The Impact of Engagement with User-generated Content in Online Discussion Forums

Varad Deolankar[1], Ali Goli[2], S. Sriram[3], and Pradeep K. Chintagunta[4]

[1]University of Michigan, varadd@umich.edu

[2]University of Washington, agoli@uw.edu

[3]University of Michigan, ssrira@umich.edu

[4]University of Chicago, pradeep.chintagunta@chicagobooth.edu

Abstract

Many online platforms that offer a core product for consumption also provide a space for the creation and consumption of user-generated content (UGC). Conditional on such an arrangement, we study whether a user's engagement with UGC has an impact on their engagement with the core product using a novel dataset from Coursera, a popular online education platform. Coursera provides its core product in the form of course material embedded within lecture videos and supplementary readings and provides a space for UGC creation and consumption via discussion threads about course materials on a discussion forum. We leverage the exogenous variation in the forum homepage as a shifter of a user's propensity to engage with UGC (i.e., the discussion forum). In particular, we use three characteristics of the content presented on the forum homepage, popularity, relevance, and variety, and use these as potential shifters of engagement with UGC. This allows us to identify the causal effect of a user's engagement with UGC on their engagement with the core product and their mastery of the content covered in that course segment. Our analysis suggests that, on average, a 10% increase in a student's UGC engagement increases that student's engagement with the core product by 4.32% and performance by 0.60%. Our findings for this platform suggest that users tend to view UGC and the core product as complements. In the specific context of online education platforms, our results suggest that discussion forums bolster student engagement with the course content as well as their learning. In this regard, we also find that the popularity and relevance of threads listed on the home page of the discussion forum have a positive effect on user engagement therein.[1]

[1]We thank seminar participants at the Department of Marketing and the Department of Economics at the University of Michigan, and attendees of the Marketing Science Conference, Chicago 2022. This paper has also benefited substantially from thoughtful comments made by Jessica Fong and Justin Huang. All errors are our own.

INTRODUCTION

Many online platforms, in addition to the main product that they offer, also provide the space for the creation and consumption of user-generated content (UGC), usually in the form of discussion forums. For example, news platforms provide curated information on current events and present opinion pieces in the form of editorials. These constitute the main product that the news platforms offer to their users. At the same time, they also allow users to post their comments and elicit reactions from their peers (a form of UGC) in their discussion forums. Other examples where the core product and UGC coexist include social news aggregation websites such as Reddit, online gaming platforms (e.g., Gaia and Zapak), fitness websites (e.g., Bodybuilding.com and MyFitnessPal), and online education platforms (e.g., Coursera, EdX, Udemy, etc.).

In light of the coexistence of the core product and UGC within the same platform, a natural question that arises is whether the two types of content are complements or substitutes. On the one hand, engagement with user-generated content can enhance the value derived from the main product and vice versa. This would render the two types of content as complements (Duke, 2023; Greenslade, 2015; Neeraj, 2023). On the other hand, consumption of one type of content could potentially distract users from consuming the other content, probably because they compete for the same available time (or other resources), thereby making them substitutes. Moreover, the borderless nature of discussion forums means that it can be hard to avoid the less useful content embedded within (Sunyer, 2014). Propelled by such concerns, some news websites (eg: Recode, Reuters, The Week, Mic, The Verge, Popular Science, Radio New Zealand, etc.) are drastically reducing support for UGC stating that it is not their primary product and that moderating it poses a major challenge (Ellis, 2014); (Goujard, 2016). Therefore, in contexts such as media consumption and online education where both effects are likely to occur, the direction of the relationship is ambiguous. If the core product and UGC are complements (substitutes), increasing user engagement with UGC will increase (decrease) their engagement with the core product. This can have downstream implications for the value that users derive from the core product, and consequently, their retention.

In this paper, we ask whether engagement with UGC has a spillover effect on that user's engagement with the core product. If such spillover effects exist, their direction (i.e., positive vs. negative) would tell us if users view the two types of content as complements or as substitutes. If the two forms of content are perceived as complements (substitutes), increasing engagement with one would increase (decrease) engagement with the other. Therefore, if we can identify factors that drive engagement with UGC, we can use these as levers that can also influence engagement with the core product. We perform our empirical analysis using data from the online education platform, Coursera. Education platforms such as Coursera collaborate with academic institutions and industry partners to offer online courses, which constitute their core product. At the same time, these platforms typically establish forums where participants can browse the discussions of past students and participate in discussions with their peers while learning (a form of UGC).

Isolating the relationship between user engagement with UGC and the core product (educational material in our context) using observational data is likely to be challenging for several reasons. First, users are likely to be different in terms of their level of motivation as well as their ability to master the course material. These differences can influence engagement with both types of content and therefore render them correlated. Second, courses (as well as different modules within a course) may differ in terms of the level of engagement that they require from their users. These differences can result in users spending more time on both types of content in some courses (or in some modules within a course) and less in others. Third, the overall proficiency of the users with respect to the educational content as well as in navigating the various learning objects might evolve over time. Similarly, the courses as well as the modules within these courses may have evolved over time in a way that would have influenced the benefit of joint consumption of the two types of content. The three factors noted above can induce a correlation between the consumption of the two types of content. Separating this correlation induced by user heterogeneity, course design, and accumulated experience, from the causal relationship that we seek to uncover, can be challenging. Finally, even if we can control for these correlated unobservables that drive engagement with both types of content, it might be difficult to isolate the direction of causality, i.e., did engagement with UGC drive engagement with the course material (i.e., the main product) or vice versa?

We use granular clickstream data from ten courses offered on the Coursera platform. Our data span 63 months from January 2015 through March 2020. During this period, each course had several cohorts of students. Furthermore, each course is broken down into multiple modules (an average of approximately 8 modules per course), with each module having its own course material (i.e., the main product) and discussion forum (i.e., UGC). For each user, our data enable us to infer the time that they spent on the educational material as well as on the discussion forums at the module level. The granularity of these data allows us to include a rich set of fixed effects that account for several confounds in the analysis; they control for cross-sectional differences (i.e., between users, courses, and modules within a course) as well as time-varying factors (e.g., changes in course content, differences across cohorts, etc.) that may induce correlation between engagement with the two types of content.

To isolate the causal effect of user engagement with one type of content on their engagement with the other type of content (i.e., pin down the direction of the causal effect), we need exogenous variables that have a direct effect on the former but not the latter. We exploit a unique aspect of the discussion forums to help with this identification. In particular, the discussion forums list multiple threads in order of the recency of the activity in each thread. As a result, the threads on the home page for the discussion forum for a module can change over time based on the recency of their activity. Thus, two users visiting the forum a few minutes apart could see different sets of threads listed and ordered on the home page. We consider three characteristics that could shift the attractiveness of content that is immediately visible to the user upon visiting the discussion forum at any given point in time - popularity, relevance, and variety of topics represented in these threads. We show that the propensity among users to click on the discussion threads within a forum (a metric of engagement with UGC) is a function of these characteristics of the threads listed on the first page. Since the threads that are displayed on the home page and the order in which they are listed can be considered as being assigned at random given the specific arrival time of a user, we use the three characteristics as exogenous shifters of the user's propensity to engage with the discussion forum (i.e., the UGC). At the same time, these shifters should not have a direct effect on their engagement with the course material (i.e., the main content) for that module.

Our results suggest that engagement with the discussion forum (i.e., UGC) increases the marginal utility from the consumption of the course material (i.e., the main product). On average, a 10% increase in engagement with UGC increases engagement with the main product by 4.32%. Therefore, users view UGC and the main product as complements. Furthermore, our results suggest that a given user's engagement with UGC has a positive impact on their performance within the course. On average, a 10% increase in engagement with UGC increases student performance by 0.60%. Overall, these results imply that, in the context of MOOC platforms, having a vibrant discussion forum (i.e., UGC) can help in increasing user engagement with the course material (i.e., the main product) as well as their performance in these courses. Therefore, online education platforms ought to internalize this benefit of discussion forums while making their design decisions. In this regard, our first stage results (i.e., the effect of the three exogenous drivers on engagement in discussion forums) provide some valuable insights on possible drivers of engagement in discussion forums. In particular, we find that the popularity and relevance of threads listed on the home page of the discussion forum have a positive effect on user engagement therein. On the other hand, we do not find a significant effect of content variety on user engagement with the discussion forum.

The rest of the paper is organized as follows. First, we will discuss the related literature.

Second, we will describe our dataset and discuss our research design in greater depth. Next, we will present our empirical analysis, and discuss the results and robustness checks. Finally, we will conclude with the implications of our findings and the limitations of our study.

RELATED LITERATURE

Firms are interested in increasing user engagement with their content because it has implications for retention and monetization. The extant literature has studied the drivers of user engagement with content in many contexts, ranging from entertainment, news, gaming, social media, and online education. Depending on the context, engagement can be measured in several ways such as time spent actively consuming the content, commenting on various aspects of the content, and sharing the content with others. We discuss what we know from the literature regarding drivers of user engagement with content. Given that our empirical investigation is in the context of online education, we discuss the literature corresponding to this setting separately.

Drivers of User Engagement

A natural way to increase engagement is by modifying the way content is designed. In the context of social media platforms, design can play an important role in driving engagement with the content posted by a brand. For example, Lee et al. (2018) find that content related to brand personality, such as humor and emotion are associated with higher engagement. On the other hand, messages with informative content such as price and deals are associated with lower engagement when present in isolation. However, when the informative content is presented along with brand personality-related attributes, engagement is higher. In a similar vein, Li and Xie (2020) find that the mere presence of image content in branded posts on Twitter affects user engagement. In addition to influencing engagement with the content posted by brands, design features can also drive peer influence. For example, using a randomized field experiment, Aral and Walker (2011) show that by designing viral features into products and marketing campaigns, firms can generate economically identifiable peer influence and contagion effects.

Researchers have also reported the role of product design in driving user engagement in contexts other than social media. In the context of multiplayer video games, Huang et al. (2019) show that matching players in game rounds can be an effective tool for increasing their engagement. Gu et al. (2022) study the impact of crowdsourcing features on user engagement in the context of a mobile gaming app. The authors report that giving users the opportunity to submit content increases their engagement and retention by empowering users to control their product experience. On the other hand, allowing users to access crowdsourced content does not have a significant effect on engagement, although it does improve retention. However, the authors find that allowing users to submit content as well as access crowdsourced content lowers engagement. Therefore, the ability to submit and access content are not complements.

In addition to product design, the literature has also documented the effect of the elements of the marketing mix, such as pricing and promotions, in driving user engagement. In the context of online newspapers, Pattabhiramaiah et al. (2019) show that the implementation of a paywall by the New York Times in 2011 resulted in a decrease in user engagement. In the context of social media, marketers routinely use influencers to increase user engagement for their products and related postings. In this context, Valsesia et al. (2020) show that conditional on having many followers, influencers who follow few others are more effective in generating user engagement than those who follow a large number of other users. While these interventions consider actions taken by firms, researchers have also investigated the effect of UGC in driving engagement with the brand. For example, Hartmann et al. (2021) study the effects of two types of selfie images on social media on engagement - consumer selfies and brand selfies. The authors find that while brand selfies increase engagement with the brand, consumer selfies increase engagement with the sender. Their findings have implications for managers seeking to use consumer selfies on social media to increase engagement with their brands. Still, in the context of UGC, Ksiazek et al. (2016) study the generation of UGC, which can also be viewed as a metric of engagement. The authors study how the exposures of YouTube news videos are associated with the number of user-content and user-user interactions generated by such videos. They find that an increase in a video's exposures is associated with an increase in the number of comments received by that video (UGC in the form of user-content interactions). In contrast, a decrease in a video's exposures is associated with an increase in the number of conversations among users in the comments section of that video (UGC in the form of user-user interactions).

In recent years, organizations have explored ways to enrich the user experience by developing platforms (websites) through which users generate and contribute content (Di Gangi and Wasko, 2009). As a result, researchers have tried to understand the consequences of customer engagement with UGC residing on these platforms in a variety of different contexts. In the context of politics, Houston et al. (2013) find that tweeting while watching a presidential debate was related to participants reporting more favorable attitudes about the democratic candidate, paying more attention to the debate, and perceiving debates to be more important. However, such behavior was not related to enjoying the debate more. Berman et al. (2019) shed further light on these findings in subsequent research that shows that sharing content on Twitter during an election debate is positively associated with engagement with the live event among sharers.

In addition to the association between UGC creation and engagement, researchers have also studied the relationship between the recognition of the UGC created by a user and their subsequent engagement with the platform. For example, Lu et al. (2022) study the effect of users receiving badges for posting UGC. They find that users increase their content generation, but decrease their content consumption immediately after receiving a digital badge. However, the authors report enduring positive effects on content generation and consumption as a result of badging. In a similar study, Huang et al. (2022) conduct a field experiment to study the relationship between the attention and recognition that a user's content receives on their subsequent engagement with the platform. The authors report that increasing the attention and recognition of a user's content can increase their engagement with the platform. Together, these results suggest that being recognized by their peers or the platform for their contribution can be an effective lever in increasing their engagement with the platform.

