Investigating the Impact of Skill-Related Videos on Online Learning

Ethan Prihar
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
ebprihar@wpi.edu

Aaron Haim
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
ahaim@wpi.edu

Tracy Shen
The Pennsylvania State University
University Park, Pennsylvania, USA
jqs5443@psu.edu

Adam Sales
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
asales@wpi.edu

Dongwon Lee
The Pennsylvania State University
University Park, Pennsylvania, USA
dongwon@psu.edu

Xintao Wu
University of Arkansas
Fayetteville, Arkansas, USA
xintaowu@uark.edu

Neil Heffernan
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
nth@wpi.edu

ABSTRACT

Many online learning platforms and MOOCs incorporate some amount of video-based content into their platform, but there are few randomized controlled experiments that evaluate the effectiveness of the different methods of video integration. Given the large amount of publicly available educational videos, an investigation into this content's impact on students could help lead to more effective and accessible video integration within learning platforms. In this work, a new feature was added into an existing online learning platform that allowed students to request skill-related videos while completing their online middle-school mathematics assignments. A total of 18,535 students participated in two large-scale randomized controlled experiments related to providing students with publicly available educational videos. The first experiment investigated the effect of providing students with the opportunity to request these videos, and the second experiment investigated the effect of using a multi-armed bandit algorithm to recommend relevant videos. Additionally, this work investigated which features of the videos were significantly predictive of students' performance and which features could be used to personalize students' learning. Ultimately, students were mostly disinterested in the skill-related videos, preferring instead to use the platforms existing problem-specific support, and there was no statistically significant findings in either experiment. Additionally, while no video features were significantly predictive of students' performance, two video features had significant qualitative interactions with students' prior knowledge, which showed that different content creators were more effective for different groups of students. These findings can be used to inform the design of future video-based features within online learning platforms and the creation of different educational videos specifically targeting higher or lower knowledge students. The data and code used in this work can be found at https://osf.io/cxkzf/.

L@S '23, July 20–22, 2023, Copenhagen, Denmark

ACM ISBN 979-8-4007-0025-5/23/07.
https://doi.org/10.1145/3573051.3593376

CCS CONCEPTS

• **Applied computing** → **Education**; **Distance learning**; **Computer-assisted instruction**.

KEYWORDS

Video Tutoring, Randomized Controlled Experiments, Multi-Armed Bandit Algorithms, Personalized Learning

ACM Reference Format:
Ethan Prihar, Aaron Haim, Tracy Shen, Adam Sales, Dongwon Lee, Xintao Wu, and Neil Heffernan. 2023. Investigating the Impact of Skill-Related Videos on Online Learning. In *Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S '23), July 20–22, 2023, Copenhagen, Denmark.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3573051.3593376

1 INTRODUCTION

There is currently a plethora of educational content available for free online. While this can empower students savvy enough to navigate to relevant content on their own, searching for relevant content can frustrate less experienced students, increasing their cognitive load and making it more difficult for them to obtain the same benefits [8]. Often, learning platforms will develop their own instructional content by working with, or crowdsourcing from experts, e.g., [3, 13], but this can be time consuming and expensive. In many cases, teachers will search for hours to find relevant instructional content to distribute to their students [11]. We are interested in reducing the cost for learning platforms to provide relevant instructional content to students and taking the burden of identifying and distributing relevant instructional content off teachers.

Prior research has shown that distributing educational videos to students has a positive impact on their learning [12, 17]. In these studies, the problem-specific videos were created by the researchers and were designed to explain how to solve the specific mathematics problems for which they were provided. Building off this prior

research, this work investigated if free and publicly available skill-related videos have a similar positive effect on students' learning. Videos aggregated from YouTube via automated searches were incorporated into the ASSISTments online learning platform [7] and provided to students upon their request. In addition to a randomized experiment investigating the effectiveness of these videos, multi-armed bandit algorithms (MABs) were used to identify which videos were most effective for each mathematics skill using the ASSISTments Automatic Personalized Learning Service (APLS) [15]. The effectiveness of the videos recommended via MAB were compared to randomly recommended videos to investigate the impact that MABs could have on the incorporation of these videos into online learning platforms.

Additionally, features of these videos, extracted using various machine learning APIs, were evaluated for their correlation with students' performance and for their ability to personalize students learning based on students' prior knowledge. For a feature to be capable of personalizing students' learning, there must be a qualitative interaction between the feature and prior knowledge. A qualitative interaction indicates that one group of students benefits more from one value of the feature while another group benefits more from a different value of the feature. For example, if high-knowledge students benefited more from long videos and low-knowledge students benefited more from short videos, then the video length feature could be used to personalize students' learning.

To summarize, this work answers the following research questions:

(1) What is the effect of incorporating publicly available skill-related videos into an online learning platform on students' performance?
(2) What is the effect of using multi-armed bandit algorithms to recommend videos on students' performance?
(3) What features of these videos are most predictive of students' performance?
(4) What qualitative interactions between video features and students' prior knowledge are most predictive of students' performance?

2 BACKGROUND

2.1 Instructional Videos

Instructional videos have been used successfully in the context of online learning many times. In a randomized controlled study in which the same problem-specific tutoring was provided to students in video or text format, it was shown that videos led to higher student performance than text [12]. Additionally, a combined analysis of five different randomized controlled experiments that compared video feedback to text feedback within the ASSISTments online learning platform found that videos were more effective than text across a variety of measures such as mastery speed and posttest score [17]. While these studies demonstrate the effectiveness of videos for problem-specific support, in this work we propose using videos to give more general, skill-related instruction.

Massive open online courses (MOOCs) are a good example of using videos not to provide specific feedback for individual problems, but to convey information on various topics in general. Many MOOCs feature videos in a wide variety of formats [19], from recordings of classroom lectures, to completely virtual presentations, to hybrid approaches, as well as various levels of integration with online assessments to enable students to practice as they learn. Not only do videos come in a variety of formats, but students use videos differently, and prefer videos formatted in a variety of ways. For example, a study of MITx MOOCs found that there was a distinct bimodal distribution in students' video usage across different courses, demonstrating differences in preference of how to use the MOOC videos [23]. Additionally, prior work has found that some students prefer classroom lecture recordings while others prefer fully digital presentations, and that these preferences are statistically significantly correlated with their motivation for enrolling in the MOOC [24]. While the study in this work is not done within a MOOC, these MOOC studies show the variety of formats and preferences for video-based content. The skill-related videos provided to students in this study may follow usage trends similar to the videos in MOOCs.

