc{InSection}(S_j, C) \wedge \textsc{SectionStrength}(C, \textit{Weak}) \wedge \textsc{WorkingTogether}(S_i, S_j) \wedge \textsc{Type}(S_i, \textit{Weak}) \Rightarrow \textsc{Type}(S_j, \textit{Weak})$$
$$w_{csw} : \textsc{InSection}(S_i, C) \wedge \textsc{InSection}(S_j, C) \wedge \textsc{SectionStrength}(C, \textit{Weak}) \wedge \textsc{WorkingTogether}(S_i, S_j) \wedge \textsc{Type}(S_i, \textit{Strong}) \Rightarrow \textsc{Type}(S_j, \textit{Weak})$$

Ruleset 9: Collaboration and Student Types

9.2.5. Predicting post-test performance

Finally, we predict whether a student will pass the post-test with the rules in Ruleset 10, which connect a students' strength to passing. As slightly more students fail than pass, we include a prior with small weight which indicates that most students will not pass (the first rule in Table 10). Together, the rules presented in this section make up the collective probabilistic model SOCIO-BEHAVIORAL.

$$w_{neg} : \neg\textsc{Pass}(S)$$
$$w_{sp} : \textsc{Type}(S_i, \textit{Strong}) \Rightarrow \textsc{Pass}(S_i)$$
$$w_{wp} : \textsc{Type}(S_i, \textit{Weak}) \Rightarrow \neg\textsc{Pass}(S_i)$$

Ruleset 10: Predicting Performance

10. EMPIRICAL EVALUATION OF PREDICTIVE MODELS

In this section, we present empirical results using the predictive models defined above to predict post-test performance. For each model, we perform both 10-fold and 3-fold cross validation, in order to assess how models perform with more (10-fold) or less (3-fold) training data. We filter our student pool to those who participated in the forums and took the post-test (approximately 16% of all students who completed the post-test). We first present the most informative features discovered using RFE. We next evaluate the predictive results of the two models. Finally, we assess the discovered values of the latent section strength and co-working variables.

10.1. INFORMATIVE FEATURES

In total, there are 196 features of student behavior. Broadly these describe forum topics, forum behavior, course behavior, and classroom/coaching environments. In predicting performance, we retain only those features which are most informative. Here, we present the most informative topics mentioned in the student forum. We also outline and comment on critical course assessments and learning environments. Here, we present those features which are informative both in the 10-fold and 3-fold data setting. We found all aspects of forum behavior be useful in predicting performance, and thus do not discuss this feature set here (see Table 1 for a complete list).

10.1.1. Topics and Performance

Recall that we found 150 topics using LDA. LDA provides a categorical score for a given document, describing to what extent this document belongs to each topic. For a student, to find the extent to which they posted about a given topic, we sum the score provided by LDA for that topic across each of the student's posts. For each student this provides a vector of length 150 where each entry expresses the extent to which this student posted about each topic. When performing RFE we consider each of the 150 topics as potentially informative. Not only does including topics in our models of student behavior improve the final predictive model, by inspecting the topics that students posted about we can gain insights into students' course experience and the differences between strong and weak students.

The topics discovered by the topic model fall into four broad categories: help requests, assignments, course material, and course activities. In Table 6, we present the ten topics which are most predictive of post-test performance. Each of these appears in the final reduced feature set, which is visible to the predictive model. The first three topics in the table fall into the help requests category. They include words such as *trouble*, *help*, and *fail*. Four of the top ten topics correspond to assignments, with top words that are descriptive of assignments from the course. For example, in assignment A_4 students are asked to write a program to count the number of hashtags, links, and attributions in a tweet, and in the topic associated with this assignment we see the words: *hashtag*, *tweet*, *attributions*, *mentions*, and *links*. Two topics represent concepts discussed in the course: object-oriented programming and hash maps. The hash maps topic is particularly interesting as hash maps are not introduced in the course, but students still use them in their projects and discuss them on the forum. The other prominent topics relate to course activities. For example, the activity topic in the table is an activity given to students to print the location of a vehicle. This is the most elaborate activity that students undertake in the course; hence it appears in the top predictive topics for predicting post-test performance.

Table 6: Top predictive topics and the words in these topics.