Drivers of User Engagement in Online Courses

Despite their promise to revolutionize education, online courses have fallen short on account of low user engagement and completion rates (Khalil and Ebner, 2014; Koller et al., 2013; Onah et al., 2014). As a result, there is a growing body of research investigating the drivers of user engagement in online education. Given that online courses are open to all learners, they attract a wide variety of participants, who differ considerably in terms of their ability and motivation. Banerjee and Duflo (2014) find that the ability of students to stay organized and force themselves to complete tasks on time is an important determinant of student engagement and performance. In addition, as Lu et al. (2022) document, participants in online courses differ considerably in terms of their learning styles. Therefore, some of the deficiencies in the effectiveness of online courses can be traced back to the characteristics of their participants. Although these intrinsic traits of participants may seem to be immutable, Patterson (2018) shows that they can be altered using behavioral interventions. Using a randomized control trial, Patterson (2018) shows that tools like commitment devices can increase student engagement in online courses.

As with the literature on the drivers of user engagement with content, researchers have also considered the link between course design and engagement. In one of the early studies on online courses, Conrad (2002) performed a descriptive investigation of the correlates of engagement. The author found that the experience in the first class or the beginning of an online course will contribute to the sense of well-being and engagement among learners. More recently, using video data from online education platforms such as Masterclass and Crash Course, Zhou et al. (2021) show that basic video characteristics (e.g., length, speaking rate, etc.), instructor's emotion and physical characteristics, and visual aesthetic features (e.g., brightness, warm hue proportion, clarity, etc.) are important predictors of engagement. Although they fall short of claiming this relationship to be causal, their results suggest that online education platforms could potentially increase engagement by adjusting the design of course videos. Extant research studying the effect of gamification as a driver of user engagement has also reported some interesting findings. While Hamari et al. (2014) reveal positive effects of gamification, Van Hentenryck and Coffrin (2014) suggest that one-size fits all style gamification programs may be perceived as annoying and can even induce users to abandon the learning platform.

In addition to content design, online education platforms and the courses therein can influence engagement using other levers that are not directly related to the content itself. For example, Goli et al. (2022) demonstrate that the mere act of paying for an online course can trigger the sunk cost effect and therefore increase the engagement of participants. Therefore, they suggest that pricing can potentially serve as a useful tool to increase engagement in online courses. Closer to our research context, researchers have also studied the relationship between participation in the discussion forum and engagement with the course content. Using an experiment that nudges users to share ideas on forums, Narang et al. (2022) examine how content sharing on the discussion forums in online courses affects the engagement of content "creators". Using a similar experimental design where users are nudged to visit the discussion forum, Zhang et al. (2017) study how engagement with content on the forums affects the quiz completion rates and performance of content "consumers" in online courses. The idea behind both these studies is that the randomly assigned nudge to engage with the discussion forum will result in increased engagement, which, in turn, can impact engagement with content.

Similar to Narang et al. (2022) and Zhang et al. (2017), our investigation is related to the idea of increasing engagement with the discussion forum and its consequence on the consumption of the course material. However, our approach differs from Narang et al. (2022) and Zhang et al. (2017) by studying the effect of exogenous variation in the intrinsic attractiveness within the discussion forum once the participant has visited the forum. In general terms, there are two ways to increase engagement with the discussion forum: (a) nudging users to engage with the forum and (b) improving the attractiveness of the content within the forum. The experimental design employed by Narang et al. (2022) and Zhang et al. (2017) demonstrates the effectiveness of (a) on engagement. Our approach complements this by considering the influence of (b) in driving engagement. As noted above, we consider three potential aspects of discussion forums that may impact user engagement. Therefore, while their approach sheds light on the effect of nudges to participate in the discussion forum, our approach has implications for the design of the forum to increase participation therein.

DATA DESCRIPTION AND MODEL-FREE EVIDENCE

We will first present a brief description of the data used in our analysis and present some summary statistics. Next, we will provide descriptive evidence for our thesis that engagement with UGC can have consequences for engagement with course material and performance. Finally, we will discuss the challenges in inferring the causal relationship between a user's engagement with the forum and their engagement with the course material.

Data Description

Our data come from ten "on-demand" courses offered by a large public university on Coursera. The observation period in our study is from January 2015 to March 2020. During this period, the courses in our panel did not undergo any major modifications, either in terms of content or structure. A total of 48,127 students participated in these courses.

On this platform, participants consume the course content in a self-paced manner. Thus,

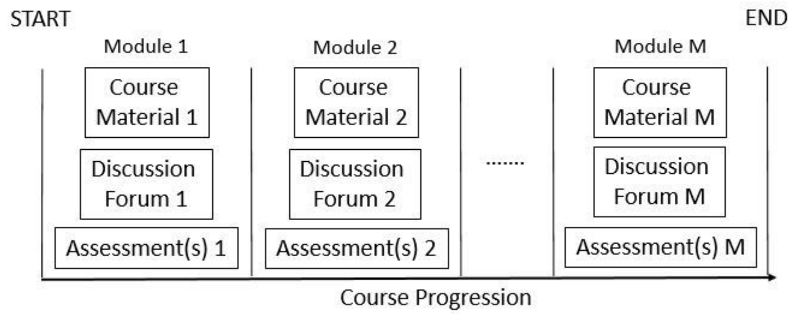


Figure 1: Course Structure on Coursera

unlike in a regular classroom setting, each enrolled student progresses through a given course at their own pace. Each course is structured as a collection of modules. Within each module, students are expected to complete a set of lecture videos and some supplementary readings and activities. Once participants have completed the course material for that module, they take one or more assessments based on the topics covered in that module. In addition, each module has its own discussion forum where participants can discuss and ask questions about the course material covered in that particular module by either creating their own threads or by adding to existing threads. In addition, participants can merely engage with existing threads. We present the typical structure of a course offered on this platform in Figure 1.

Our dataset is longitudinal because we observe how an enrolled student progresses through the various modules within a course. We use two kinds of data in our empirical analysis. The first kind is the click stream data, which enables us to track the time spent by each student on the course material (i.e., the main product) corresponding to the various modules within the course. As noted above, each module has its own course material and discussion forum. The course material come in the form of lecture videos, readings, and other activities. In the discussion forum (i.e., the UGC), participants can engage in three different ways: (a) starting new threads, (b) adding to existing threads, and (c) consuming content in the existing threads. The click stream data allow us to infer the time spent by each user in accessing content in the module-specific course material. We use this time spent by each participant in consuming content as a measure of engagement. With regards to the discussion forums, we use the number of discussion threads that a user clicked (i.e., (c) above) as a metric of their engagement.[2]

The second kind of data we use includes information on each participant's performance. For each participant, in every module, we observe their performance in the assessments for that module. For modules that have multiple assessments, we compute student performance as a weighted average of the grade in each assessment. Together, the two kinds of data give us three key metrics that are related to a student's experience with the course: (a) engagement with the course material (i.e., the main product), (b) engagement with the forum (i.e., the UGC), and (c) their performance in the course. As discussed above, our objective is to isolate the causal effect of (b) on (a), and subsequently on (c).

Summary Statistics

We present the summary statistics that describe the distribution of the key pieces of information across the ten courses in Table 1. The first two rows represent the characteristics of the courses. The information in Table 1 suggests that a typical course has about 7.6 modules and 4812 students enrolled. However, there is considerable heterogeneity across courses on both these dimensions. The remaining rows reflect different aspects of user behavior within these courses in terms of their (a) engagement with the course material (i.e., the main product), (b) engagement with the discussion forum (i.e., the UGC), and (c) completion of the course and performance therein.

With regards to user engagement with the course material (i.e., the main product), a typical student spent a little over two hours (125 minutes) per module. For engagement with UGC, we find that an average of 787.4 students (16.4% of those registered in the course) visited the discussion forum corresponding to at least one of the modules within the course. Of those who visited the discussion forum, an average of 532.8 (67.6% of those who visited the forum) clicked on at least one of the threads and viewed the content therein.

[2]Subsequently, we verify the robustness of the results by modifying the metric to include all three modes of engagement, i.e., (a)-(c). The result of this robustness check is reported in Appendix A.

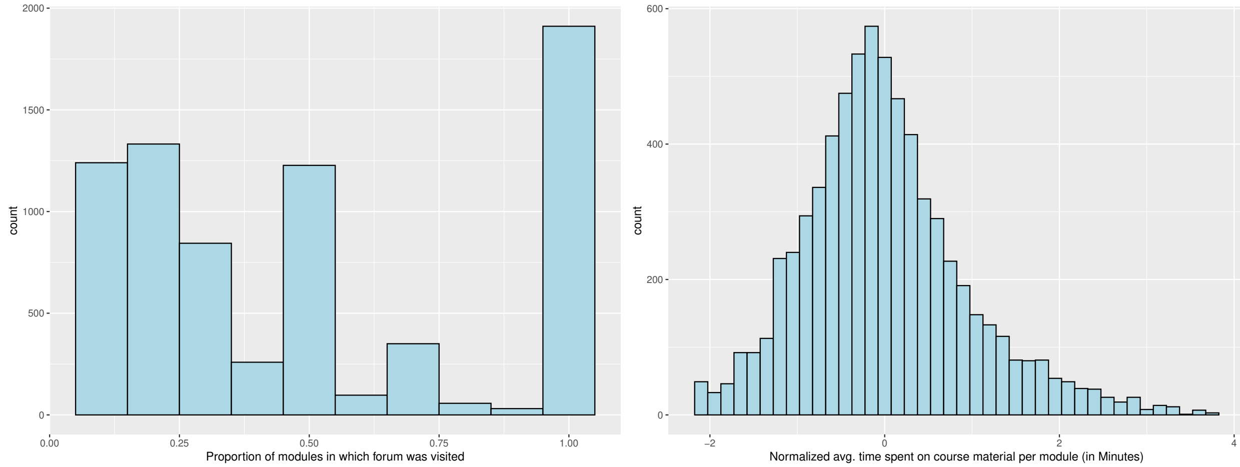
Table 1: Descriptive statistics across the ten courses

Statistic	Mean	Std. Dev.	Min	Max
Number of students	4,812.700	5,278.634	653	14,519
Number of modules	7.600	2.875	4	12
Avg portion of course completed	0.390	0.121	0.179	0.601
Avg Time spent on content per module (in minutes)	125.122	45.755	77.129	218.160
Number of Forum Visitors	787.400	1,036.310	57	3,081
Number of Thread Viewers	532.800	742.315	19	2,267
Percentage of students completing the course	17.5	9.1	3.2	32.3
Average overall course grade (out of 100)	38.104	13.832	17.872	57.301

The final two rows of Table 1 shed light on the course completion rates and the performance of participants within the course. These summary statistics reveal that only about 17.5% of participants taking a course ended up completing it. There is considerable heterogeneity across courses, ranging from a low completion rate of 3.2% for some courses to a high of 32.3%. Similarly, the average grade for the participants was 38.104%, with a range of 17.872% to 57.301% across courses. Together, the data in Table 1 highlight the fact that courses differ considerably in terms of their design, student engagement with the course material and the discussion forum, and their performance in the course.

Let us now take a closer look at the distribution of some of the key metrics across individuals and modules within courses. First, we consider heterogeneity across individual users in terms of their engagement. We consider users that had at least one engagement with the forums and present the distribution of the proportion of modules within a course where each user visited the discussion forum in Figure 2a. The information in these figures suggests that there is considerable heterogeneity across users in terms of their propensity to engage with the discussion forum as well as the course material as shown in Figure 2b. In particular, while some users consistently engage with the discussion forum, others are largely dormant. This heterogeneity in engagement across users is probably driven by differences in their intrinsic motivation to engage and/or in their ability to master the material.

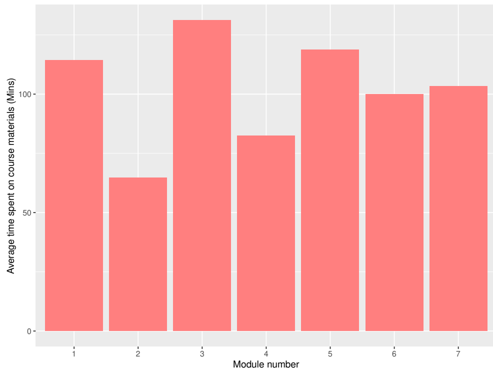
Next, we consider how the engagement metrics vary across modules within a course. In Figure 3, we present this information for one course in our data. Once again, the data



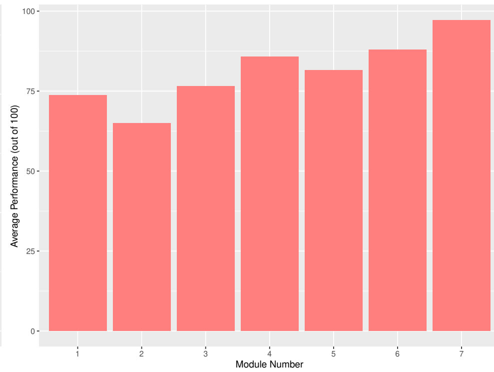
(a) Discussion forums (b) Course material

Figure 2: Users differ in terms of their intrinsic tendency to engage with course discussion forums and material.