2.2 Multi-Armed Bandit Algorithms

Multi-Armed bandit algorithms (MABs) are a simple type of reinforcement learning where the algorithm takes one of multiple possible actions, is given a numeric reward based on criteria defined by the researcher, and models the relationship between each action and the expected reward. Over time, a MAB uses its model to try and maximize the reward it receives by taking actions with the highest expected reward [20]. MABs assume that the reward received for an action is independent of the sequence of actions taken, unlike more complicated reinforcement learning algorithms.

Research has shown in simulation that MABs would be able to increase students' learning during randomized experiments performed within online learning platforms, but would also increase the false-discovery rate of significant experiment results [18]. Although there are methods to adjust how a MAB operates to correct for some of the increase in false-discovery rate [25, 26], to avoid any bias, this work includes a randomized controlled experiment to investigate the effectiveness of providing students with skill-related videos. However, MABs have been shown via a large-scale randomized experiment to sightly improve students' performance by learning the most effective problem-specific support messages for middle-school mathematics problems [15]. Compared to randomly receiving one of multiple relevant problem-specific supports, students that received the support recommended by the MAB got the next problem in their assignment correct more often [15]. Therefore, to both maximize the benefit of skill-related videos and to study the effects of MABs on student performance in a different but similar context to the previous study, this work also studies the effect of using a MAB to recommend skill-related videos to students.

2.2.1 Thompson Sampling. The MAB used in this work is Thompson sampling. Thompson sampling was used in previous studies comparing MABs to random selection [15, 18] and has outperformed other algorithms when recommending content to students [15, 18]. Thompson sampling models the expected reward of each action it can take as a distribution of the rewards it has received for that action before. Each time Thompson sampling receives a reward for taking an action, that reward is used to update the action's prior distribution. Thompson sampling selects which

action to take by randomly sampling from each action's prior reward distribution, and then takes the action corresponding to the highest random sample [22]. By randomly sampling from the prior distributions, Thompson sampling balances learning more about actions that have not been taken frequently with taking actions that lead to the highest reward on average. At the beginning of Thompson sampling's use it will know very little about each action, and thus each prior distribution will have a high variance. The high variance will lead to random samples far from the mean reward of each action, which will make Thompson sampling's choice of action very similar to random selection. However, once each action has been taken many times, the variance of the prior reward distributions tends to decrease, and Thompson sampling will begin to take the action with the highest expected reward more frequently. The Thompson sampling algorithm used in this work is Beta-Bernoulli Thompson Sampling (BBTS), which models the prior distribution of a binary reward as a Beta distribution, and has been proven to be asymptotically optimal in [9]. BBTS has been used successfully in the past to recommend problem-specific support to students [15].

2.3 The ASSISTments APLS

The experiments in this work were performed within ASSISTments, an online learning platform that focuses on middle-school mathematics. Since 2021, ASSISTments has been able to use MABs to personalize the content provided to students through the Automatic Personalized Learning Service (APLS) [15]. The APLS allows for algorithms to make content recommendations for students in real-time. The APLS has the capacity to incorporate features of students, problems, and the content itself to its decision of what content to provide to a student. When multiple recommendation algorithms are available in the APLS, one is selected randomly, which enables randomized experiments between algorithms [15]. In this work, a random selection model and a BBTS model were added to the APLS for recommending videos. This way, the APLS administers the experiment comparing MABs to random selection, and the random selection model administers the experiments comparing videos.

Each night, the APLS calculates a reward for each recommendation made in the past 24-hours and updates each recommendation algorithm using these rewards. If a student was able to complete the next problem on their first try without any additional tutoring, the algorithm receives a reward of 1 for its video recommendation. Otherwise, the algorithm receives a reward of 0. In the studies in this work, the algorithms received rewards regardless of whether or not the student observed the skill-related video because both the random selection model and the BBTS model had the option to recommend no video. If a reward was only given when students viewed the videos, a reward could never be calculated for recommending no video. The downside of this is that the population of students that never observed the skill-related videos, while not biasing the prior reward distributions, added noise, making it more difficult to learn the differences in effectiveness between videos.

3 SKILL-RELATED VIDEOS

3.1 The Show Video Button

Prior to this work, ASSISTments only had the capacity to offer students problem-specific support. Given that it has been shown multiple times that the problem-specific support in ASSISTments benefits students [13, 17, 21], it would have been potentially detrimental to replace this problem-specific support with skill-related videos. Instead of replacing this tutoring, a new button was added to the ASSISTments Tutor. The ASSISTments Tutor is shown in Figure 1. Figure 2 shows the explanation, in yellow, that appears when a student clicks the Show Explanation button, which is the pre-existing button used to request problem-specific support. This tutoring only explains how to solve the specific problem on screen.

The new Show Video button is to the left of the Show Explanation button. When a student clicks on the Show Video button, a new tab containing a skill-related video opens in the student's web browser. Viewing a skill-related video does not directly explain how to solve the specific problem in the Tutor, and therefore, there is no penalty for requesting a skill-related video, unlike the problem-specific support, which removes a fraction of a student's score when requested. To familiarize students with the new Show Video button, an information icon, shown in Figure 1 directly to the left of the Show Video button, was provided. When students hover over the information icon, the message "Clicking this button does not reduce your score. It shows a video to help you solve the problem" is displayed. Figure 3 shows an example of a video opened in a new tab when a student clicks the Show Video button.

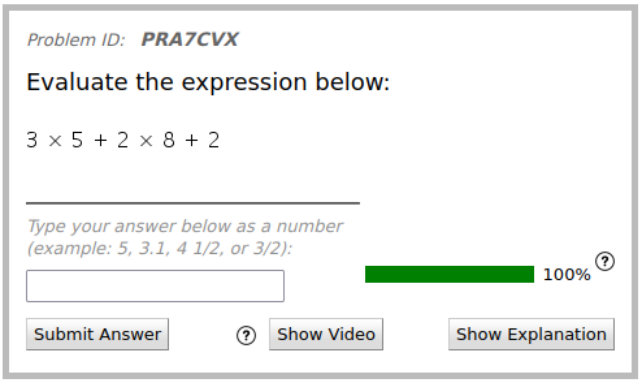

Figure 1: A mathematics problem in the ASSISTments Tutor. The new Show Video button appears to the left of the pre-existing Show Explanation button.