Topic Label	**Top Words**
Help requests	trouble, don't, perfectly, won, updated
Help requests	hope, helps, change, find
Help requests	fail, expected, updated, supposed
Assignment content (A_4)	hashtag, tweet, attributions, mentions, links
Lecture (hashmaps)	Map, key, Getvalue, Hashmap, entry
Course Activity (vehicle activity)	vehicle, location, backward, forward, GetLocation
Assignment content (A_6)	ArrayList, words, remove, equals, size
Assignment content (A_{10})	strand, size, TurnOn, green, BurntOut
Assignment content (A_{14})	sort, insertion, swap, insert, algorithm
Lecture (OOP and Methods)	object, constructor, methods, parameter, returns

Figure 13 gives the distribution of passed and failed students across the ten most predictive topics given in Table 6. We observe that passing students post about the course activity on vehicles more than failing students. Since activities only contribute to a small portion of their grade, participation in activities is a good measure for students' level of motivation and learning.

Additionally, we observe that failing students are far more likely to write posts which fall in the help category. Looking at some of the posts in this category, we find that these posts are often short and use help words, but do not contain detailed information about the specific assignment problem in question. This finding suggests that analyzing the posts for linguistic cues is helpful in understanding students' metacognition.

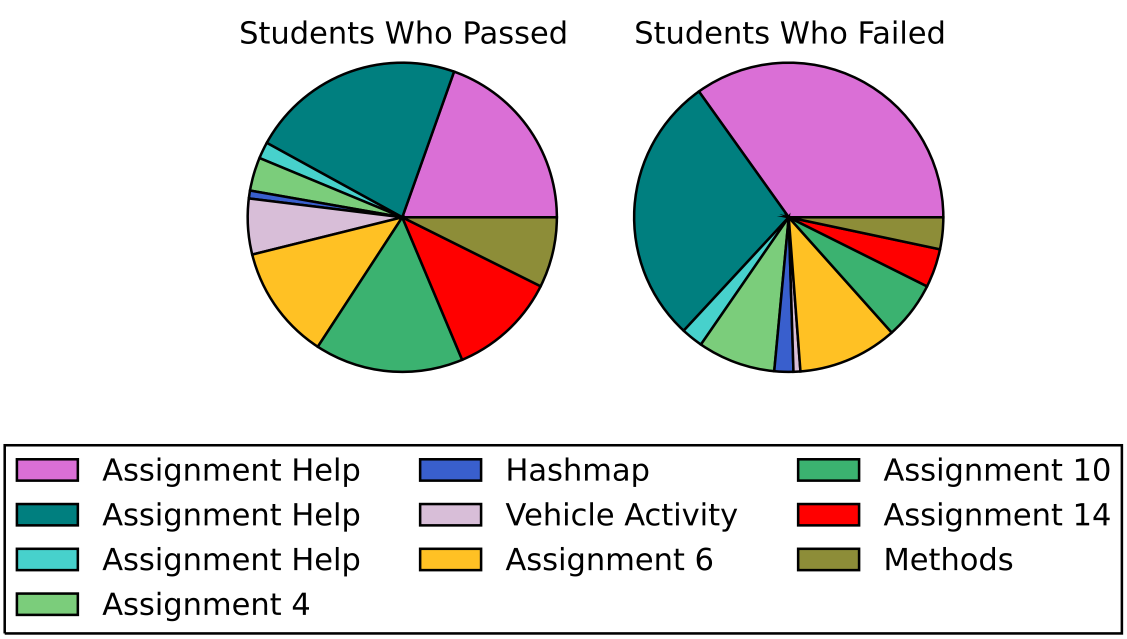

Figure 13: Commonly discussed topics on the student forum. Topic occurrence in posts authored by students who pass differs from that in posts authored by students who fail.

The third important take away from this analysis is that this topic distribution can help discover patterns in student behavior. For example, passing students post about assignment A_{10} more than failing students. But, failing students post more about assignment A_4. As assignments tend to get harder as the course progresses, the difference in behavior can be attributed to failing students needing help on the easier assignments, while the savvier students focus on the harder assignments.

10.2. Assignments and Performance

Here, we describe the most predictive assignments that we use in the two models. We find that assignments A_4, A_8, A_9, and A_{10} are the most predictive assignments in the 10-fold setting. These assignments are on core concepts and hence may be the most critical assignments in the course. This observation is bolstered by the fact that these assignments are referenced in the forums more than other assignments. Two of these assignments feature in the top ten predictive topics given in Table 6. Only A_{10} is also selected in the 3-fold setting. This assignment is the first which covers inheritance, a topic which many students struggle with and which is heavily covered on the AP exam. This stresses the importance of assessing students' understanding of inheritance *early* and providing additional resources for those students.