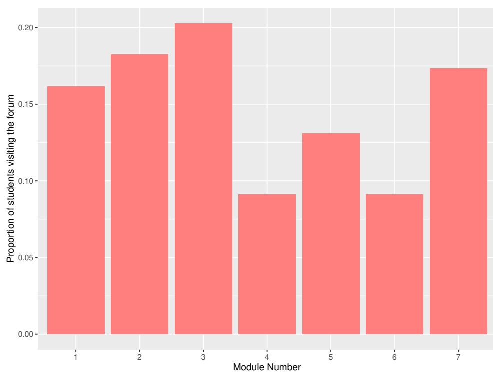
suggest that modules within a course differ considerably in terms of the amount of time participants spend consuming course material, visiting the discussion forum, viewing threads in the discussion forum, and engaging with the user-generated content. This heterogeneity is probably attributable to differences in the need for engagement (i.e., due to the nature of the course material) or to the ability of the material to foster engagement (i.e., due to design). As we discuss subsequently, we need to control for such heterogeneity across individuals and modules while parsing out the causal effects of interest.



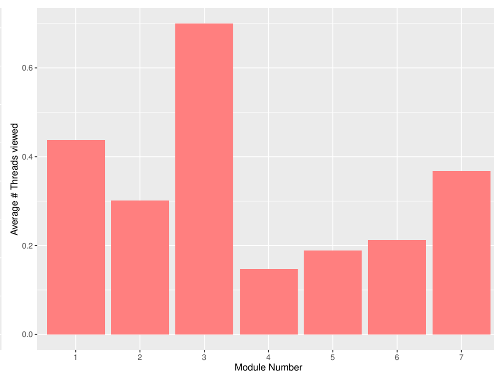
(a) Time spent on course material



(b) Performance



(c) Portion of students visiting the forum



(d) Threads viewed

Figure 3: Variation in engagement with the core product, UGC, and performance across different modules within a given course.

Lastly, we also consider how student behavior and outcomes evolve over the life cycle of a course. In Figure 4, we present this information for one module within a course as a function of its age (the x-axis). For this particular module, the data suggest that the engagement metrics and course outcomes increase with the age of the course. This evolution could possibly be attributed to growing proficiency in navigating online courses among learners. In our empirical analysis, we control for such potential confounding due to the age of the course.[3]

[3] Another explanation is that the design of the course changed over time in such a way that it fostered greater engagement and also resulted in higher performance. While this explanation does not invalidate our identification strategy, we did verify that the courses used in our study did not undergo any major design changes.

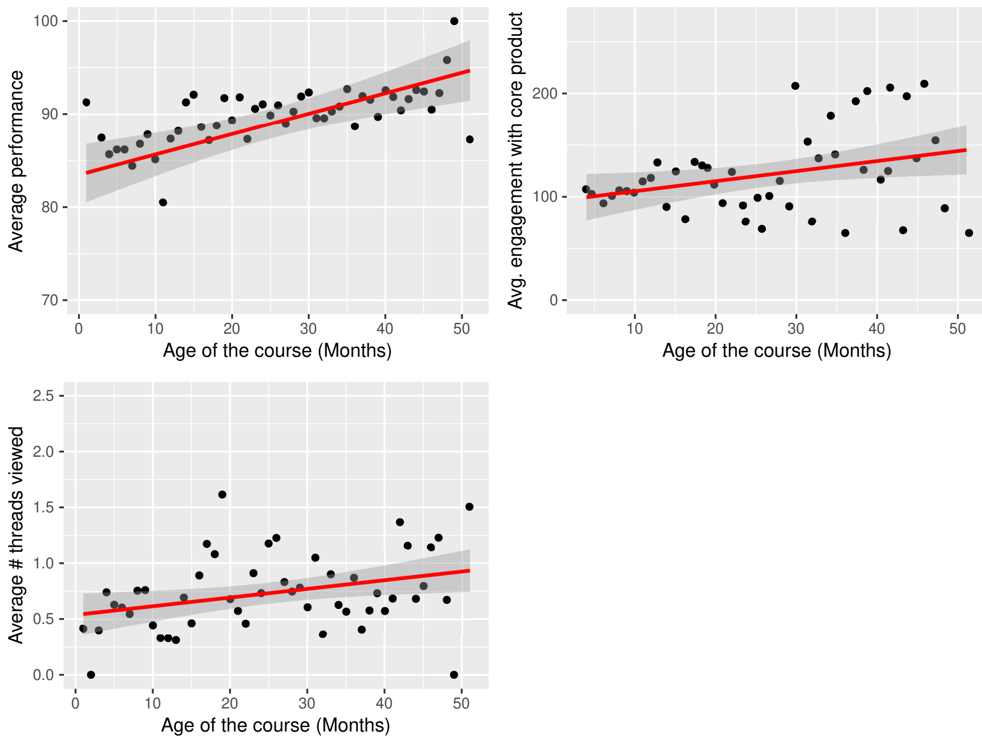


Figure 4: Student behavior and outcomes within one course evolve over time

Model-Free Evidence

As noted in the introduction, our objective is to assess if engagement with UGC (i.e., the discussion forum) affects engagement with the main product (i.e., the course material). As a first-cut analysis, we investigate if there is a difference between users who engage with the discussion forum vs. those who do not in terms of their engagement with the course material. To this end, we first consider these outcomes at the course level. In Figures 5, we compare the engagement and performance for two groups of users: those who visited the discussion forum at least once during the course vs. those who did not. These data suggest that, on average, those who visited the discussion forum spent more time with the course material and also performed better in terms of learning outcomes.

We extend this analysis by considering engagement with the discussion forum at the module level. This will help us assess the contemporaneous relationships between forum

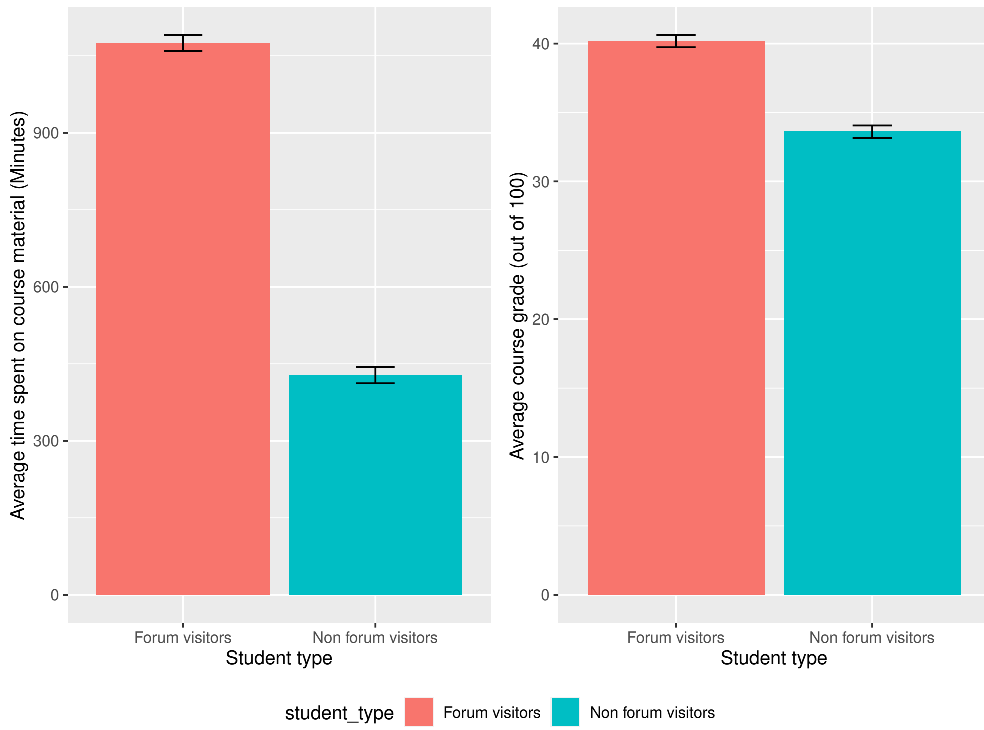


Figure 5: Relationship between engagement with the discussion forum and time spent on the material and performance at the course level.

engagement and engagement with the course material and performance. Furthermore, we distinguish between three levels of engagement with the discussion forum pertaining to a course module: (a) did not visit the discussion forum, (b) visited the discussion forum, but did not view any of the threads, and (c) visited the discussion forum and viewed at least one thread. We present the results from this analysis in Figure 6. Once again, we observe that those who actively engaged in the discussion forum corresponding to a module (by clicking on individual threads in the forum) were significantly more engaged with the course material than the other two groups of users. Interestingly, users who visited the discussion forum, but did not engage with the content were also more engaged than their peers who did not visit the discussion forum. We find a similar pattern when we consider performance within the module, although the differences are muted.

While it is tempting to interpret these patterns as evidence of the causal effect of engagement with the discussion forum on engagement with the course material and performance,

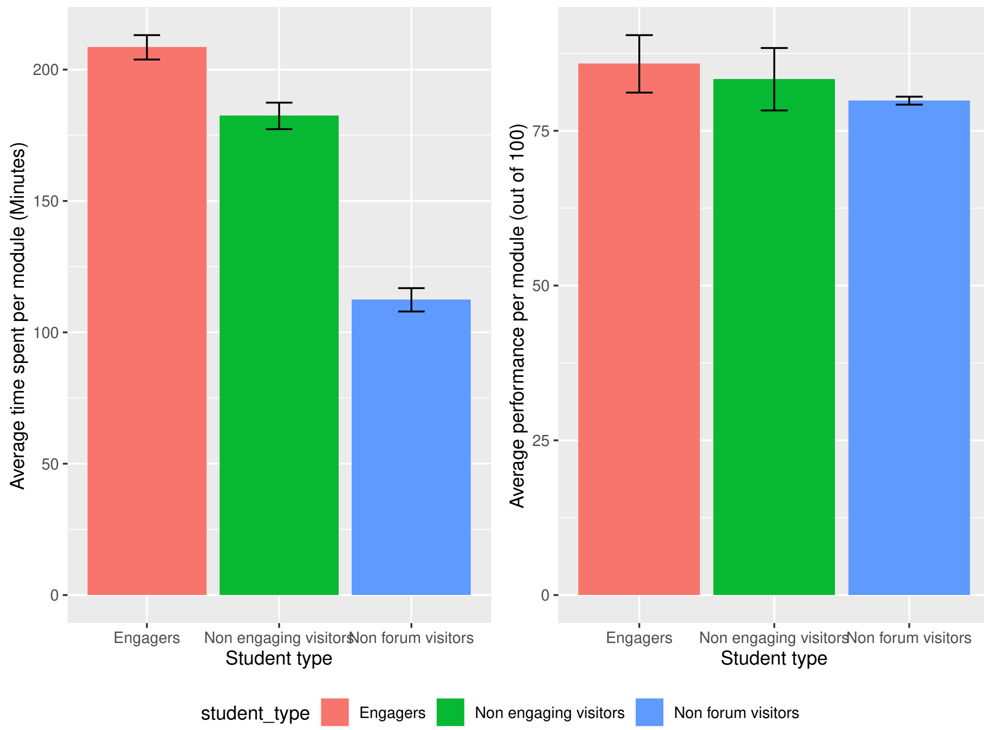


Figure 6: Relationship between engagement with the discussion forum and time spend on the material and performance at the course level.

we need to be careful about drawing such a conclusion for several reasons. First, recall that the patterns in Figure 2a presented earlier revealed that while some users actively engaged with the discussion forum, others were largely dormant. Such heterogeneity might arise due to differences among students in terms of their level of motivation as well as their ability to master the content. If users who were more active in the discussion forum were also systematically more engaged with the course material, then the observed correlation between these modes of engagement is probably attributable to these intrinsic differences, rather than to a causal relationship between them.

Second, in addition to heterogeneity across individuals, we also documented systematic differences across courses as well as modules within a course. For example, courses (as well as different modules within a course) may differ in terms of the level of engagement that they require from their users or in terms of the extent to which they foster engagement. These differences can result in users spending more time on both types of content in some courses (or in some modules within a course) and less in others. As we discussed earlier, the information in Figure 3 provided some evidence of such heterogeneity across modules within a course, probably because of the nature of the content. We need to control for these differences before isolating a causal relationship.