3.2 Video Incorporation

To incorporate skill-related videos into the ASSISTments APLS, the following steps had to be taken.

(1) Skill Labeling: Tag every problem in ASSISTments with the most relevant Common Core Skill Code [1].
(2) Video Filtering: Identify publicly available YouTube videos relevant to each skill.
(3) Feature Extraction: Create features of the videos and incorporate them into the APLS in order to investigate their impact on student performance.

3.2.1 Skill Labeling. The Common Core State Standards for Mathematics [1] discretize the United States mathematics curriculum into a tree of branching codes, where each leaf refers to a specific

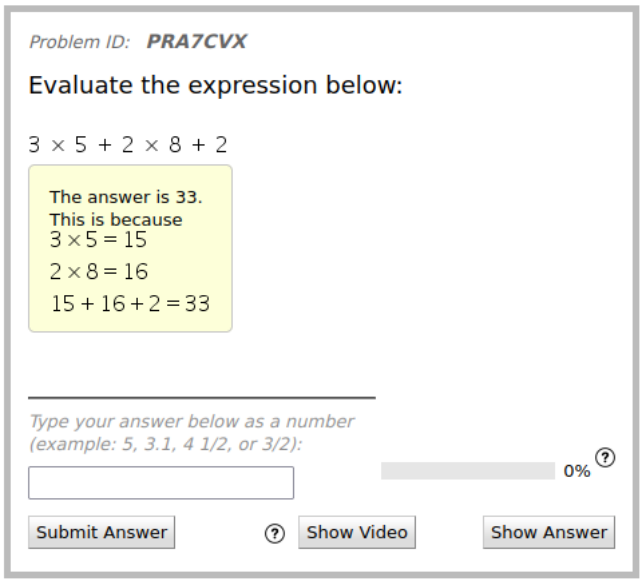

Figure 2: A mathematics problem in the ASSISTments Tutor with an explanation highlighted in yellow.

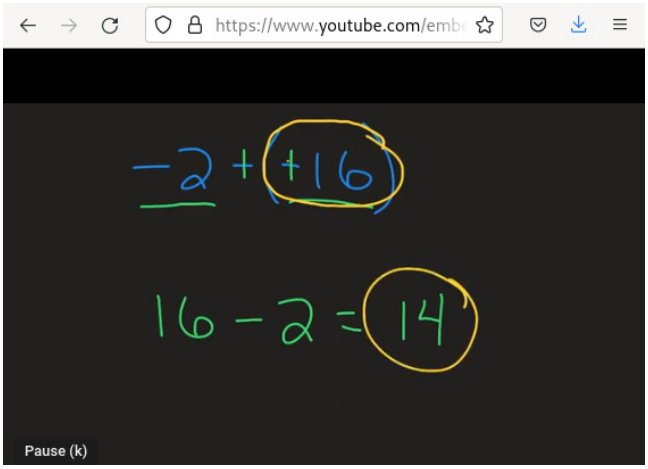

Figure 3: An example of a skill-related video.

concept that students must learn during their mathematics education. For example, the Skill Code 7.G.A.1 refers to a 7th grade geometry problem (7.G). The letter A refers to a section of the 7th grade geometry curricula, specifically the section described as "Draw, construct, and describe geometrical figures and describe the relationships between them." The number 1 is the final part of the skill code which refers to the skill in section A described as "Solve problems involving scale drawings of geometric figures, including computing actual lengths and areas from a scale drawing and reproducing a scale drawing at a different scale.".

For each 6th grade through 8th grade mathematics problem in the Engage New York[1], Illustrative Mathematics[2], and Utah Middle School Math Project[3] curricula, two teachers labeled each mathematics problem with the Common Core Skill Code most relevant to solving the problem. If the two teachers agreed, then that was the final skill code incorporated into ASSISTments. If the teachers disagreed, a third teacher was used to decide which of the two skill codes was correct. Essentially, two out of three teachers had to agree on the skill code for each mathematics problem before it was labeled. In total, 16,167 mathematics problems were tagged with their most relevant skill.

3.2.2 Video Filtering. After all the mathematics problems were tagged with their most relevant skill code, the skill code descriptions were used as the search term in YouTube in order to find relevant videos for each skill. The first ten results of each search were collected and shown to middle-school math teachers. The teachers were instructed to select the first five relevant videos for each skill. If less than five videos were relevant, then the teachers were instructed to go to YouTube and find the remaining videos themselves. Even though part of this work was to investigate how well BBTS would be able differentiate between more and less effective videos, the videos were still evaluated by teachers because at no point in this work would it have been acceptable for students to have been shown noneducational content. This process was used to find five relevant videos for each skill. The number five was chosen somewhat arbitrarily, with the goal of having enough videos for there to be variations between them, but few enough videos that BBTS would have time to learn the effectiveness of each video. In total, 1,315 videos were collected for 263 skills.

3.2.3 Feature Extraction. Once five videos for each skill were collected, a variety of machine learning APIs and YouTube metadata was used to create features for each video. Two APIs, Speechace[4] and DeepAffects[5], were used to extract features related to the voice of the speaker in the video if there was one. The Azure Face API[6] was used to examine qualities of the face in the video if the speaker included their face. Lastly, YouTube metadata[7] was used to extract features related to the length and appeal of the videos. The number of dislikes for a video was made private by YouTube on November 10th, 2021[8], but these features were extracted prior to that change. Of the dozens of features available through these sources, 12 were included as features in the APLS and used for further analysis of the experimental results. If all the features had been included, the false discovery rate of features that significantly impact student performance would have been much higher. The following 12 features were chosen because of their relevance to the educational quality of the videos, as determined qualitatively by a combination of middle-school mathematics teachers and researchers. Essentially,

[1] http://www.nysed.gov/curriculum-instruction/engageny
[2] https://illustrativemathematics.org/
[3] http://utahmiddleschoolmath.org/
[4] https://docs.speechace.com/
[5] https://docs.deepaffects.com/docs/introduction.html
[6] https://azure.microsoft.com/en-us/products/cognitive-services/face/
[7] https://www.youtube.com/
[8] https://blog.youtube/news-and-events/update-to-youtube/

these features are based on engagement statistics and things the educators that were consulted had heard students express preferences for in the past.