10.3. Learning Environment and Performance

Whether or not a student was coached is an informative feature. However, there can be differences in the skill levels of coaches as they all have different backgrounds in computer science and different teaching methods. We found that the number of times a coach answers a question on the coach forum, the days they spend online, and the size of their classroom were also useful in predicting the scores of coached students. The most decisive feature was the number of times a coach answers a question. One possible explanation for this is that these answers can approximate a coach's prior knowledge or involvement. Evaluating both students' and coaches' prior knowledge can provide more insight.

Additionally, we found that coaches with small sections were more likely to produce strong students. Yet, there were exceptions to this trend where a few coaches with large sections produced strong students. This might be an artifice of this dataset as there were a few coaches teaching large sections who had prior experience teaching computer science. There are many open questions about the relationship between prior experience and teaching ability, and furthermore, between ability and ideal section size.

10.4. Predictive Results

Here we present the precision, recall, and F-measure of the two predictive models, the SVM classifier and the collaborative model Socio-behavioral. We show the predictive results in Table 7. We see that Socio-behavioral outperforms the SVM overall, and for each student group as well.

We also see that Socio-behavioral is more robust to the training/testing environment. For example, moving from the 10-Fold to the 3-Fold setting, where this is less training data and more testing data per fold, we see the performance of the SVM drop by 4.5%. However, for Socio-behavioral performance only drops by .49%.

Both models are better able to predict the performance of coached rather than independent students. This is most likely due to the increased heterogeneity of the cohort of independent

Table 7: Performance of SOCIO-BEHAVIORAL and SVM in inferring post-test performance. Significant improvements ($p < 0.05$) of one model over another within the same train/test split environment are shown in bold (e.g., SOCIO-BEHAVIORAL relative to SVM in the 3-Fold condition).

	3-Fold			10-Fold		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
	All					
SVM	77.3	75.9	76.6	72.3	88.1	80.2
SOCIO-BEHAVIORAL	**82.0**	**82.4**	**82.2**	**75.6**	**89.6**	**82.6**
	Coached					
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
SVM	72.6	74.2	73.4	68.0	89.6	78.8
SOCIO-BEHAVIORAL	**78.4**	**80.5**	**79.5**	**72.2**	**90.3**	**81.2**
	Independent					
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
SVM	86.0	79.0	82.5	83.8	84.9	84.5
SOCIO-BEHAVIORAL	**88.6**	**85.7**	**87.1**	**84.4**	**87.6**	**86.0**

students. However, as these students might not have classroom support, improving the model to better understand their needs is important future work.

10.5. INFERRING CLASS STRENGTH WITH SOCIO-BEHAVIORAL

In addition to inferring if students will pass the post-test, in SOCIO-BEHAVIORAL we also infer the latent strength of each class or section with a coach. Here, we would like to evaluate how well this model discerns section strength. While true section strength is an unobserved variable, here we introduce a proxy score to evaluate SOCIO-BEHAVIORAL. For each class we calculate the percentage of students who pass the post-test. We introduce a proxy score SECTIONSTRENGTHPROXY(section), which is 1 if a section has a pass rate greater than 50% and 0 otherwise. To evaluate the predicted section strengths against SECTIONSTRENGTHPROXY(section), we round our predicted strength values with a .5 threshold. These results are shown in Table 8.

Table 8: Performance of SOCIO-BEHAVIORAL in inferring section strength.

	Precision	Recall	F-Measure
3-Fold	67.9	1.00	80.9
10-Fold	55.9	82.6	66.7

In Table 8, we compare the 3-Fold and 10-Fold versions of SOCIO-BEHAVIORAL. In this case, we see that it is more difficult to infer section strength in the 10-Fold case, where each fold has fewer students in the same class. In both cases, recall is much higher than precision, and precision can be further optimized. In both cases, the models were trained to optimize their ability to predict students' scores, not to infer section strength. Directly optimizing for this prediction would most likely lead to improved performance. Additionally, gathering more information about coaches teaching styles would better allow us to understand which styles are most successful.

A section's strength is most likely a combination of the abilities and achievements of its constituent students and of the coach leading the section. Coaches can influence students' learning, and determining what aspects of coaching are most effective can have a large impact on learning. Here, we model several potential indicators of a coach's strength and show that these offer reasonable performance. While these indicators might help course designers in choosing coaches to approach with extra resources, they are not informative enough to lend insight into successful classroom strategies. For example, we do not know how often strong instructors decided to meet with their students, nor do we know their prior knowledge.