Third, as we saw in Figure 4, the overall proficiency of the users with respect to the educational content as well as in navigating the various learning objects might have evolved over time. Similarly, the courses as well as the modules within these courses may have evolved over time in a way that would have influenced the benefit from joint consumption of the two types of content. These require us to control for the evolution of each course and its modules over time. Finally, even if we can control for these correlated unobservables that drive engagement with both types of content, it might be difficult to isolate the direction of causality, i.e., did engagement with UGC drive engagement with the course material (i.e., the main product) or vice versa? In the next section, we will discuss our approach to addressing these challenges with a combination of a rich set of fixed effects and an instrumental variable approach to isolate the causal relationship between user engagement in the discussion forum and the course material.

EMPIRICAL ANALYSIS

In this section, we first present a simple regression model to illustrate the relationship between UGC and course material consumption more formally. Subsequently, we present instrumental variables that rely on exogenous variation in the forum rankings that will help us establish a causal link between the consumption of UGC and the course material.

Model Specification

Our goal is to understand if engagement with UGC, i.e., discussion forums, could increase consumption of course material and also improve students' performance. We consider the following regression model:

$$sinh^{-1}(\mathbf{Y}_{icm}) = \alpha \cdot \mathbf{V}_{icm} + \beta \cdot sinh^{-1}(\mathbf{E}_{icm}) + \eta_{cmw_{(i)}} + \eta_i + \epsilon_{icm}. \tag{1}$$

where i, c, w, and m index individual, course, week, and module, respectively. $\mathbf{Y}_{icm}$ is an outcome of interest, that is either time spent on module m of course c by individual i, or individual's i performance within module m of course c. $\mathbf{V}_{icm}$ is a dummy that indicates if individual i visited module m's forum from course c. $\mathbf{E}_{icm}$ is the number of threads user i clicked on during module m of course c. Finally, η_i and $\eta_{cmw_{(i)}}$ are user and course-module-week fixed effects. These fixed effects absorb the persistent differences across different users, and also non-parametrically control for the evolution of content available on the course forums as the material ages.

The parameter of interest (β) aims at measuring the marginal effect of $\mathbf{E}$ on $\mathbf{Y}$, that is the percentage change in course material consumption as a result of a percent shift in engagement with UGC. This is typically achieved using a log-log specification. However, since our outcome ($\mathbf{Y}_{icm}$) and independent variable $\mathbf{E}_{icm}$ do obtain zero as a value, we use the inverse hyperbolic sine (IHS) function:

$$sinh^{-1}(y) = \log\left(y + \sqrt{y^2+1}\right).$$

The IHS transformation alleviates the influence of outliers and the zero-inflation concern that is associated with adding an arbitrary value to the dependent variable while conducting inference, see Bellemare and Wichman (2020), Burbidge et al. (1988), Jayachandran et al. (2017), and Johnson (1949), for more discussion on IHS and Bahar and Rapoport (2018), Clemens and Tiongson (2017), and McKenzie (2017) for a few of its applications.[4]

Our objective is to use the variation in the data to measure the causal effect of UGC consumption on the consumption of the core product, which is measured by β in model (1).

[4]In the online appendix B we explain how to convert the coefficients β to elasticities.

Note that our data is at the user-module-course level. There are a number of different factors that could create spurious correlations between UGC consumption and the consumption of course material in each module. Below we list these issues along with our proposed solution:

- **Unobserved user heterogeneity:** Students could vary across different dimensions such as their intrinsic motivation or ability. These cross-sectional differences could affect their engagement with both UGC and course content. We use user-fixed effects to absorb the effect of these confounds.

- **Unobserved course-module heterogeneity:** Different modules within a course may require different levels of time investment and engagement. These factors could be correlated with both engagement with UGC and time spent watching lectures. Furthermore, course forums are shared across student cohorts. This means that as time passes by more content is added to the forums that could affect user engagement with course forums. We employ course-module-week fixed effects to control for confounds such as module difficulty or time-varying content of the forums.

- **Reverse causality:** Exposure to lectures could induce students to then engage with course forums to resolve questions or ambiguities that arise during the lectures. This would mean that OLS estimates from specification 1 could be biased due to the spurious correlation that is caused by the effect of content consumption on engagement with UGC. Note that including user and course-module-week fixed effects does not solve this issue. To address this endogeneity issue we use instrumental variables which we introduce in the next section.

We present the OLS estimates for specification (1) in Table 2. The effect of user engagement with the forum on time spent on lectures, see columns (1)-(3), shrinks as we include more granular fixed effects, The effect on time spent attenuates from 0.243 to 0.042, which means that unobserved user heterogeneity and course-module differences do indeed bias the estimates and including these fixed effects absorbs the effect of confounds that vary at user, and course-module-week level. The effect of UGC consumption on performance, however, remains similar as we include finer fixed effects, see columns (4)-(6). Nevertheless, these estimates may still be biased because of the reverse causality issues discussed above. In the next section, we employ an instrumental variable approach to address this issue.

Table 2: The effect of user engagement with course forums (OLS estimates)

	Dependent variable					
	$\sinh^{-1}$(Time spent on course material)			$\sinh^{-1}$(Performance)		
	(1)	(2)	(3)	(4)	(5)	(6)
Engagement with forum (β)	0.243***	0.053**	0.042**	0.011**	−0.002	0.012**
	(0.018)	(0.022)	(0.019)	(0.005)	(0.006)	(0.005)
Visited forum (α)	1.456***	0.668***	0.483***	0.009	0.044***	0.030***
	(0.035)	(0.035)	(0.030)	(0.009)	(0.010)	(0.008)
Constant	3.788***			5.085***		
	(0.017)			(0.003)		
User FE		X	X		X	X
Course-Module-week FE			X			X
Observations	115,036	115,036	115,036	115,036	115,036	115,036
R^2	0.037	0.671	0.744	0.0001	0.581	0.647
Adjusted R^2	0.037	0.569	0.664	0.0001	0.451	0.537

Note: *p<0.1; **p<0.05; ***p<0.01

All standard errors are clustered at user level.

Instrumental variable regressions

While cross-sectional differences across users and modules can be absorbed using high dimensional fixed effects, the reverse causality issue discussed above requires us to use exogenous variation that affects forum usage but does not directly impact course material consumption. As discussed in the "Drivers of User Engagement in Online Courses" section, there are two approaches to increase user engagement with course forums: (a) nudging users to engage with forums, and (b) improving the attractiveness of the content within the forum. Previous research by Narang et al. (2022) and Zhang et al. (2017) have relied on experimental variation to demonstrate (a), however, our approach differs from these papers as we study the

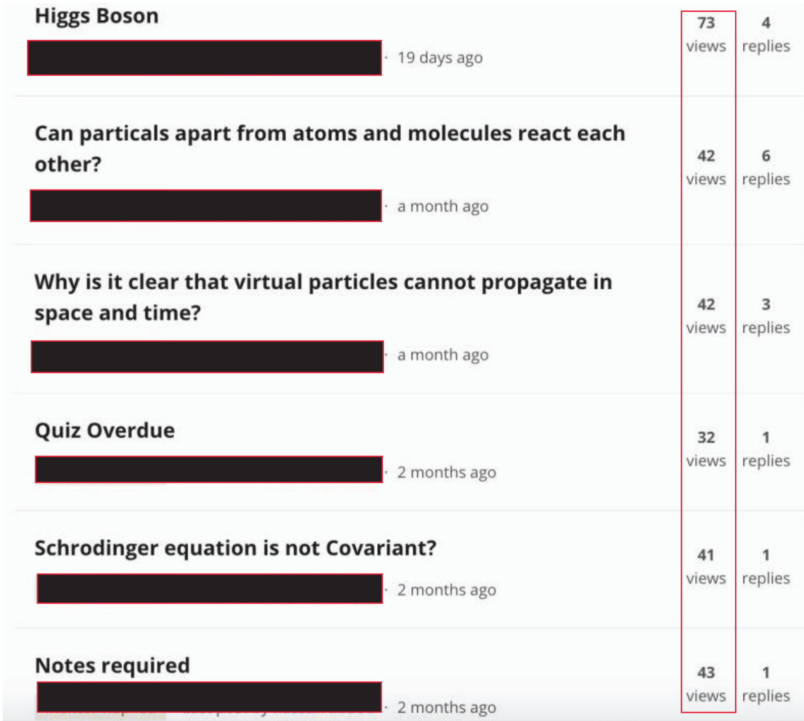


Figure 7: Forum Home Page for a Module

effect of exogenous variation in the intrinsic attractiveness *within* the discussion forum once the participant lands on the forum page. In particular, we employ instrumental variables that can shift user engagement with the forum threads by shifting forums' attractiveness.

Definition of the instrumental variables

Each module in a given course has a separate discussion forum where students can discuss the course material covered in that particular module. Once students visit a discussion forum within a module, they are greeted by its home page which consists of a set of discussion threads. The threads on the homepage are sorted by activity recency, i.e., the thread with the most recent activity (post or comment) bubbles up to the top of the list. The discussion threads list is paginated, with at most fifteen threads per page.

Our goal is to construct metrics that capture the attractiveness of the forum homepage of each module when a student visits it. As mentioned above, the rank ordering of threads in the forum changes as a new thread is created or new posts/comments are added to an older thread. Hence, users who arrive at the same forum homepage for a module, albeit within a few minutes of each other, could face a different set of threads as the forum evolves. We conjecture that the probability that a user engages with the UGC (i.e., clicks on a thread within the discussion forum) would depend on the state of the forum at that time. We define three measures to capture the state of the forum home page at the time of a user's first forum visit, namely, popularity, relevance, and variety. Of these, popularity is a metric of prior user engagement with the threads on the visible part of the discussion forum at the time of the user's visit. As we discuss below, we determine this directly from the numeric information that is readily available both to the user and the researcher. For the remaining two factors, relevance and variety, we use text-based methods to infer them from the threads on the visible part of the discussion forum at the time of a user's visit. Our premise is that upon visiting the homepage of the discussion forum, the user is also exposed to the titles of threads listed on the homepage. To construct these text-based metrics, for each course, we first estimate a BERTopic model (Grootendorst, 2022) on a corpus comprising of the thread titles residing in the forums of that course and then utilize the output of the BERTopic model. We discuss the details regarding the estimation of the BERTopic model in appendix C. All metrics are defined only for users that visit a module's forum and are intended to capture the state of the forum home page at the time of their first forum visit. [5] We discuss the construction of these measures in detail below.

Popularity

Experimental studies have documented the positive impact of popularity cues on individuals' online shopping choices in contexts such as software and music downloads (Hanson and Putler, 1996, Salganik and Watts, 2008). More recently, studies have also documented that

[5]Our objective is to measure the impact of the state of the discussion forum for a module on a user's engagement with the forum and the course content. 89.32% of forum visitors in our sample visit the discussion forum for a module only on a unique date. If a user visits the discussion forum for a module on multiple occasions, we can define the state of the discussion forum in several possible ways. We explore the robustness of our results by restricting our analysis to the instances where users visited the forum only on a unique date.

such cues increase engagement/sharing on social media (Epstein et al., 2022). In our context, we define popularity as the total views of the threads that were present on the forum home page at the time of the user's visit. To a user who arrives at a forum at any given time, this metric would act as a signal of the benefits from engagement.

Let $\mathcal{F}_{icm}$ be the set of threads that appear on the first page of the forum m when user i from course c first visits it. Also, let $v_{icm}^{(t)}$ be the number of views for thread t at the time where user i visits the forum's homepage for module m in course c, e.g., see Figure 7. We define our first metric as follows:

$$Popularity_{icm} = \frac{1}{1000} \sum_{t \in F_{icm}} v_{icm}^{(t)}. \tag{2}$$

Relevance

We define relevance as the appeal of topics spanned by the threads present on the forum home page at the time of a user's first visit. As Chernev et al., 2012 notes, "... some assortments comprise options that are, on average, of higher quality and, hence, are likely to be perceived as more attractive (e.g., Nordstrom, Neiman Marcus, and Whole Foods). In contrast, other assortments comprise options that are, on average, of lower quality and are likely to be perceived as relatively less attractive (e.g., dollar stores, Value City, and K-Mart). In addition, some assortments can be perceived as more attractive because the items they carry match customer preferences". In a similar vein, given the assortment of threads that the user encounters on their first visit, we characterize appeal based on the level of engagement exhibited by *all users* on the forum for the set of topics spanned by these threads.

Let $\mathcal{T}_c$ be the set of threads in the forums of course c. Let $\mathcal{S}_c$ be the set of topics in the forums of course c. $\forall j \in \mathcal{S}_c$, we first compute a topic quality score by combining two pieces of information as follows:

- Let $v^{(t)}$ be the number of views accumulated by thread t at the end of the entire observation period.
- For each thread t, we obtain a vector of probabilities from the topic model that characterizes the probability of a thread title belonging to a topic: $\mathbf{p_t} = \left[p_1^t, p_2^t, \ldots, p_{|S_c|}^t\right]$.