- **Length**: The length, in seconds, of the video, determined using YouTube metadata.
- **View Count**: The number of views of the video, determined using YouTube metadata.
- **Percent Likes**: The ratio of likes to views, determined using YouTube metadata.
- **Percent Dislikes**: The ratio of dislikes to views, determined using YouTube metadata.
- **Percent Comments**: The ratio of comments to views, determined using YouTube metadata.
- **Pronunciation Score**: A score from 0-100 that assesses how well the words in the video are pronounced, determined using Speechace API.
- **Unknown Pronunciation Score**: A binary indicator for whether or not Speechace was unable to calculate a pronunciation score.
- **Male Tone**: A binary indicator for whether or not the tone of the speaker sounded as though they were male, determined using the DeepAffects API.
- **Reading Tone**: A binary indicator for whether or not the tone of the speaker sounded as though they were reading, determined using the DeepAffects API.
- **Passionate Tone**: A binary indicator for whether or not the tone of the speaker sounded passionate, determined using the DeepAffects API.
- **Unknown Tone**: A binary indicator for whether or not DeepAffects was unable to analyse part of the tone.
- **Face Included**: A binary indicator of whether or not there was a face included in the video, determined using Azure Face API.

4 METHODOLOGY

4.1 Empirical Studies

Two randomized controlled experiments were performed using the ASSISTments APLS between March 3rd, 2022 and July 18th, 2022. The first experiment investigated the impact of skill-related videos on student performance, and the second experiment investigated the impact of using a MAB, specifically BBTS, to recommend skill-related videos compared to randomly recommending skill-related videos. Both studies were run simultaneously at the problem level, on different subsets of the student population. When a student started a problem, they were first randomized with equal probability between receiving a randomly recommended video or a BBTS recommended video. Students randomized to a BBTS recommended video were the treatment population for the second experiment, and BBTS was used to recommend one of the five relevant videos for the skill the problem was tagged with or no video (six options per recommendation). Students randomized to a randomly recommended video were the control population for the second experiment, and were randomly given one of the five relevant videos for the skill the problem was tagged with or no video with equal probability (1/6 chance of receiving each video, 1/6 chance of receiving no video). Students in the control population of the second experiment that were randomized to no video were considered the control population for the first experiment, and students randomized to any of the five videos were considered the treatment population.

Essentially, all students participated in the second experiment, and the half of students that were given randomly recommended videos participated in the first experiment as well. Both experiments were intent-to-treat analyses because the Show Video button was visible or not based on which condition a student was in. Because the presence of the button could have an effect on students' behavior, a student was included in the analysis if they were randomized into a condition, regardless of whether or not they viewed the skill-related video. Both experiments used next-problem correctness as the dependent measure. Correctness is a binary indication of whether the student got the problem correct on their first try without any additional support (1) or not (0). Next-problem correctness was chosen because it is an immediate measure that has been shown in prior work to be an effective surrogate for learning, and it correlates with other measures of learning such as posttest score and mastery speed[13, 15–17]. Additionally, while one could use students' engagement with the videos as a dependent measure, e.g., number of videos requested or time spent watching videos, students' preferences do not always correlate with their learning [6, 17]. Therefore, next-problem correctness was chosen, as it provides an immediate and effective measure of learning.

4.1.1 Video Vs. No Video Analysis. To analyse the results of the first experiment, a mixed-effects logistic regression model [4] was fit to predict students next-problem correctness given the following inputs.

(1) A constant.
(2) The average correctness of the student across the prior weeks problems.
(3) The average correctness of the problem a skill-level video was provided (or not provided) for across the prior weeks instances of students completing the problem.
(4) The average correctness of the next problem used to calculate the dependent measure across the prior weeks instances of students completing the problem.
(5) A binary indication of whether or not the student was in the treatment (1) or control (0) condition.
(6) A random effect for each skill's impact on the treatment effect.

Inputs 2, 3, and 4 are all covariates meant to remove variations in the results from students with different prior knowledge and problems of different difficulty. Input 5 measures the average effect of offering students the opportunity to request a skill-related video, and each of the skill-level random effects in Input 6 measures the effect of offering students the opportunity to request a skill-related video for each skill separately. The random effects were included because each skill has a different set of five videos available for it, and it could be that some skills had very helpful videos while other skills did not, which would not be captured by Input 5.

The coefficient and statistical significance of Input 5 can be used to measure the impact of providing students with the opportunity to request skill-related videos on their performance, and the coefficients and statistical significance for the random effects can be used to determine the skill-level impact of this new feature.

4.1.2 BBTS Vs. Random Selection Analysis. To analyse the results of the second experiment, a mixed-effects logistic regression model [4] was fit to predict students next-problem correctness given the same inputs as the mixed-effects model for the first experiment but with the treatment variable now being whether or not BBTS (1) or random selection (0) was used to determine which video was made available to the student, and the following additional inputs.

(1) The number of recommendations made so far by the selected model for the given skill.
(2) The interaction between Input 1 and whether or not the student was in the treatment (1) or control (0) condition.
(3) A random effect for each skill's impact on Input 1.
(4) A random effect for each skill's impact on Input 2.

Unlike the first experiment, where we do not expect the effect of having a video available to change over time, we do expect the effect of the videos provided through BBTS to change over time because at the beginning of BBTS's use, it makes basically random recommendations, but over time, BBTS learns to recommend the most effective videos.

The coefficient and statistical significance of Input 2 captures this change over time and measures the impact of using BBTS to select videos compared to randomly selecting videos. The mixed effects in Input 4 capture how the impact of using BBTS to select videos differs across skills.

4.2 Video Feature Analysis

In addition to measuring the impact that videos and the methods used to select them had on student performance, this work used the data from the first experiment to investigate which features of videos made them more or less effective. A logistic regression [10] was fit using only the data from samples where students viewed the randomly recommended videos to predict students' next problem correctness given the following inputs.

(1) A constant.
(2) Random effects for the average effectiveness of videos for each skill.
(3) The average correctness of the student across the prior weeks problems.
(4) The average correctness of the problem a skill-level video was provided (or not provided) for across the prior weeks instances of students completing the problem.
(5) The average correctness of the next problem used to calculate the dependent measure across the prior weeks instances of students completing the problem.
(6) All of the video features except for Unknown Pronunciation Score and Unknown Tone.