10.6. Collaboration Dynamics

Students have multiple opportunities to collaborate in this course. Coached students might meet other students both in classroom settings and outside of them. Students can seek help on the online student forum and often receive answers from their peers. As collaborations have been shown to enhance learning in some settings, here we explore the potential effect of working together.

Consider a group of students who are seeking to solve the same assignment. We consider these students as working together if they are actively collaborating to reach the answer. They might share portions or all of their solutions with each other. They might turn to others with whom they are working when they are stuck. In our online setting we have no ground truth labels of this working together dynamic. Instead, we model it as a latent variable and infer the extent to which two students might be working together with the model described in Section 9.

Our primary evidence of students working together is their forum posts. A potential sign of working together typically observed by course instructors is the following. A student will ask for help on the course forum, and in a short window of time after this student posts a question, another student will post a question with highly similar code to the first post. Course instructors describe that students strategize that by posting the same question by different authors it might be answered more quickly. Many MOOCs rely on peers to address each other's questions. This MOOC hires instructors to respond to every post. Thus, while in many courses over-posting might result in less visibility for questions, in this course it is not a concern as every question is guaranteed to be answered within 24 hours. Thus, here we look for posts with high similarity as an indicator of co-working behavior.

In inferring these collaboration dynamics, we introduced a post similarity function. Recall we are interested in posts that suggest that students are working on the same assignment together. Following the observation that students will post highly similar questions about the same question we arrive at a set of criteria for assessing post similarity. First, we only consider question posts which are the first posts in a thread. These posts often contain the code the student will submit for the assignment. Next, we only consider the similarity between posts where the authors are in the same section. Additionally, these questions must be posted in a narrow timeframe, which suggests posting coordination. Finally, for two messages to be similar they must have high text similarity. This provides a very specific indicator of students working together, which is meant to capture the observed phenomenon of students working on one assignment and independently posting the code which they wrote together. Text similarity is calculated as the cosine similarity between term frequency-inverse document frequency (TF-IDF; Jurafsky and

Table 9: Number of interactions between students of different types. Here, we use students' post-test scores to label them as being strong or weak and assign students to be working together according to the output of SOCIO-BEHAVIORAL.

	Weak	*Strong*
Weak	9	5
Strong	5	5

Martin, 2000) vectors. In constructing the TF-IDF vectors we consider the entire corpus.

Next, we inspect which kinds of students are predicted to work together by SOCIO-BEHAVIORAL. We consider any pair of students (S_i, S_j) to be working together, if the model assigned a truth value of at least 0.5, to $\text{WORKINGTOGETHER}(S_i, S_j)$. To inspect the types of students involved in an interaction we use post-test performance to label students as *Strong* or *Weak*, where *Strong* are those who pass the AP and *Weak* students are those who do not. In total, we find 19 pairs of students out of 119 potential pairs. Of these, the largest interactions are between weak students.

The absolute numbers here are very small. We do see some strong-weak interactions are inferred. In the inferred interactions between strong and weak students, the weak students are all unexpected low learners with high course scores and low post-test scores. Thus, these interactions might suggest that the students are receiving help in submitting assignments, which might hamper their ability to properly internalize the course material. Furthermore, of the weak-weak interactions the majority are also between unexpected low learners. However, we also see interactions between strong students. This is a positive sign that students can collaborate and assist each other.

10.7. PREDICTING UNEXPECTED STUDENT TYPES

In Section 7, we saw two groups of students whose course performance did not match their post-test performance. Here, we inspect our ability to predict the performance of these groups. In Table 10 we analyze the accuracy of each model in predicting post-test performance for these groups of students.

In Table 10, we see that it is very difficult to predict the performance of unexpected student types. The behavior of unexpected low learners matches that of strong students, yet these students fail the post-test, while unexpected high students do not participate in the course, yet obtain high scores. By incorporating both student interactions and encoding relationships be-

Table 10: Prediction accuracy for unexpected student types. Significant improvements over competitors are shown in bold ($p < .05$).

	Accuracy	
	SVM	SOCIO-BEHAVIORAL
All	73.7	**76.8**
Unexpected low learners	52.3	**54.1**
Unexpected high learners	37.5	**52.7**

tween student characteristics and success, SOCIO-BEHAVIORAL is able to outperform the SVM model, though there is much room for improvement.