We then define the quality score for topic j as:

$$TopicQualityScore_j = \sum_{t \in T_c} p_j^t * v^{(t)}. \tag{3}$$

Now, let $\mathcal{F}_{icm}$ be the set of threads that appear on the first page of the forum m when user i from course c first visits it. Let $p_{icm} = \left[p_{icm}^1, p_{icm}^2, \ldots, p_{icm}^{|S_c|}\right]$ characterize the distribution of topics on the home page at the time of user i's first forum visit to module m's forum in course c where $p_{icm}^j = \frac{\sum_{t \in F_{icm}} p_j^t}{|F_{icm}|}$.

We compute relevance by combining the information regarding the distribution of the topics on the homepage with the information regarding the quality of those topics as follows:

$$Relevance_{icm} = \frac{1}{1000} \sum_{j \in S_c} (p_{icm}^j * TopicQualityScore_j). \tag{4}$$

Variety

We define variety as the diversity of topics spanned by the threads present on the forum homepage at the time of the user's visit. As a metric of the topic diversity, we use entropy (Holtz et al. (2020), Chen et al., 2023). Let $p_{icm} = \left[p_{icm}^1, p_{icm}^2, \ldots, p_{icm}^{|S_c|}\right]$ characterize the distribution of topics on the home page at the time of user i's first forum visit to module m's forum in course c where $p_{icm}^j = \frac{\sum_{t \in F_{icm}} p_j^t}{|F_{icm}|}$. We compute Shannon entropy as follows:

$$Variety_{icm} = -\sum_{j \in S_c} p_{icm}^j log(p_{icm}^j). \tag{5}$$

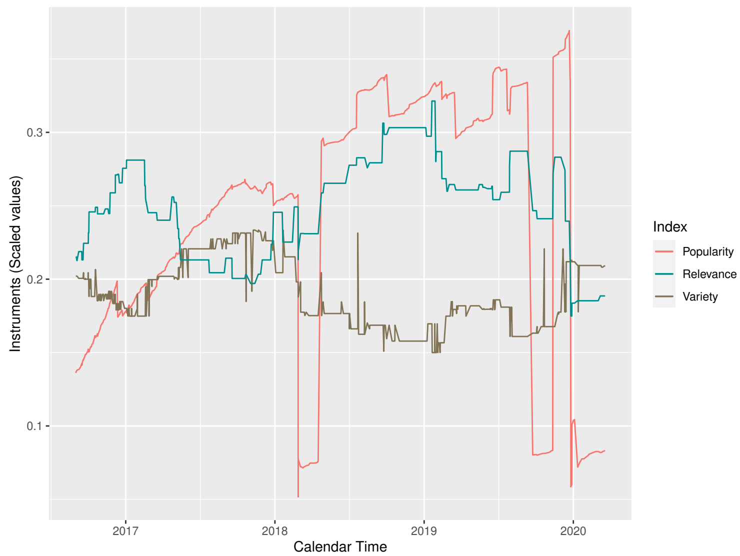
All else equal, the entropy metric is a function of the total number of topics (S_c) as well as the distribution of these topics among the threads. Hence, our variety metric considers both the richness and evenness of topic distribution on the homepage. Finally, the correlation between popularity and relevance is 0.28, popularity and variety is 0.39, and relevance and variety is -0.15 which suggests that the three measures capture different properties of UGC encountered on the forum home page.

Exogeneous Variation in the instruments

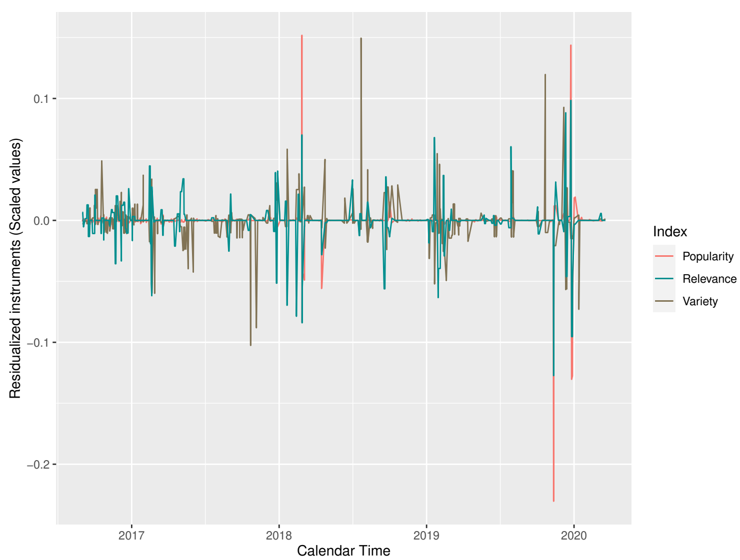
To understand how the instruments evolve, we present the time series of these measures for one module within a course in our panel in Figure 8. We observe sharp discontinuous drops/jumps in the time series plot. Sharp drops (jumps) in instrumental variables, see Figure 8. These drops and jumps in instruments are due to changes in the composition of threads presented on the homepage and are due to a thread creation or activity in one of the threads that was not on the homepage. This variation in the composition of threads on the homepage and the resulting changes in instruments is the main source of exogenous variation used in our study.

Our goal is to use these three measures as instruments that shift a forum's attractiveness and thereby affect user engagement with the forums. There are three concerns that one might have about using these metrics as instruments that shift a forum's attractiveness:

- **Instruments signal quality of course material:** One might be concerned that forum attractiveness might signal the quality of the course material which could induce users to directly increase their engagement with the course material. We rule this mechanism out in the online appendix D. The idea is to use the variation in instruments across users who visit a forum's homepage but do not click on any threads. If instruments are indeed signaling the quality of the course material, users who don't engage with the threads must be treated as well and a relationship between forum attractiveness and user engagement with course content must be observed which is not



(a) Evolution of the potential instrumental variables



(b) Residual of the potential instrumental variables

Figure 8: The top panel shows the evolution of the instrumental variables for a discussion forum. The discontinuous jumps in these variables are driven by changes in the composition of threads on the first page. The bottom panel displays the residualized instruments and demonstrates that general time trends are absorbed by our time-varying fixed effects.

the case in our data.

- **Trends and spurious correlations:** While the discontinuous jumps/drops in the instruments present a source of exogenous variation, there could be a time trend as the course ages. This trend could be correlated with other confounds that affect students' engagement and performance. For instance, the students who enroll later on could be systematically different from those who enrolled in the course when it was relatively new on the platform. Thus, the age of the course on the platform has the potential to confound our empirical analysis. To address this issue we use two sets of high-dimensional fixed effects: (a) user, and (b) course-module-week fixed effects. The user fixed effects address the user selection issue and ensure that we are relying on within-user variation. The time-varying course-module-week fixed effects subsume seasonalities or other time-varying factors and ensure that we rely on within-week variation when investigating the effect of the instrumental variable on user engagement with forums. Figure 8b shows that the fixed effects were effective in absorbing general time trends in our instruments (see Figure 8a) and the identification would rely on the changes in detrended instruments in Figure 8b.

- **Sequential interactions and aggregation:** Note that our data is aggregated at the user-course-module level, meanwhile the instruments are calculated based on the first visit of each individual to a forum's homepage. Therefore, instruments are essentially shifting the engagement level for the first visit within each module. Since visiting a forum page could prompt users to consume course content, and consuming course content can, in turn, induce users to return to forums to seek help regarding course content, aggregating forum engagement and course material consumption to module level could violate the exclusion restrictions and bias our estimates, see Figure 9 for illustration. To allay these concerns, in the online appendix E, we repeat the analysis for the set of observations with a single visit to the forum's homepage and find that

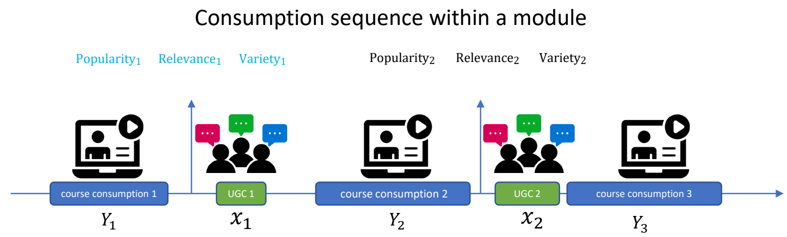


Figure 9: Sequential interactions

our results remain robust.

In the subsequent sections, we present the results from the first-stage regressions and then discuss the estimates from the second-stage regressions.

First-stage regressions

In this section, we present the first-stage regression specification and discuss the results. We consider the following specification:

$$sinh^{-1}(\mathbf{E}_{icm}) = \tilde{\alpha} \cdot \mathbf{V}_{icm} + \tilde{\beta} \cdot \mathbf{V}_{icm} \cdot \mathcal{I}_{icm} + \eta_{cmw_{(i)}} + \eta_i + \epsilon_{icm}. \tag{6}$$

where i, c, and m index users, courses, and modules, respectively. $\mathcal{I}_{icm}$ are the instrumental variables that are defined for users where $\mathbf{V}_{icm} = 1$ and they capture the state of the forum when the user first visits a forum's homepage. The rest of the parameters and fixed effects are defined similar to specification (1). Moving from left to right in Table 3 we progressively add fixed effects and report the estimates from (6). The results in Table 3 suggest that while popularity and relevance shift users' engagement with forums, we do not find a significant effect of variety on engagement with forums.[6]

[6] In specification (3), this result translates into an elasticity estimate that suggests that on average, a 10% increase in popularity and a 10% increase in relevance increases the number of threads viewed by 15.19% and 3.43% respectively. See appendix B for a discussion on recovering elasticity estimates from IHS-IHS specifications.

Table 3: The effect of potential instrumental variables on engagement with threads on the course forums

	Dependent variable:		
	sinh^{-1}(Engagement with forum)		
	(1)	(2)	(3)
Forum_Visited	0.823***	0.914***	0.794***
	(0.086)	(0.088)	(0.078)
Forum_Visited x sinh^{-1}(Popularity)	0.010	0.082***	0.074***
	(0.018)	(0.020)	(0.018)
Forum_Visited x sinh^{-1}(Relevance)	0.143***	0.054***	0.062***
	(0.019)	(0.020)	(0.019)
Forum_Visited x sinh^{-1}(Variety)	−0.001	−0.047	0.006
	(0.053)	(0.054)	(0.048)
User FE		X	X
Course-Module-Week FE			X
Observations	115,036	115,036	115,036
R^2	0.529	0.723	0.765
Adjusted R^2	0.529	0.639	0.647

Note: *p<0.1; **p<0.05; ***p<0.01
All standard errors are clustered at user level.

Second-stage regressions

We now re-estimate (1) using popularity and relevance as instrumental variables for engagement with the forum $sinh^{-1}(\mathbf{E}_{icm})$. The estimates from this analysis are displayed in Table 4, which are the counterparts to the OLS estimates reported in Table 2. Moving from left to right, we progressively add more granular fixed effects. Our preferred specifications for the effect on time spent on course material and performance are models (3) and (6), respectively — both of these include fixed effects by user and course-module-week. Using the procedure in appendix B to convert these estimates to elasticity, we find that a 10% percent increase in engagement level with forum content increases time spent on course material and performance by 4.32% and 0.60% percent, respectively. The IV estimates for the effects of engagement on course content consumption and performance are statistically different and significantly larger than the OLS estimates documented in Table 2, which suggests that high-dimensional fixed effects in (1) were not adequate for addressing the endogeneity issues discussed above. To assess whether popularity and relevance suffer from the problems of being weak instruments, we checked the F-statistic on these excluded instruments (See 4). The first-stage F-test suggests that the instruments have power and do not suffer from the problems of being weak instruments.

Table 4: The effect of user engagement with course forums (IV estimates)

	Dependent variable:					
	$\sinh^{-1}$(Time spent on course material)			$\sinh^{-1}$(Performance)		
	(1)	(2)	(3)	(4)	(5)	(6)
Engagement with forum (β)	4.866***	0.694***	0.502***	4.565***	0.096***	0.070***
	(0.053)	(0.025)	(0.021)	(0.048)	(0.010)	(0.009)
Visited forum (α)	0.510**	−0.808**	0.715***	−0.033	−0.028	0.307***
	(0.207)	(0.322)	(0.224)	(0.070)	(0.106)	(0.105)
Constant	4.068***			4.989***		
	(0.012)			(0.004)		
User FE		X	X		X	X
Course-Module-Week FE			X			X
Observations	115,036	115,036	115,036	115,036	115,036	115,036
R^2	0.021	0.653	0.772	0.00001	0.561	0.649
Adjusted R^2	0.021	0.423	0.532	−0.00001	0.270	0.279
F-statistic on the excluded instruments	64.446	31.079	34.992	64.446	31.079	34.992

Note: *p<0.1; **p<0.05; ***p<0.01

All standard errors are clustered at user level.

Identification Strategy

Our identification strategy hinges on two premises:

- The *precise* time a user lands on a given module's forum homepage is "as good as random."
- The state of a given module's forum, fluctuates by virtue of the forum sorting rule

and this sorting rule generates exogenous variation in the set of topics a user faces, see Figure 8.