In the regression, Inputs 1 and 2 allow for the average likelihood of getting the next problem correct after viewing a video to vary based on skill. This is important because different skills could be easier or harder to explain via video, and the model should be able to take this into account. Inputs 3, 4, and 5 are covariates to account for the variance in students' propensity to get the next problem correct. The video features "Unknown Pronunciation Score" and "Unknown Tone" were excluded from the logistic regression because every feature investigated for its impact on student learning increased the severity of the hypothesis correction used in this analysis, these two features would not have provided interpretable findings.

The coefficients and confidence intervals of the video features were used to determine if they had an impact on student performance. The Benjamini-Hochberg procedure [2] was used to correct the false discovery rate of significant features.

4.3 Opportunities for Personalization

In addition to exploring the impact that different video features had on students' performance, this work used the data from students that requested randomly selected videos to look for qualitative interactions between features of the videos and the students' prior knowledge. A qualitative interaction exists if one group of students benefits more from one type of content, while another group of students benefits more from a different type of content. For example, a qualitative interaction between students' prior knowledge and video length would exist if high-knowledge students got the next problem correct more often after viewing long videos and low-knowledge students got the next problem correct more often after viewing short videos. These qualitative interactions are each an opportunity to personalize students' learning. To identify any qualitative interactions in the data, the same method used in [15] to identify statistically significant qualitative interactions between students and the content available to them was used. Using this method, the regression $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \oplus x_2)$ is fit, where x_1 is a video feature, converted to a binary indicator of whether or not the value is above or below average for that feature, x_2 is a binary indicator of whether or not the student's prior correctness is above or below average, and y is the student's next problem correctness. Using this model, a qualitative interaction exists if β_3^2 is greater than β_1^2, which is derived with more detail in [15]. p-values for the statistical significance of these qualitative interactions were calculated using a bootstrapping approach in which the regression above was fit 10,000 times on different samples of equal size to the original data sampled with replacement from the original data. The distribution of $\beta_3^2 - \beta_1^2$ was used to perform a one-sample t-test to determine the p-value of the null hypothesis: $\beta_3^2 - \beta_1^2 \leq 0$. The p-values for the significance of different video features' qualitative interactions were corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure [2].

5 RESULTS

From March 3rd, 2022 to July 18th, 2022, 479,032 video recommendations were made to 18,267 students as they completed one of 27,589 problems. More problems were included in the experiments than were tagged for this work because some problems in ASSISTments were already tagged with their most relevant skill. On average, about 1,835 recommendations were made per skill, and each video was recommended an average of about 369 times. Unfortunately, out of all these recommendations, only 3,196 videos were actually requested by students. The vast majority of the time, students did not request videos. Compared to the about 15% of the time that students request problem-specific support, students only requested skill-related videos about 0.7% of the time.

Of the 2,383 students that requested at least one video, only 22% percent of those students requested a second, and less than

1% of those students requested at least 5 videos. Figure 4 shows this trend in skill-related video requests compared to problem-specific support requests. Students were not only less interested in skill-related videos from the start, but after requesting one video, students were much less likely to request another compared to the trend for problem-specific supports. Additionally, about 51% percent of the time that a video was requested, the problem-level support for the same problem was requested afterwards. Due to the intent-to-treat design of the randomized experiments, students' lack of interest in videos added a tremendous amount of noise to the results.

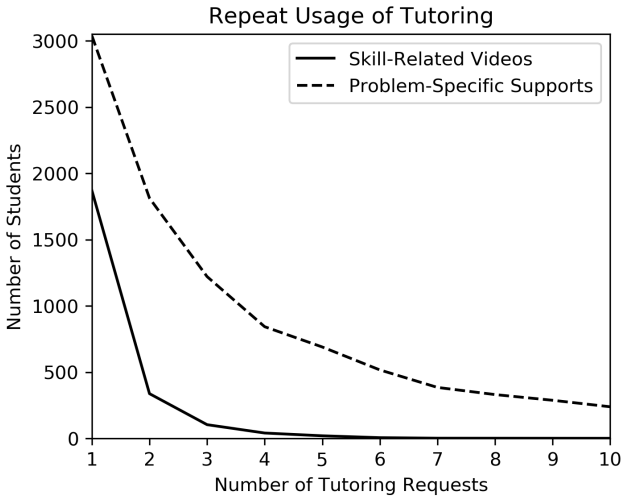

Figure 4: The number of students that requested from one to ten instances of tutoring for both skill-related videos and problem-specific support.

5.1 Video Vs. No Video

In the first experiment, 280,646 samples of a student being randomized when they started a problem between having the option to request a skill-related video or not were collected. In the control condition, there were 46,707 instances of one of 11,840 students completing one of 13,491 problems without the option to request a video. In the treatment condition, there were 233,939 instances of one of 16,974 students completing one of 23,119 problems with the option to request a video. There are more samples in the treatment than the control because students were randomized with equal probability to each of the five relevant videos or no video. Therefore, there are about five times more samples in the treatment condition than the control.

Using the model described in Section 4.1.1, the coefficient and 95% confidence interval for the average treatment of being shown a video was about 0.0002 ± 0.0250, which is far from being statistically significant. Figure 5 shows the coefficients and confidence intervals for the random effects of being offered a skill-related video for each skill separately, sorted from lowest to highest coefficient. Even when examining the effect of offering students skill-related videos on a per-skill basis, there were no significant effects. The model fit to determine these coefficients was a logistic regression, so the coefficients in Figure 5 should not be interpreted as effect sizes, they should solely be interpreted as indications that there were no statistically significant effects, which makes determining effect size moot.

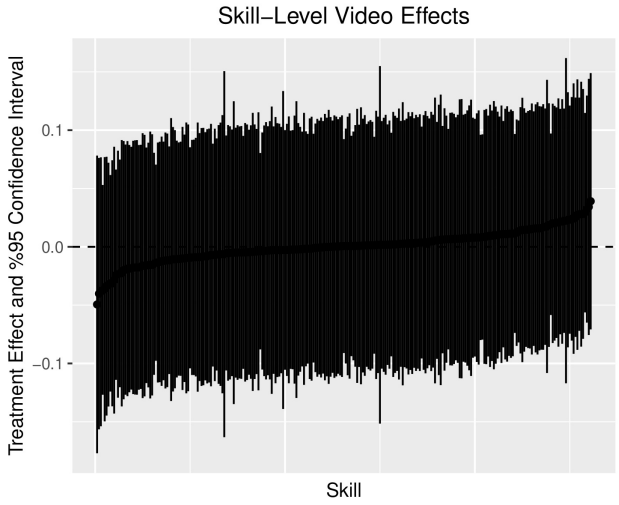

Figure 5: The coefficients and 95% confidence intervals for the random effects of offering students skill-related videos compared to not offering videos, sorted from lowest to highest coefficient.