A question regarding unexpected low learners is whether they might be receiving help from their classmates. To further investigate this, we look at the student co-working pairs discovered by SOCIO-BEHAVIORAL. In all of the weak-strong interactions, we see that the weak students are unexpected low learners. This supports the theory that these students are working with strong students who may be over-helping them. Furthermore, in the weak-weak interactions we see that the majority of students (14/18) are unexpected low learners. Thus, weak students may also be depending on each other to the detriment of their learning.

11. DISCUSSION

We employed both 3-Fold and 10-Fold cross-validation for predicting performance. With three folds, there is less training and more testing data than in the 10-Fold case. Thus, the F-Measure is expected to decrease in the 3-Fold setting. This is what we saw in the SVM, with an F-Measure of 80.2 in the 10-Fold setting and 76.6 in the 3-Fold setting (a decrease of 4.5%). However, for SOCIO-BEHAVIORAL the decrease was much lower, just 0.49%.

Next, we investigated whether we might improve predictions by modeling collaboration dynamics. Indeed, this was the case. In the 3-Fold setting, where there are more instances of classroom structure in the test folds, we saw improvements in F-Measure over the SVM model of 7.4% (overall), 8.3% (coached students), and 5.7% (independent students). As the collaborative model can take advantage of classroom information and interactions between coached students, we expect to see the largest increase with regards to coached students.

In this collaborative model, we are also able to incorporate inferred section strength and pairs of students inferred to be working on assignments together. We saw that the inferred strong sections had higher numbers of passing students than the inferred weak sections. Additionally, we saw that the most common co-working interaction was between weak students and that the majority of all interactions involved one unexpected low learner. While collaborations can be beneficial for learning, high school students may need more guidance in forming successful teams. However, we only uncovered a small total number of such interactions. To better measure the effect of collaboration, it would be useful to observe changes in students' strength after working together. For example, do we see stronger students aiding weaker students to the extent that they can also pass the post-test? Our inferred co-working pairs are a static snapshot of evolving relationships. In future work, we will take temporal aspects into account when studying these effects.

11.1. LIMITATIONS

In this work, we have analyzed observational data to discover relationships between aspects of a high-school MOOC and post-test performance. We then used many of these relationships to build several predictive models of student success. This work could be strengthened by conducting explicit randomized trials to isolate specific aspects of learning, for example the effect of certain coaching strategies. We view our exploratory data analysis as work that highlights areas for further study. A central contribution of our work is the use of collective inference to strengthen predictive performance. As the relationships of interest to us are unobserved we introduce a latent model which can both infer these crucial dynamics and use them in inference.

Additional latent-variable approaches could be informative in this setting as well. However, PSL offers a scalable and expressive appoach. Bayesian networks (Friedman et al., 1999) could be relevant, yet they require different assumptions to template than PSL, which is conveniently templated with logical rules. Bayesian Knowledge Tracing (Corbett and Anderson, 1994) is a popular modeling framework for knowledge acquisition, but the standard form does not apply to our goal of predicting post-test performance from heterogenous observations of student behavior. Finally, Markov logic networks (Richardson and Domingos, 2006) also employ logic to template a different class of Markov random fields, however this framework does not admit reasoning over continuous latent variables, as does PSL.

We show that by modeling socio-behavioral dynamics in a data-driven manner, we can achieve better performance than when such dynamics are ignored. However, while we can clearly say that section strength is an indicator of student strength, much more work is needed to understand the circumstances which influence this relationship. While we can study the online characteristics of different coaches, such as the number of times they answer a question on the coaches' forum, we do not have access to their offline behavior.

A central hindrance is our inability to measure all of the aspects of the educational environment that can contribute to student learning. For example, we have an incomplete idea of the learning environments provided by each coach. Additionally, we do not have access to the complete social network of this high-school MOOC and rely on partial measurements to inform our model of collaboration. Finally, there are many latent constructs that can inform the study of learning. Here, we explore the effect of only some of the many potential latent constructs. While in future work we hope to gather more data in order to further our understanding of student learning in this MOOC environment, this work presents several contributions designed to confront these limitations.