Using a similar identification strategy, Donnelly et al. (2021) and Levin et al. (2017) have leveraged randomness in customers' arrival time in a supermarket that changes prices on Tuesdays and high-frequency shifts in prices of gasoline to identify price elasticities in each context. Similarly, Vana and Lambrecht (2021) have leveraged randomness in the arrival of online reviews to identify the effect of one additional review on purchase likelihood on e-commerce platforms. In our case, the composition of threads available on a forum's homepage could change discontinuously as new threads bubble up or old ones get updated by the users. This means that the set of threads available on the homepage of a module's forum for two users who arrive within a short time period of one another could be different.

To capture the exogenous variation in the composition of threads on the forum homepage we define popularity, relevance, and variety. In essence, these metrics gauge the properties of threads on the forum's homepage at any point in time. As demonstrated in Figure 8, these metrics could change dramatically as the composition of the threads changes due to the platform's sorting rule. We argue that conditional on the fixed effects used in our analyses popularity and relevance qualify as valid instruments for user engagement. First, note that detrended measures (Figure 8b contain exogenous variation that is caused due to the changes in the set of threads on the homepage as threads enter/exit the page. To qualify as a valid instrument, these measures must satisfy the following conditions: i) relevance: a given student's first impression of the state of the forum must be correlated with that student's forum usage, and ii) exclusion: The state of the forum must affect the focal student's engagement with course material and the focal student's performance only through its effect on the focal student's UGC engagement. Our results in Table 3 show that popularity and relevance are indeed relevant and as expected higher popularity and higher relevance led to higher engagement with the forum.

ROBUSTNESS CHECKS

In this section, we present more evidence supporting the validity of the instruments and our identification strategy. We first demonstrate that our results could not be a mere effect of attrition, and then we investigate different aspects of the instruments used in our analyses.

Selection bias due to attrition

Not every enrollee in the course completes the course. A lot of students drop out on MOOC platforms. Once a student drops out, we don't observe their behavior. Thus, the parameter estimates obtained from our econometric model fitted on the full sample will reflect both changes in the composition of the sample and changes in the behavior of students. The research objective of the paper is to document findings based on changes in behavior rather than findings based on the changes in the composition of the sample.

We assess the sensitivity of our results to changes in the composition of the sample by estimating our model on subsamples stratified by the course completion rate. Specifically, we fit our model on the sample of students who completed at least 10% of the course, fit the model on the sample of students who completed at least 20% of the course, and so on and so forth. The idea is that if the parameter estimates of our model are reflecting changes in behavior, then the results should persist across these subsamples. Tables 5 and 6 report the results from estimating the model on different subsamples stratified by the course completion rate. The results persist across the different subsamples, which suggests that parameter estimates reflect changes in behavior rather than changes in the composition of the sample.[7]

[7] Each module consists of course items such as lecture videos, supplementary readings, etc. The completion rate is computed using course progress data as the percentage of course items completed by the student.

Table 5: Estimation of second stage equation on subsamples stratified by the course completion rate for time spent on course material

	Dependent variable:								
	$\sinh^{-1}$(Time spent on course material)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Engagement with forum (β)	0.502***	0.477***	0.445***	0.430***	0.409***	0.406***	0.392***	0.378***	0.347***
	(0.021)	(0.021)	(0.021)	(0.022)	(0.023)	(0.025)	(0.026)	(0.027)	(0.028)
Visited forum (α)	0.715***	0.775***	0.836***	0.632**	0.632**	0.484	0.501*	0.523*	0.397
	(0.224)	(0.238)	(0.272)	(0.268)	(0.271)	(0.299)	(0.300)	(0.314)	(0.343)
Observations	115,036	107,438	90,921	80,306	72,945	67,292	61,613	54,853	45,690
R^2	0.772	0.754	0.698	0.684	0.680	0.685	0.691	0.704	0.725
Adjusted R^2	0.532	0.527	0.477	0.479	0.486	0.498	0.507	0.524	0.543

Note: *p<0.1; **p<0.05; ***p<0.01

All regressions include User and Course-Module-Week fixed effects. All standard errors are clustered at user level.

Table 6: Estimation of second stage equation on subsamples stratified by the course completion rate for performance

	Dependent variable:								
	$\sinh^{-1}$(Performance)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Engagement with forum (β)	0.070***	0.069***	0.056***	0.034***	0.027***	0.023***	0.012	0.008	0.001
	(0.009)	(0.009)	(0.009)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.007)
Visited forum (α)	0.307***	0.291***	0.213*	0.338***	0.275***	0.095	0.056	0.035	0.089
	(0.105)	(0.109)	(0.111)	(0.102)	(0.085)	(0.074)	(0.065)	(0.065)	(0.069)
Observations	115,036	107,438	90,921	80,306	72,945	67,292	61,613	54,853	45,690
R^2	0.649	0.629	0.616	0.614	0.626	0.637	0.639	0.660	0.669
Adjusted R^2	0.279	0.287	0.335	0.362	0.398	0.422	0.425	0.452	0.450

Note: *p<0.1; **p<0.05; ***p<0.01

All regressions include User and Course-Module-Week fixed effects. All standard errors are clustered at user level.

Validity of the exclusion restriction

The exclusion restriction requires that the state of the forum at the time of the first forum visit must affect a forum visitor's engagement with course material and performance only through its influence on the forum visitor's engagement with UGC. There should not be a direct effect of the state of forum on the dependent variables. We conduct some additional analysis to assess the validity of the exclusion restriction in our setting.

We conjecture that two behavioral mechanisms can drive the effect of the state of the forum on engagement with UGC. First, observing an attractive state of the forum can improve the focal user's perception of the quality of UGC available in the forum. This improvement in the perceived quality of UGC can make the focal student engage more with the UGC by viewing more discussion threads. Second, observing an attractive state of the forum might act as a signal of the quality of the course material. This, in turn, can directly impact the user's engagement with the course material. This interpretation, wherein the state of the discussion forum homepage at the time of the visit can directly influence engagement with the course material would violate the exclusion restriction.

In our context, we observe three types of users in terms of the way they engage with the discussion forum for each module (see Figure 10): (a) those who do not visit the discussion forum, (b) those who visit the discussion forum, but do not click on any of the threads, and (c) those who visit the forum and click on one or more of the threads. If the signaling argument holds, merely visiting the discussion forum homepage and observing the state therein would influence engagement with the course material. In that case, we should observe a relationship between the state of the forum and the engagement with the course material if we consider the subsample of users in (b).

In online appendix D, we report the results from the estimation of our model for the subsample of users who visited the forum but didn't engage with any UGC by clicking on threads. These results suggests that there is no systematic relationship between the

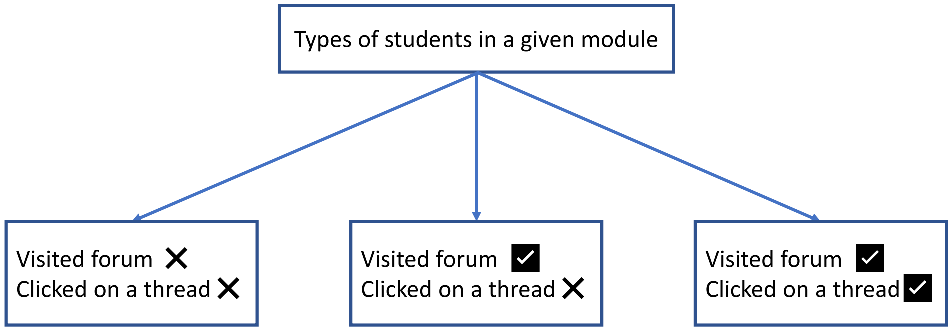


Figure 10: Types of students within a course module

instrumental variables and the dependent variables for this subsample. This provides us greater confidence regarding the validity of the exclusion restriction.

DISCUSSION

We have investigated how engagement with UGC affects student behavior and outcomes on a popular online education platform. In this section, we provide the implications of our findings, highlight limitations, and provide directions for future research. Our finding that the discussion forum and the course material tend to be viewed as complements by users has some potential implications for online education platforms. As online education platforms are looking for ways to address the low engagement and course completion problems, discussion forums can be a potential solution. In particular, increasing user engagement with UGC can be a useful lever that these platforms can use for increasing engagement with the course material. Therefore, we believe that the debate around the value of maintaining discussion forums for online courses needs to consider the evidence from our research. However, we acknowledge that our data do not include instances where there is entry/exit of UGC. Therefore, while our study can provide insights about the effect of engagement with discussion forums on engagement with the course material, it cannot directly comment on the value of having these forums per se.

Our study also sheds light on how online education platforms can increase user engagement with the discussion forum. While prior research (e.g., Narang et al. (2022) and Zhang et al. (2017)) has considered nudging users to visit the forum as a potential lever, we document that what users see when they visit the discussion forum is an important driver of their subsequent engagement. In particular, our results suggest that a good first impression in terms of the popularity and relevance of the available UGC increases engagement with UGC. Figure 11 illustrates this relationship. Therefore, our results have potential implications for the design of discussion forums. Since our results suggest that the popularity and relevance of topics that are represented on the homepage tend to increase engagement, optimizing the default homepage along these dimensions appears to hold promise. However, given that our data do not include such a scenario, it is not clear how users would respond to such a structural change in the way information is presented. It would be worthwhile for online education platforms to conduct experiments with alternative approaches for designing the default presentation of information on the forum homepage.

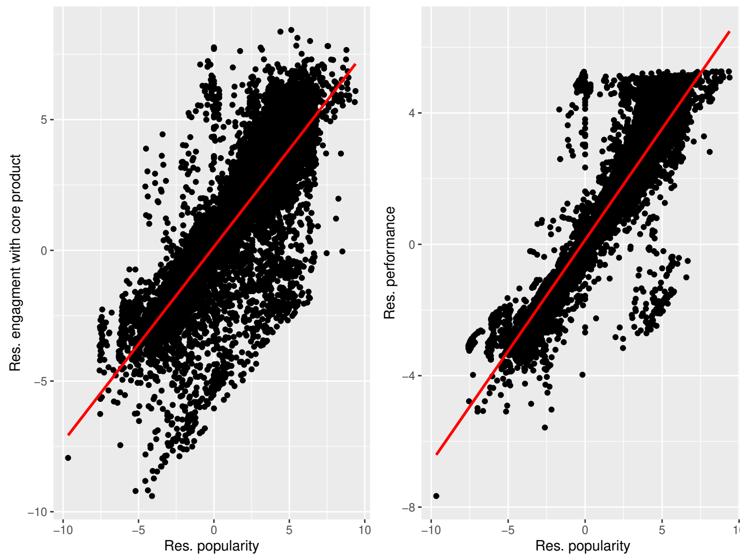
The findings from our research can potentially shed light on the relationship between engagement with UGC and the core product in a broader set of contexts beyond online education. Examples of such platforms where the core product coexists with UGC include social news aggregation platforms such as Reddit, fitness platforms such as Bodybuilding.com, MyFitnessPal, etc., gaming platforms such as Zapak, Gaia, etc. Extrapolating our findings to these contexts would suggest that maintaining a vibrant UGC forum and encouraging engagement therein might serve as a potential tool for increasing user engagement with the core product. However, we need to exercise caution in extrapolating to contexts beyond online education for two reasons. First, while discussion forums in online courses may help users navigate the course material (and hence be the reason behind the complementarity documented here), we cannot assume that such an association would exist in other contexts. Rather, discussion forums can be distractions as feared. Second, unlike the discussion forums in online courses, UGC in other contexts can be polarizing, and at times, toxic. In such instances, the beneficial effects of engagement with UGC can become questionable.

Nevertheless, while our results cannot comment on the exact nature of relationships in a broad array of contexts, they highlight the need for content-based platforms to investigate such relationships to inform approaches to improve user engagement.

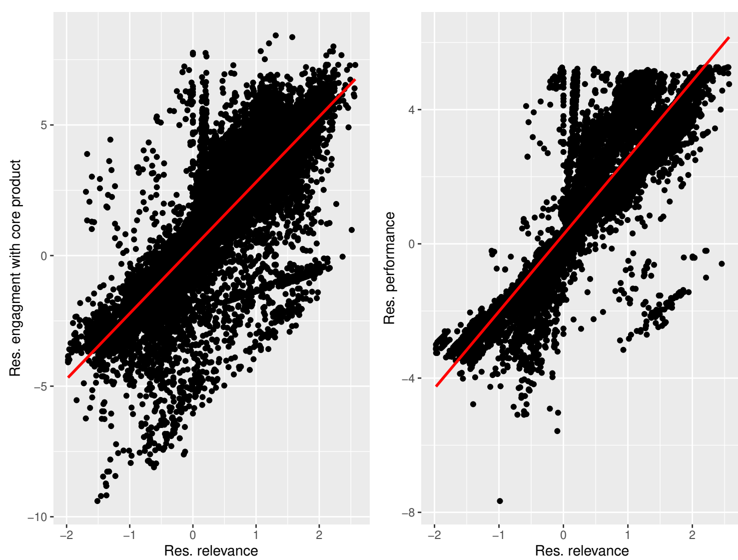
Our study suffers from a few limitations. First, as noted earlier, our study has abstracted away from potential substitutability or complementarity across modules and has just focused on contemporaneous engagement with UGC. A perfectly plausible hypothesis is that engagement with the previous module's UGC in the current module affects performance and engagement with course material in the current module. However, we focus on the consequences of engagement with the current module's UGC on behaviors and outcomes in the current module. Future research can investigate such potential substitutability or complementarity across different periods in the customer journey. Research along these lines can inform principles that should govern the optimal design of discussion forums.