5.2 BBTS Vs. Random Selection

In the second experiment, 559,917 samples of a student being randomized when they started a problem were collected. Students were randomized between BBTS or random selection determining which video (or lack thereof) they could request. In the control condition, there were 280,646 instances of one of 17,377 students completing one of 24,276 problems with the option to request a randomly recommended video. In the treatment condition, there were 279,271 instances of one of 17,309 students completing one of 24,315 problems with the option to request a BBTS recommended video. There are about an equal number of samples in each condition because students were randomized with equal probability to receive BBTS recommendations or random recommendations.

Using the model described in Section 4.1.2, the coefficient and 95% confidence interval for the average impact over time of using BBTS to recommend videos was about -0.10 ± 0.14, which is again, far from being statistically significant. Figure 6 shows the coefficients and confidence intervals for the random effects of the impact over time of using BBTS to recommend videos for each skill separately, sorted from lowest to highest coefficient. Even when examining the effect of using BBTS to recommend videos on a per-skill basis, there were no significant effects. The model fit to determine these coefficients was a logistic regression, so the coefficients in Figure 6 should not be interpreted as effect sizes, they should solely be interpreted as indications that there were no statistically significant effects, which makes determining effect size moot.

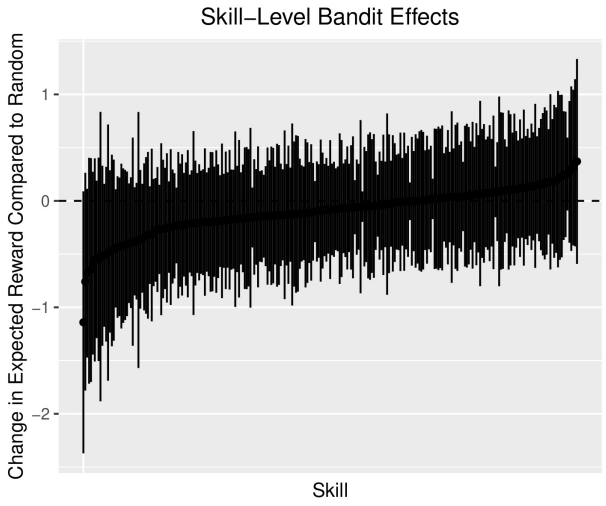

Figure 6: The coefficients and 95% confidence intervals for the random effects of the impact over time of using BBTS to recommend videos compared to randomly recommending videos, sorted from lowest to highest coefficient.

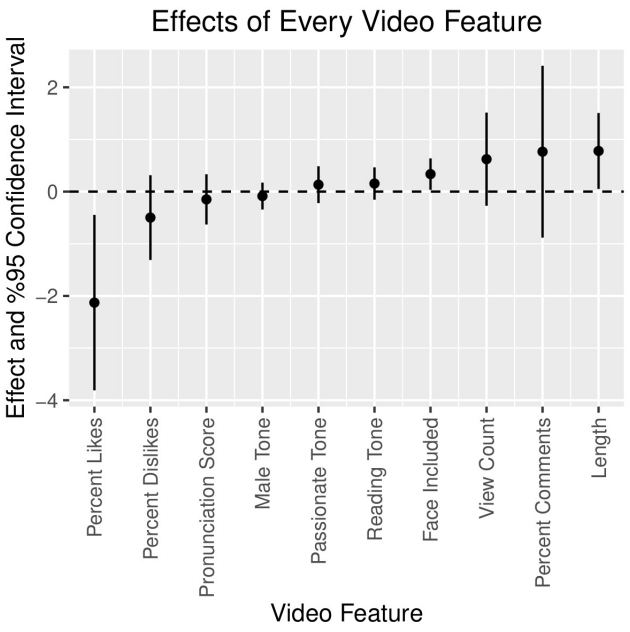

Figure 7: The coefficients and 95% confidence intervals for the impact of each video feature on students' propensity to get the next problem correct.

5.3 Video Features

In total, 1,677 randomly recommended videos were requested by 1,372 different users across 1,303 problems. Using the model described in Section 4.2, Figure 7 shows the coefficients and 95% confidence intervals for the video features. The confidence intervals in Figure 7 are calculated prior to any hypothesis correction. After hypothesis correction using the Benjamini-Hochberg procedure [2], none of the video features were significant predictors of students' next-problem correctness. The model fit to determine these coefficients was a logistic regression, so the coefficients in Figure 7 should not be interpreted as effect sizes, they should solely be interpreted as indications of which features were significant prior to correcting for multiple hypotheses.

5.4 Opportunities for Personalization

Using the methodology described in Section 4.3, of the ten potential qualitative interactions between students' prior knowledge and video features, two qualitative interactions were present and statistically significant. Both qualitative interactions are shown in Figure 8. In both plots, students with above-average prior correctness outperform students with below-average prior correctness on average, regardless of video features. However, students with below-average prior correctness benefited more from videos with above-average pronunciation scores and male toned speakers while students with above-average prior correctness benefited more from videos with below-average pronunciation scores and non-male toned speakers.

While these findings are statistically significant (both have $p < 0.001$ after correction), they are only correlational. If all other features of the videos were held constant, and the only difference was the speakers tone or pronunciation, then it would be possible to look for causality, but this is not the case for these skill-related YouTube videos. There are likely many covariates outside of the feature set created in this work that are correlated with pronunciation score and tone that effect these results. However, finding any opportunities to personalize students' learning at scale is rare, and it is interesting that even though so few students seemed to engage with the skill-related videos, there were still significant differences between the effectiveness of certain videos for specific groups of students.