Expanding upon this, a further limitation with this work is that we lack details about each coach's strategies. Understanding the pedagogical tactics of each coach will better allow us to model section strength. For example, it would be useful to know coaches' prior knowledge and subject competency, how often they meet with students, how they use the forum resources, and what kind of collaborative dynamics they foster. As a first step in this direction we have modeled latent section strength. This latent variable is designed to explain many of the attributes of coaches which are in our case unobserved. In future work we hope to combine enriched observational data with this latent-variable model to better understand the influence of coaching on student learning.

Collaboration can be an essential aspect of student learning. As we have observed that students often work on the same assignment together, often simultaneously posting the same code, we included this dynamic as a possible form of student interaction in our proposed model. This is only one potential form of student interaction. While there are many other interactions to explore, this one was deemed important in the context of this MOOC upon consultation with course instructors. Obtaining observations outside of the forum would allow us to better examine a fuller range of collaboration dynamics. For example, observing which students work together in the classroom could be more informative of collaboration. Modeling social ties such as friendship would also be very beneficial. Finally, with more multi-modal traces of student behavior, we could better differentiate collaboration dynamics. In our case, being able to differentiate between the dynamics of sharing information (such as assessment answers or code snippets) and collaboratively improving understanding (such as discussing course concepts) would allow us to better understand student learning in this environment.

In this work, we use final post-test performance as our guiding metric. A limitation of this metric is that it does not capture changes in course mastery. Another potential metric of learning would be improvement on a post-test from a pre-test. Furthermore, our performance metric does fully capture other aspects of how students participate in online courses. Finally, a limitation with this metric is that it reduces the number of students whose course mastery can be analyzed, as not all students participated in the post-test. A key contribution our work is to introduce a flexible modeling framework for this domain which can readily incorporate multiple measures of student behavior. In the future, we hope to also expand our metrics of learning to incorporate these diverse behaviors. For example, a more holistic measure of how students benefit from MOOCs could include their study patterns, social behavior, and grasp of critical concepts.

Here we model student strength from observed individual characteristics and behaviors. A limitation of this modeling is that it does not model psychological traits, which can influence student behavior in online environments, such as motivation and conscientiousness. In future work we plan to survey students with respect to their motivational styles and other psychological constructs. A benefit of our model is that one can template interpretable hypotheses about how unobserved factors might be influencing student learning. These hypotheses can lend strength to even limited data. Additionally, we expect that models of student learning will continuously evolve. Here, we have presented a particular model of student learning within a general modeling framework. We foresee that the specifics of this model will evolve with the science of learning.

12. Conclusion

Our work shows that MOOCs are a viable option for high school students. Forty-seven percent of students who took the post-test passed it. Of these students, approximately 29% were, to the best of our knowledge, self-directed. We found that post-test performance was correlated with course performance, such that students who earned a high course score also earned a high post-test score. While we can say that MOOCs work for some high school students, the particularities of this group must be understood. Towards this end, we have characterized high and low learners by their course and forum behavior, as well as by the topics that they post about.

Our collective socio-behavioral model incorporates findings from the data to predict post-test performance. Critically, this model uses latent variables to uncover hidden structure in this domain. By modeling latent collaborative dynamics and inferring post-test scores collectively, we were better able to predict students' performance. We also improve our predictions of coached students' performance by inferring latent section strength.

This collective probabilistic approach is very general. Currently, we incorporate relationships between student and coach behavior, and student performance and show the effectiveness of this approach. In future work, it may be advantageous to encode additional contextual knowledge about coaches' experience, styles, and knowledge. Additionally, we could adapt this model to include knowledge of other internal student states, such as evolving course comprehension and motivation. Finally, this modeling approach can be used to design course interventions. If we can predict performance early in a student's experience, we may be able to personalize interventions to the needs of the student. Strong students might need more challenging, conceptual exercises, while other students might benefit from review or focused instructor attention. Online education introduces the opportunity to provide personalized learning experiences at scale. We view this work as a step in that direction.

References

Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. 2014. Engaging with massive online courses. In *International Conference on World Wide Web (WWW)*. 687–698.

Anderson, T. 2003. Getting the mix right again: An updated and theoretical rationale for interaction. *The International Review of Research in Open and Distributed Learning 4,* 2. Retrieved from: http://www.irrodl.org/index.php/irrodl/article/view/149/230.

Andrews-Todd, J., Forsyth, C., Steinberg, J., and Rupp, A. 2018. Identifying profiles of collaborative problem solvers in an online electronics environment. In *International Conference on Educational Data Mining (EDM)*, K. E. Boyer and M. Yudelson, Eds. 239–245.