Second, since we don't control for the endogeneity of the decision to visit the forum, we need to view the quantifications as suggestive. Nevertheless, the evidence that students who were induced to engage more with UGC due to exposure to a better forum (a forum with higher popularity and higher relevance presented upfront), also engaged more with the course material and performed better, reassures us that engagement with UGC positively affects both engagement with the core product and performance. This interpretation of our results is valid because a student's decision to visit a module's forum while learning that module is uncorrelated with the state of that module's forum at the time of that student's first forum visit. This is because every module has a separate forum and the state of a given module's forum is only known to the student after they have visited that module's forum.

Third, to leverage the panel nature of our dataset and use individual fixed effects to control for unobservable heterogeneity, we have to assume that the user-module match is fixed within a course. Finally, we are unable to comment on the value of the discussion forum itself as we only have data from a world where discussion forums exist. We do not have access to data on student behavior and outcomes from a world where discussion forums



(a) Effect of popularity on engagement with course material and performance



(b) Effect of relevance on engagement with course material and performance

Figure 11: Due to module-specific heterogeneity and potential confounding due to the age of the course, a scatter plot with raw data will not provide a clear picture regarding the relationship between instrumental variables and the dependent variables of interest. Thus, for cleaner exposition, we residualized the instruments and the outcome measures by the course-module-week fixed effects for the observations where a forum was visited to construct this scatterplot.

don't exist or where student access to the discussion forum was randomly manipulated. Such a dataset can be very useful in understanding the value of discussion forums and the boundary conditions of when having a discussion forum is desirable and when that isn't the case. This knowledge can be useful in informing online platforms' content design policies.

References

Aral, S. and Walker, D. (2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management science* 57 (9): 1623–1639.

Bahar, D. and Rapoport, H. (2018). Migration, knowledge diffusion and the comparative advantage of nations. *The Economic Journal* 128 (612): F273–F305.

Banerjee, A. and Duflo, E. (2014). Economics education in the digital age: The implications of online technologies and MOOCS. In: *American Economic Review: Papers and Proceedings.* Vol. 104. 5: 514–518.

Bellemare, M. F. and Wichman, C. J. (2020). Elasticities and the inverse hyperbolic sine transformation. *Oxford Bulletin of Economics and Statistics* 82 (1): 50–61.

Berman, R., Melumad, S., Humphrey, C., and Meyer, R. (2019). A tale of two Twitterspheres: Political microblogging during and after the 2016 primary and presidential debates. *Journal of Marketing Research* 56 (6): 895–917.

Burbidge, J. B., Magee, L., and Robb, A. L. (1988). Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association* 83 (401): 123–127.

Chen, G., Chan, T., Zhang, D., Liu, S., and Wu, Y. (2023). The Effects of Diversity in Algorithmic Recommendations on Digital Content Consumption: A Field Experiment. *Available at SSRN 4365121.*

Chernev, A. et al. (2012). Product assortment and consumer choice: An interdisciplinary review. *Foundations and Trends® in Marketing* 6 (1): 1–61.

Clemens, M. A. and Tiongson, E. R. (2017). Split decisions: Household finance when a policy discontinuity allocates overseas work. *Review of Economics and Statistics* 99 (3): 531–543.

Conrad, D. L. (2002). Engagement, excitement, anxiety, and fear: Learners' experiences of starting an online course. *The American journal of distance education* 16 (4): 205–226.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Di Gangi, P. M. and Wasko, M. (2009). The co-creation of value: Exploring user engagement in user-generated content websites. In: *Proceedings of JAIS theory development workshop. sprouts: working papers on information systems*. Vol. 9. 50: 9–50.

Donnelly, R., Ruiz, F. J., Blei, D., and Athey, S. (2021). Counterfactual inference for consumer choice across many product categories. *Quantitative Marketing and Economics* 19 (3): 369–407.

Duke, D. (2023). Why User-Generated Content Is Winning. `https://www.forbes.com/sites/forbesbusinesscouncil/2023/03/13/why-user-generated-content-is-winning/?sh=4d8aeb096e94`.

Ellis, J. (2014). What happened after 7 news sites got rid of online comments. `https://www.niemanlab.org/2015/09/what-happened-after-7-news-sites-got-rid-of-reader-comments/`.

Epstein, Z., Lin, H., Pennycook, G., and Rand, D. (2022). How many others have shared this? Experimentally investigating the effects of social cues on engagement, misinformation, and unpredictability on social media. *arXiv preprint arXiv:2207.07562*.

Goli, A., Chintagunta, P. K., and Sriram, S. (2022). Effects of payment on user engagement in online courses. *Journal of Marketing Research* 59 (1): 11–34.

Goujard, C. (2016). Why news websites are closing their comments section. `https://medium.com/global-editors-network/why-news-websites-are-closing-their-comments-sections-ea31139c469d`.

Greenslade, R. (2015). Is it really wise for news websites to stop people from commenting? `https://www.theguardian.com/media/greenslade/2015/sep/25/is-it-really-a-good-idea-to-turn-for-news-websites-to-turn-off-comments`.

Griffiths, T., Jordan, M., Tenenbaum, J., and Blei, D. (2003). Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems* 16.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Gu, Z., Bapna, R., Chan, J., and Gupta, A. (2022). Measuring the impact of crowdsourcing features on mobile app user engagement and retention: A randomized field experiment. *Management Science* 68 (2): 1297–1329.

Hamari, J., Koivisto, J., and Sarsa, H. (2014). Does gamification work?–a literature review of empirical studies on gamification. In: *2014 47th Hawaii international conference on system sciences*. Ieee: 3025–3034.

Hanson, W. A. and Putler, D. S. (1996). Hits and misses: Herd behavior and online product popularity. *Marketing letters* 7: 297–305.

Hartmann, J., Heitmann, M., Schamp, C., and Netzer, O. (2021). The power of brand selfies. *Journal of Marketing Research* 58 (6): 1159–1177.

Holtz, D., Carterette, B., Chandar, P., Nazari, Z., Cramer, H., and Aral, S. (2020). The engagement-diversity connection: Evidence from a field experiment on spotify. In: *Proceedings of the 21st ACM Conference on Economics and Computation*: 75–76.

Houston, J. B., Hawthorne, J., Spialek, M. L., Greenwood, M., and McKinney, M. S. (2013). Tweeting during presidential debates: Effect on candidate evaluations and debate attitudes. *Argumentation and Advocacy* 49 (4): 301–311.

Huang, J. T., Kaul, R., and Narayanan, S. (2022). The Causal Effect of Attention and Recognition on the Nature of User-Generated Content: Experimental Results from an Image-Sharing Social Network.

Huang, Y., Jasin, S., and Manchanda, P. (2019). “Level up”: Leveraging skill and engagement to maximize player game-play in online video games. *Information Systems Research* 30 (3): 927–947.

Jayachandran, S., De Laat, J., Lambin, E. F., Stanton, C. Y., Audy, R., and Thomas, N. E. (2017). Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation. *Science* 357 (6348): 267–273.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika* 36 (1/2): 149–176.

Khalil, H. and Ebner, M. (2014). MOOCs completion rates and possible methods to improve retention-A literature review. In: *EdMedia+ innovate learning*. Association for the Advancement of Computing in Education (AACE): 1305–1313.

Koller, D., Ng, A., Do, C., and Chen, Z. (2013). Retention and intention in massive open online courses: In depth. *Educause review* 48 (3): 62–63.

Ksiazek, T. B., Peer, L., and Lessard, K. (2016). User engagement with online news: Conceptualizing interactivity and exploring the relationship between online news videos and user comments. *New media & society* 18 (3): 502–520.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Lee, D., Hosanagar, K., and Nair, H. S. (2018). Advertising content and consumer engagement on social media: Evidence from Facebook. *Management Science* 64 (11): 5105–5131.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36 (4): 1234–1240.

Levin, L., Lewis, M. S., and Wolak, F. A. (2017). High frequency evidence on the demand for gasoline. *American Economic Journal: Economic Policy* 9 (3): 314–47.

Li, Y. and Xie, Y. (2020). Is a picture worth a thousand words? An empirical study of image content and social media engagement. *Journal of Marketing Research* 57 (1): 1–19.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lu, S., Xie, Y., and Chen, X. (2022). Immediate and enduring effects of digital badges on online content consumption and generation. *International Journal of Research in Marketing*.

McInnes, L., Healy, J., and Astels, S. (2017). hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* 2 (11): 205.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

McKenzie, D. (2017). Identifying and spurring high-growth entrepreneurship: Experimental evidence from a business plan competition. *American Economic Review* 107 (8): 2278–2307.

Narang, U., Yadav, M. S., and Rindfleisch, A. (2022). The "idea advantage": how content sharing strategies impact engagement in online learning platforms. *Journal of Marketing Research* 59 (1): 61–78.

Neeraj (2023). User-Generated Content(UGC) Stats And Facts – 2023. `https://taggbox.com/blog/user-generated-content-facts-and-stats/`.

Onah, D. F., Sinclair, J., and Boyatt, R. (2014). Dropout rates of massive open online courses: behavioural patterns. *EDULEARN14 proceedings* 1: 5825–5834.

Pattabhiramaiah, A., Sriram, S., and Manchanda, P. (2019). Paywalls: Monetizing online content. *Journal of marketing* 83 (2): 19–36.

Patterson, R. W. (2018). Can behavioral tools improve online student outcomes? Experimental evidence from a massive open online course. *Journal of Economic Behavior & Organization* 153: 293–321.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Salganik, M. J. and Watts, D. J. (2008). Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market. *Social psychology quarterly* 71 (4): 338–355.

Sunyer, J. (2014). The threat facing online comments. `https://www.ft.com/content/9c0cf256-e197-11e3-b7c4-00144feabdc0`.

Valsesia, F., Proserpio, D., and Nunes, J. C. (2020). The positive effect of not following others on social media. *Journal of Marketing Research* 57 (6): 1152–1168.

Van Hentenryck, P. and Coffrin, C. (2014). Teaching creative problem solving in a MOOC. In: *Proceedings of the 45th ACM technical symposium on Computer science education*: 677–682.

Vana, P. and Lambrecht, A. (2021). The effect of individual online reviews on purchase likelihood. *Marketing Science* 40 (4): 708–730.

Zhang, D. J., Allon, G., and Van Mieghem, J. A. (2017). Does social interaction improve learning outcomes? Evidence from field experiments on massive open online courses. *Manufacturing & Service Operations Management* 19 (3): 347–367.

Zhou, M., Chen, G. H., Ferreira, P., and Smith, M. D. (2021). Consumer Behavior in the Online Classroom: Using Video Analytics and Machine Learning to Understand the Consumption of Video Courseware. *Journal of Marketing Research* 58 (6): 1079–1100.

Appendix A - Alternative measure of UGC engagement

In this appendix, we report the results from estimating our model using an alternative measure of engagement with UGC. This alternative measure is computed as the sum of both consumption of UGC and creation of UGC. Although this measure weighs consumption and creation of UGC equally, we find that the results using this measure of engagement with UGC are similar to the results obtained using just the consumption of UGC as the measure of engagement with UGC.

Table 7: The effect of popularity, relevance, and entropy on user engagement on course forum.

	Dependent variable:		
	$\sinh^{-1}$(Engagement with forum)		
	(1)	(2)	(3)
Forum_Visited	1.560***	0.818***	0.512***
	(0.090)	(0.095)	(0.086)
Forum_Visited x $\sinh^{-1}$(Popularity)	−0.017	0.049***	0.061***
	(0.012)	(0.012)	(0.011)
Forum_Visited x $\sinh^{-1}$(Relevance)	0.152***	0.053**	0.042**
	(0.020)	(0.021)	(0.020)
Forum_Visited x $\sinh^{-1}$(Variety)	−0.334***	−0.146**	−0.023
	(0.061)	(0.061)	(0.056)
User FE		X	X
Course-Module-Week FE			X
Observations	115,036	115,036	115,036
R^2	0.461	0.714	0.778
Adjusted R^2	0.461	0.525	0.545

Note: *p<0.1; **p<0.05; ***p<0.01

All standard errors are clustered at user level.