6 DISCUSSION

From this work it seems that students are not interested in engaging with skill-related videos. It is unlikely that students were uninterested in the videos simply because they were videos because prior research in ASSISTments offered students a choice between video-based or text-based problem-specific support and found that about 29% of students chose the videos [5]. The presence of problem-specific support, which is more direct, relevant, and shorter, likely made students see the extra videos as a waste of time. Even though viewing the problem-specific support lowered students' scores while the skill-related videos did not, most students use ASSISTments for in-class work or homework assignments, which are generally low-stakes assignments meant to help prepare them for tests that are more impactful to their grades. Students might not care about their homework score and prioritize getting the most direct and relevant advice over general advice that may or may not be as helpful. An important distinction between the videos in this work and the videos in MOOCs is that MOOC videos are meant to be the primary instructional material, whereas in this work the videos were supplemental instructional material. This likely had an impact on students motivation to engage with

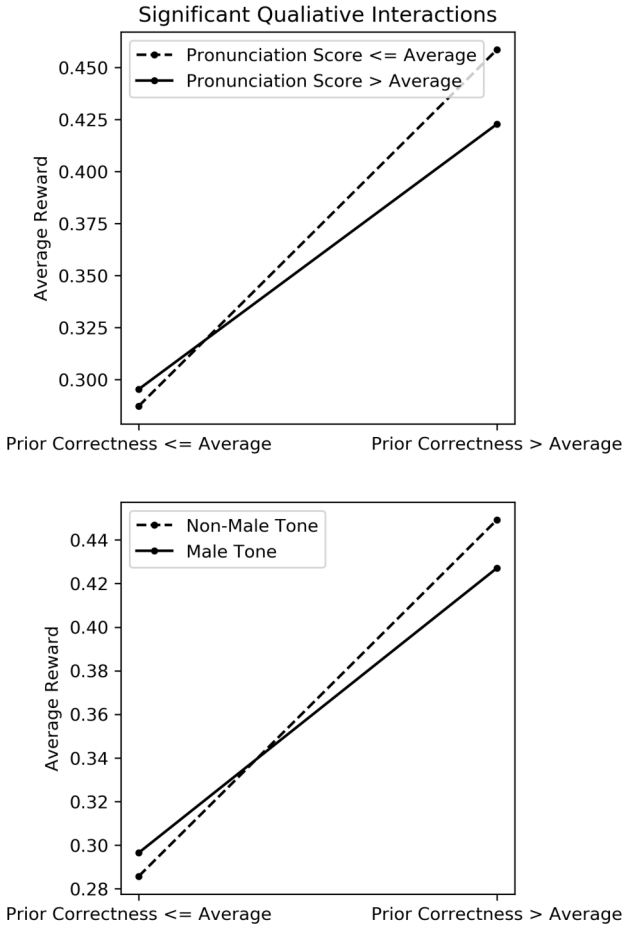

Figure 8: The two significant qualitative interactions between students' prior correctness and video features.

the videos because their teachers were probably providing them with primary instruction in a way they were more familiar and engaged with.

Regarding the analysis, using an intent-to-treat design made it very difficult to observe any effect of skill-related videos or of using BBTS to recommend them. Students only requested a video about 0.7% of the time. Unless seeing that a video is available but not requesting it effects students' propensity to get the next problem correct, 99.3% of the data in the treatment condition was equivalent to the data in the control condition. The amount of noise this adds to the analyses made the confidence intervals too large to see any effects, even on a per-skill basis.

By only including data from instances where students requested a randomly recommended video, this work was able to investigate the impact that different video features had on student performance. This part of the analysis was not an intent-to-treat design, and instead looked only at the impact that the videos had on the treated, i.e., the students that requested them. Interestingly, even though no video feature was a significant predictor of students' next-problem correctness, two video features, Male Tone and Pronunciation Score, had a significant qualitative interaction with students' prior correctness. These findings are almost certainly not causal because other features of the videos were not controlled for. Students with below-average prior correctness benefited more from videos with above-average pronunciation scores and male toned speakers while students with above-average prior correctness benefited more from videos with below-average pronunciation scores and non-male toned speakers. There were a handful of videos in this study in which a woman with a southern accent effectively explained a variety of mathematics skills. It likely that this woman, and similar content creators in the data, happen to explain concepts at a level that was more appropriate for students with higher knowledge, and because this woman has a lower pronunciation score and a non-male tone, the data reflects that these features have qualitative interactions with students' prior knowledge. In reality it is likely not the features themselves that led to these qualitative interactions, but the content creators that happened to correlate with those features.

7 LIMITATIONS AND FUTURE WORK

The results of these studies do not imply that skill-related videos are ineffective, but rather that there was no effect in this particular use case. This work only looked at the impact of skill-related videos on middle-school mathematics students within ASSISTments. It could be that without the problem-specific support that ASSISTments provides, skill-related videos would have a larger effect. It could also be that different age or socioeconomic groups are impacted differently than the population in this study. More studies should be conducted to investigate the impact of skill-related videos in different contexts, and to ensure that if there is an impact in a particular context, that this impact is fairly distributed amongst different groups of students.

While the intent-to-treat analysis was necessary to unbiasedly compare videos to no videos, it was not as necessary to investigate the impact of using BBTS to recommend videos compared to random selection. If BBTS was not allowed to recommend no video, then BBTS could have been updated only when students actually requested videos, and these samples could have been compared to only the times that students requested randomly recommended videos. This would have likely resulted in a larger effect by removing about 99.3% of the data used to updated the BBTS model in which students never requested videos. This would have allowed the BBTS model to learn the trends in the data more easily, and would have likely led to a larger difference over time between BBTS recommendations and random recommendations. Moving forward, more experiments comparing BBTS to random selection in ways that are more fair to BBTS should be conducted.

Additionally, better covariates for predicting students' next-problem correctness could be created to help remove some of the noise in the intent-to-treat analysis. The covariates used in all the models in this work for students' prior knowledge and problem difficulty had Pearson correlations [14] with students' next-problem correctness of only around 0.2. Serious work could be done to thoroughly investigate different combinations of student and problem past performance measures in order to create more predictive covariates.

Lastly, the videos in this work were collected from YouTube via algorithmic searches and teacher ratings. If, in the future, one wished to perform a causal analysis of the significance of different video features and their qualitative interactions with students, it would be better to create the videos from scratch. If everything except one video feature of interest was held constant, the analyses in Sections 4.2 and 4.3 could be regarded as causal for that feature.