Aviv, R., Erlich, Z., Ravid, G., and Geva, A. 2003. Network analysis of knowledge construction in asynchronous learning networks. *Journal of Asynchronous Learning Networks 7,* 3, 1–23.

Bach, S. H., Broecheler, M., Huang, B., and Getoor, L. 2017. Hinge-loss Markov random fields and probabilistic soft logic. *Journal of Machine Learning Research 18,* 109, 3846–3912.

Baker, R., Corbett, A., Koedinger, K., and Wagner, A. 2004. Off-task behavior in the cognitive tutor classroom: When students "game the system". In *SIGCHI Conference on Human Factors in Computing Systems*. 383–390.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, 993–1022.

Blom, J., Verma, H., Li, N., Skevi, A., and Dillenbourg, P. 2013. MOOCs are more social than you believe. *33*, 1–3.

Bock, M. and O'Dea, V. 2013. Virtual educators critique value of MOOCs for K-12. https://www.edweek.org/ew/articles/2013/02/06/20moocs.h32.html.

Chen, G., Davis, D., Hauff, C., and Houben, G.-J. 2016. On the impact of personality in massive open online learning. In *User Modeling Adaptation and Personalization*. 121–130.

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

Er, E., Gómez-Sánchez, E., Bote-Lorenzo, M. L., Asensio-Pérez, J. I., and Dimitriadis, Y. 2019. Informing the design of collaborative activities in MOOCs using actionable predictions. In *ACM Conference on Learning@ Scale (L@S)*. 27:1–27:4.

Friedman, N., Getoor, L., Koller, D., and Pfeffer, A. 1999. Learning probabilistic relational models. In *IJCAI*. Vol. 99. 1300–1309.

Gardner, J. and Brooks, C. 2018. Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction 28,* 2, 127–203.

Gitinabard, N., Khoshnevisan, F., Lynch, C., and Wang, E. Y. 2018. Your actions or your associates? Predicting certification and dropout in MOOCs with behavioral and social features. In *International Conference on Educational Data Mining (EDM)*, K. E. Boyer and M. Yudelson, Eds. 404–410.

Hanzaki, M. R. and Epp, C. D. 2018. The effect of personality and course attributes on academic performance in MOOCs. In *European conference on technology enhanced learning*. 497–509.

Huang, J., Dasgupta, A., Ghosh, A., Manning, J., and Sanders, M. 2014. Superposter behavior in MOOC forums. In *ACM Conference on Learning @ Scale Conference (L@S)*. 117–126.

Jurafsky, D. and Martin, J. H. 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. *Prentice Hall*.

KELLOGG, S., BOOTH, S., AND OLIVER, K. 2014. A social network perspective on peer supported learning in MOOCs for educators. *The International Review of Research in Open and Distributed Learning 15,* 5, 263–289.

KENNEDY, G., COFFRIN, C., DE BARBA, P., AND CORRIN, L. 2015. Predicting success: How learners' prior knowledge, skills and activities predict MOOC performance. In *International Conference on Learning Analytics And Knowledge (LAK)*. 136–140.

KIZILCEC, R. F., PIECH, C., AND SCHNEIDER, E. 2013. Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In *International Conference on Learning Analytics and Knowledge (LAK)*. 170–179.

KLÜSENER, M. AND FORTENBACHER, A. 2015. Predicting students' success based on forum activities in MOOCs. In *International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS)*. Vol. 2. 925–928.

KUMAR, R. AND ROSE, C. P. 2011. Architecture for building conversational agents that support collaborative learning. *IEEE Transactions on Learning Technologies 4,* 1, 21–34.

KURHILA, J. AND VIHAVAINEN, A. 2015. A purposeful MOOC to alleviate insufficient CS education in Finnish schools. *Transactions of Computing Education 15,* 2, 1–18.

LI, N., VERMA, H., SKEVI, A., ZUFFEREY, G., BLOM, J., AND DILLENBOURG, P. 2014. Watching MOOCs together: investigating co-located MOOC study groups. *Distance Education 35,* 2, 217–233.

LI, X., WANG, T., AND WANG, H. 2017. Exploring n-gram features in clickstream data for MOOC learning achievement prediction. In *International Conference on Database Systems for Advanced Applications*. 328–339.

LONGSTAFF, E. 2017. Ritual in online communities: A study of post-voting in MOOC discussion forums. *International Journal of Human Computer Interaction 33,* 8, 655–663.

LOYA, A., GOPAL, A., SHUKLA, I., JERMANN, P., AND TORMEY, R. 2015. Conscientious behaviour, flexibility and learning in massive open on-line courses. *Procedia-Social and Behavioral Sciences 191*, 519–525.

MCCALLUM, A. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu/.

MEGA, C., RONCONI, L., AND DE BENI, R. 2014. What makes a good student? How emotions, self-regulated learning, and motivation contribute to academic achievement. *Journal of Educational Psychology 106,* 1, 121–131.

MOTZ, B., BUSEY, T., RICKERT, M., AND LANDY, D. 2018. Finding topics in enrollment data. In *International Conference on Educational Data Mining (EDM)*, K. E. Boyer and M. Yudelson, Eds. 424–430.

NAJAFI, H., EVANS, R., AND FEDERICO, C. 2014. MOOC integration into secondary school courses. *The International Review of Research in Open and Distributed Learning 15,* 5, 306–322.

PEDREGOSA ET AL., F. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12,* Oct, 2825–2830.

RAMESH, A., GOLDWASSER, D., HUANG, B., DAUME III, H., AND GETOOR, L. 2014. Learning latent engagement patterns of students in online courses. In *AAAI Conference on Artificial Intelligence*. 1272–1278.

RICHARDSON, M. AND DOMINGOS, P. 2006. Markov logic networks. *Machine learning 62,* 1-2, 107–136.

ROSÉ, C., WANG, Y.-C., CUI, Y., ARGUELLO, J., STEGMANN, K., WEINBERGER, A., AND FISCHER, F. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning 3,* 3, 237–271.

ROSÉ, C. P., CARLSON, R., YANG, D., WEN, M., RESNICK, L., GOLDMAN, P., AND SHERER, J. 2014. Social factors that contribute to attrition in MOOCs. In *ACM Conference on Learning @ Scale Conference (L@S)*. 197–198.

ROSÉ, C. P., GOLDMAN, P., ZOLTNERS SHERER, J., AND RESNICK, L. 2015. Supportive technologies for group discussion in MOOCs. *Current Issues in Emerging eLearning 2,* 1, 5. Available at: https://scholarworks.umb.edu/ciee/vol2/iss1/5.

SIMON, B., PARRIS, J., AND SPACCO, J. 2013. How we teach impacts student learning: Peer instruction vs. lecture in CS0. In *ACM Technical Symposium on Computer Science Education (SIGCSE)*. 41–46.

STAUBITZ, T. AND MEINEL, C. 2019. Graded team assignments in MOOCs: Effects of team composition and further factors on team dropout rates and performance. In *ACM Conference on Learning@ Scale (L@S)*. 5:1–5:10.

SU, Y.-S., HUANG, C. S.-J., AND DING, T.-J. 2016. Examining the effects of MOOCs learning social search results on learning behaviors and learning outcomes. *Eurasia Journal of Mathematics, Science and Technology Education 1,* 9, 2517–2529.

SUN, Y., NI, L., ZHAO, Y., SHEN, X.-L., AND WANG, N. 2018. Understanding students engagement in MOOCs: An integration of self-determination theory and theory of relationship quality. *British Journal of Educational Technology*, 1–19.

TOMKINS, S., RAMESH, A., AND GETOOR, L. 2016. Predicting post-test performance from online student behavior: A high school MOOC case study. In *International Conference on Educational Data Mining (EDM)*, T. Barnes, M. Chi, and M. Feng, Eds. 239–246.

VELETSIANOS, G., COLLIER, A., AND SCHNEIDER, E. 2015. Digging deeper into learners' experiences in MOOCs: Participation in social networks outside of moocs, notetaking and contexts surrounding content consumption. *British Journal of Educational Technology 46,* 3, 570–587.

WANG, X., YANG, D., WEN, M., KOEDINGER, K., AND ROSÉ, C. P. 2015. Investigating how student's cognitive behavior in MOOC discussion forums affect learning gains. In *International Conference on Educational Data Mining (EDM)*, O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, and M. Desmarais, Eds. 226–233.

WEN, M., YANG, D., AND ROSÉ, C. P. 2015. Virtual teams in massive open online courses. In *International Conference on Artificial Intelligence in Education*, C. Conati, N. Heffernan, A. Mitrovic, and F. M. Verdejo, Eds. 820–824.