8 CONCLUSION

Overall, it did not appear that offering students the option to request skill-related videos had a positive impact on their performance. This mostly stemmed from students' lack of interest in the skill-related videos. Students only requested a skill-related video about 0.7% of the time, compared to the about 15% of the time that they requested problem-specific tutoring, which implies that students prefer concise advice directly related to the task at hand, regardless of the impact it has on their score. Although this work did not find any significant impact of providing skill-related videos to students, it was able to analyse which features of videos correlated most with students' performance when they did request a video. This analysis found that while there were no video features that significantly predicted students' performance, there were two video features that had qualitative interactions with students' prior knowledge. These qualitative interactions implied that particular content creators created videos that were more helpful for higher-knowledge students, while other content creators made videos that were more effective for lower-knowledge students. Moving forward, the educational research community can take away two main findings from this work. The first is that students are unlikely to be interested in content that they do not see as directly relevant to them. Therefore, when creating or curating tutoring for students, taking the effort to ensure each piece of content is direct and relevant is likely to pay off. Secondly, it seems possible to create videos that are better for higher or lower knowledge students. This should motivate randomized controlled studies to determine which aspects of video based learning specifically influence videos' effectiveness for different groups of students. Uncovering the causal mechanisms behind these qualitative interactions paves the way for more effective forms of personalized learning.

ACKNOWLEDGMENTS

We would like to thank NSF (e.g., 2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, & 1535428), IES (e.g., R305N210049, R305D210031, R305A170137, R305A170243, R305A180401, & R305A120125), GAANN (e.g., P200A180088 & P200A150306), EIR (U411B190024 & S411B210024), ONR (N00014-18-1-2768), NHI (R44GM146483), and Schmidt Futures. None of the opinions expressed here are that of the funders.

REFERENCES

[1] Murat Akkus. 2016. The Common Core State Standards for Mathematics. *International Journal of Research in Education and Science* 2, 1 (2016), 49–54.
[2] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300.
[3] Joshua A Dijksman and Salman Khan. 2011. Khan Academy: the world's free virtual school. In *APS March Meeting Abstracts*, Vol. 2011. A14–006.
[4] Andrew Gelman and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.
[5] Ashish Gurung. 2021. *Examining Student Effort on Hint through Response Time Decomposition*. Ph. D. Dissertation. Worcester Polytechnic Institute.
[6] Aaron Haim and Neil Heffernan. 2022. Student Perception on the Effectiveness of On-Demand Assistance in Online Learning Platforms. In *Educational Data Mining Conference*.
[7] Neil T Heffernan and Cristina Lindquist Heffernan. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education* 24, 4 (2014), 470–497.
[8] Nina Hollender, Cristian Hofmann, Michael Deneke, and Bernhard Schmitz. 2010. Integrating cognitive load theory and concepts of human–computer interaction. *Computers in human behavior* 26, 6 (2010), 1278–1288.
[9] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. 2012. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*. Springer, 199–213.
[10] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. 2002. *Logistic regression*. Springer.
[11] Rebecca Mullen and Linda Wedwick. 2008. Avoiding the digital abyss: Getting started in the classroom with YouTube, digital stories, and blogs. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas* 82, 2 (2008), 66–69.
[12] Korinn Ostrow and Neil Heffernan. 2014. Testing the multimedia principle in the real world: a comparison of video vs. Text feedback in authentic middle school math assignments. In *Educational Data Mining 2014*.
[13] Thanaporn Patikorn and Neil T Heffernan. 2020. Effectiveness of crowd-sourcing on-demand assistance from teachers in online learning platforms. In *Proceedings of the Seventh ACM Conference on Learning@ Scale*. 115–124.
[14] Karl Pearson. 1895. VII. Note on regression and inheritance in the case of two parents. *proceedings of the royal society of London* 58, 347-352 (1895), 240–242.
[15] Ethan Prihar, Aaron Haim, Adam Sales, and Neil Heffernan. 2022. Automatic Interpretable Personalized Learning. In *Proceedings of the Ninth ACM Conference on Learning@ Scale*. 1–11.
[16] Ethan Prihar, Thanaporn Patikorn, Anthony Botelho, Adam Sales, and Neil Heffernan. 2021. Toward Personalizing Students' Education with Crowdsourced Tutoring. In *Proceedings of the Eighth ACM Conference on Learning@ Scale*. 37–45.
[17] Ethan Prihar, Manaal Syed, Korinn Ostrow, Stacy Shaw, Adam Sales, and Neil Heffernan. 2022. Exploring Common Trends in Online Educational Experiments. In *Proceedings of the 15th International Conference on Educational Data Mining*. 27.
[18] Anna Rafferty, Huiji Ying, Joseph Williams, et al. 2019. Statistical consequences of using multi-armed bandits to conduct adaptive educational experiments. *Journal of Educational Data Mining* 11, 1 (2019), 47–79.
[19] Jan Renz, Matthias Bauer, Martin Malchow, Thomas Staubitz, and Christoph Meinel. 2015. Optimizing the video experience in moocs. In *EDULEARN15 Proceedings*. IATED, 5150–5158.
[20] Herbert Robbins. 1952. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* 58, 5 (1952), 527–535.
[21] Jeremy Roschelle, Mingyu Feng, Robert F Murphy, and Craig A Mason. 2016. Online mathematics homework increases student achievement. *AERA open* 2, 4 (2016), 2332858416673968.
[22] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. 2018. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning* 11, 1 (2018), 1–96.
[23] Daniel T Seaton, Sergiy Nesterko, Tommy Mullaney, Justin Reich, Andrew Ho, and Isaac Chuang. 2014. Characterizing video use in the catalogue of MITx MOOCs. *Proceedings of the European MOOC Stakeholder Summit* (2014), 140–146.
[24] Doraisamy Gobu Sooryanarayan and Deepak Gupta. 2015. Impact of learner motivation on mooc preferences: Transfer vs. made moocs. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 929–934.
[25] Jiayu Yao, Emma Brunskill, Weiwei Pan, Susan Murphy, and Finale Doshi-Velez. 2021. Power Constrained Bandits. In *Machine Learning for Healthcare Conference*. PMLR, 209–259.
[26] Yang Zhi-Han, Shiyue Zhang, and Anna Rafferty. 2022. Adversarial bandits for drawing generalizable conclusions in non-adversarial experiments: an empirical study. In *Proceedings of the 15th International Conference on Educational Data Mining*. 353.