Table 8: The effect of user engagement with course forums (IV estimates)

	Dependent variable:					
	$\sinh^{-1}$(Time spent on course material)			$\sinh^{-1}$(Performance)		
	(1)	(2)	(3)	(4)	(5)	(6)
Engagement with forum (β)	4.745***	0.681***	0.495***	4.367***	0.094***	0.069***
	(0.048)	(0.025)	(0.021)	(0.044)	(0.010)	(0.009)
Visited$_F orum(\alpha)$	2.288***	−0.522	0.673***	0.119	0.163	0.334***
	(0.305)	(0.320)	(0.189)	(0.097)	(0.108)	(0.091)
Constant	4.080***			4.990***		
	(0.013)			(0.004)		
User FE		X	X		X	X
Course-Module-Week FE			X			X
Observations	115,036	115,036	115,036	115,036	115,036	115,036
R^2	0.007	0.660	0.772	0.0001	0.561	0.649
Adjusted R^2	0.007	0.434	0.532	0.0001	0.270	0.277

Note: *p<0.1; **p<0.05; ***p<0.01

All standard errors are clustered at user level.

Appendix B - Elasticity calculation

In this appendix, we illustrate how we translate our parameter estimates into elasticity estimates

Consider a regression of the form,

$sinh^{-1}(y) = \alpha + \beta * sinh^{-1}(x) + \epsilon$

After estimating such a regression, to compute elasticities, we need to compute the derivative. To compute the derivative, we need to recover the left-hand side. To recover the left-hand side, we revert the IHS transformation by applying the hyperbolic sine transformation on both sides.

Therefore,

$$y = sinh(\hat{\alpha} + \hat{\beta} * sinh^{-1}(x) + \hat{\epsilon})$$

To compute the derivative in IHS-IHS models, we can apply the chain rule of differentiation to obtain,

$$\frac{\hat{dy}}{dx} = \frac{\hat{\beta} * cosh(\hat{\alpha} + \hat{\beta} * sinh^{-1}(x) + \hat{\epsilon})}{\sqrt{x^2+1}}$$

Since the argument inside the cosh function is $sinh^{-1}(y)$,

$$\frac{\hat{dy}}{dx} = \frac{\hat{\beta} * cosh(sinh^{-1}(y))}{\sqrt{x^2+1}}$$

Since hyperbolic functions satisfy the fundamental identity:

$$cosh(t)^2 - sinh(t)^2 = 1,$$

$$\implies cosh(t)^2 = 1 + sinh(t)^2,$$

$$\implies cosh(t) = \sqrt{1 + sinh(t)^2},$$

Now, we can compute the derivative as,

$$\frac{\hat{dy}}{dx} = \frac{\hat{\beta} * \sqrt{1 + sinh(sinh^{-1}(y))^2}}{\sqrt{x^2+1}}$$

Since $sinh(sinh^{-1}(t)) = t$,

$$\frac{\hat{dy}}{dx} = \frac{\hat{\beta} * \sqrt{1+y^2}}{\sqrt{x^2+1}}$$

Now, we can compute the elasticity as,

$$\frac{\hat{dy}}{dx} * \frac{x}{y} = \frac{\hat{\beta} * \sqrt{1+y^2}}{\sqrt{x^2+1}} * \frac{x}{y}$$

Note that the elasticity will be undefined when y=0. Hence, following the practice in the extant literature using IHS transformations, we report the average elasticity estimates for the forum visitors in our sample.

Appendix C - Topic model estimation

To characterize common themes in a corpus of text, topic models have proven to be a powerful unsupervised learning tool. Traditional models, such as Latent Dirichlet Allocation (LDA) (Griffiths et al., 2003), describe a document as a bag-of-words and model each document as a mixture of latent topics. However, one limitation of LDA is that it disregards semantic relationships among words. This limitation inhibits this model from performing well on short texts that are classically noisy and sparse, and therefore lack sufficient information for LDA-based models which rely mainly on statistical learning from information embedded within word co-occurrences for topic discovery.

Since our corpus is made up of short texts (i.e., thread titles), we utilize semantic information encoded within pre-trained text embeddings. Text embedding techniques have rapidly become popular in natural language processing. Specifically, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and their variants (e.g., (Lee et al., 2020), (Liu et al., 2019), (Lan et al., 2019)), have shown great results in generating contextual vector representations of words, sentences, and paragraphs. The semantic properties of these vector representations allow the meaning of texts to be encoded in such a way that the vectors representing similar texts are close together in the vector space.

We use a recently developed technique BERTopic (Grootendorst, 2022) that leverages the semantic properties of BERT-based sentence vector representations in conjunction with non-linear, graph-based dimension reduction and clustering methods to estimate a topic model.

Following the BERTopic framework, first, we encode each thread title into 384-dimensional vectors using a BERT-based sentence transformer (Reimers and Gurevych, 2019). Subsequently, we reduce these high-dimensional embeddings into low-dimensional embeddings using a dimension reduction technique and then cluster these low-dimensional embeddings using a clustering technique. Finally, a class-based TF-IDF procedure models the importance of words in the identified clusters. This generates a topic-word distribution for each cluster, which is then interpreted by the researcher as a topic.

Note that the modular nature of the BERTopic framework allows the researcher to use a dimension reduction and clustering method adequate for their application. We stick to the default dimension reduction method (Uniform Manifold Approximation, and Projection (UMAP) (McInnes et al., 2018) and clustering method (Hierarchical Density-based spatial clustering for applications with noise (HDBSCAN) (McInnes et al., 2017) in BERTopic. Since we use HDBSCAN, which is a soft clustering technique, the topic model outputs a title-topic probability matrix that characterizes the distribution of topics in a thread title. An illustrative example of the topic model's performance for one of the courses in our panel is provided in figure 12

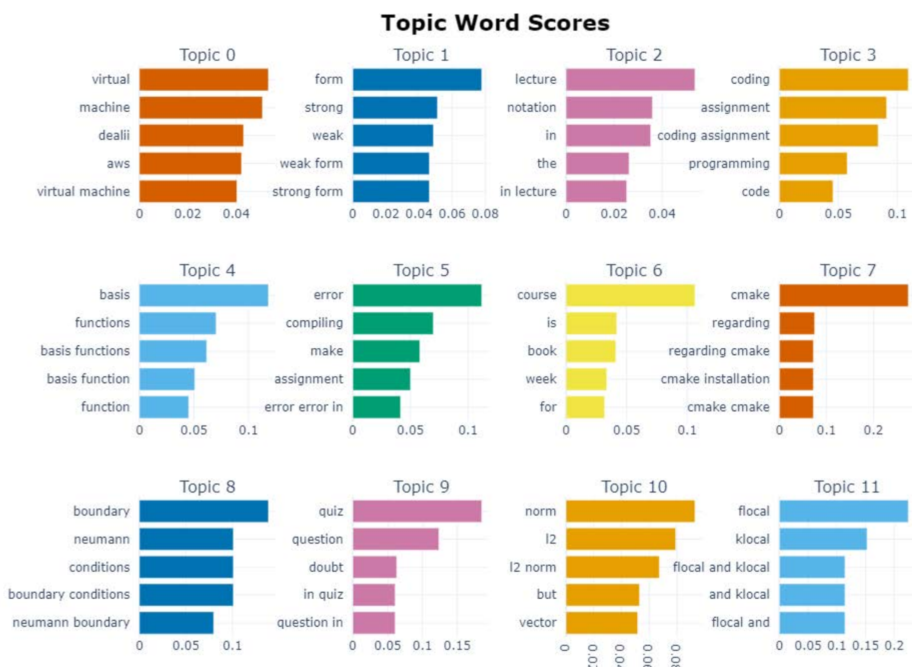


Figure 12: Illustrative example of the performance of the topic model

Appendix D - Ruling out signaling of course material

In our analysis in the body, we showed that popularity and relevance do shift users' engagement with forum content. However, one threat to our analyses is that these instruments may reflect the quality of course material in a given course module. This could induce users to directly increase their engagement with the course material. In this appendix, we use the variation in our instruments across users who visit a forum's homepage but do not click on any threads to rule this mechanism out. If the instrumental variables are indeed signaling the quality of course material, users who don't engage with the threads must be treated as well and the instrumental variables would be associated with changes in user engagement with course content.

To investigate this, we consider two sets of observations:

- User-course module-week tuples where the user visited forum's homepage and clicked on at least one thread.
- User-course module-week tuples where the user visited forum's homepage but did not click on any thread.

We show the results from estimating the reduced form regression on the first sample in Table 9. Subsequently, we present the results from estimating the reduced form regression in the second subsample in Table 10. The estimates reported in columns (3) and (6) of tables 9-10 demonstrate that (a) the instruments do shift the outcomes in the case where users engaged with the forum, and (b) the instruments do not have an effect when users did not click on threads but visited forum's homepage. This observation means that mere exposure to a vibrant and attractive forum homepage cannot directly lead to higher engagement with course content.

Table 9: Effect of the instrumental variable on the dependent variables when users engaged with the forum

	Dependent variable:					
	$\sinh^{-1}$(Time spent on course material)			$\sinh^{-1}$(Performance)		
	(1)	(2)	(3)	(4)	(5)	(6)
$\sinh^{-1}$(Popularity)	0.418***	0.350***	0.196***	0.061***	0.029***	0.039***
	(0.016)	(0.034)	(0.021)	(0.005)	(0.006)	(0.010)
$\sinh^{-1}$(Relevance)	0.593***	0.154***	0.127***	0.067***	0.050***	0.028***
	(0.012)	(0.021)	(0.019)	(0.005)	(0.006)	(0.004)
Constant	5.434***			5.062***		
	(0.051)			(0.019)		
User FE		X	X		X	X
Course-Module-Week FE			X			X
Observations	27,880	27,880	27,880	27,880	27,880	27,880
R^2	0.021	0.714	0.948	0.002	0.772	0.933
Adjusted R^2	0.021	0.604	0.915	0.002	0.684	0.890

Note: *p<0.1; **p<0.05; ***p<0.01
All standard errors are clustered at user level.

Finally we present, the 2SLS results using the first subsample in Table 11, that is user-module-week tuples where the forum was visited and at least one thread was clicked on, and computed the elasticities similar to those calculated for estimates in Table 4. The results are consistent with those obtained using the full sample. On average, a 10% increase in a user's UGC engagement increases that user's engagement with the core product by 4.64% (versus 4.32% for the full sample) and performance by 0.65% (0.60% for the full sample).

Table 10: No evidence of violation of exclusion restriction in the subsample of forum visits where no thread was viewed

	Dependent variable:					
	$\sinh^{-1}$(Time spent on course material)			$\sinh^{-1}$(Performance)		
	(1)	(2)	(3)	(4)	(5)	(6)
$\sinh^{-1}$(Popularity)	0.370***	−0.044	−0.218	0.018*	0.014	−0.031
	(0.033)	(0.039)	(0.161)	(0.010)	(0.011)	(0.022)
$\sinh^{-1}$(Relevance)	0.288***	0.254***	0.041	−0.007	0.010	0.035
	(0.028)	(0.023)	(0.509)	(0.007)	(0.007)	(0.067)
Constant	5.091***			5.073***		
	(0.095)			(0.027)		
User FE		X	X		X	X
Course-Module-Week FE			X			X
Observations	2,825	2,825	2,825	2,825	2,825	2,825
R^2	0.027	0.917	0.997	0.0004	0.915	0.999
Adjusted R^2	0.026	0.504	0.906	−0.0003	0.489	0.997

Note: *p<0.1; **p<0.05; ***p<0.01

All standard errors are clustered at user level.

Table 11: The effect of user engagement with course forums where users engaged with the forum (IV estimates)

	Dependent variable:					
	$\sinh^{-1}$(Time spent on course material)			$\sinh^{-1}$(Performance)		
	(1)	(2)	(3)	(4)	(5)	(6)
Engagement with forum (β)	4.998***	0.665***	0.493***	4.558***	0.085***	0.069***
	(0.055)	(0.028)	(0.022)	(0.049)	(0.010)	(0.010)
Constant	4.579***			4.956***		
	(0.207)			(0.070)		
User FE		X	X		X	X
Course-Module-Week FE			X			X
Observations	27,880	27,880	27,880	27,880	27,880	27,880
R^2	−0.140	0.560	−0.007	−0.017	0.772	0.923
Adjusted R^2	−0.141	0.390	−0.650	−0.017	0.684	0.874

Note: *p<0.1; **p<0.05; ***p<0.01

All standard errors are clustered at user level.

Appendix E - Sequential interactions with forum's content and aggregation

In this appendix, we try to assess the sensitivity of our parameter estimates to potential violations of the exclusion restriction stemming from sequential interactions with the forum within a module. We do this by estimating our model on the subsample of forum visits, where the forum was visited on a single day. Since we consider the value of the instruments after the first exposure to the forum within each module in our analyses, examining this subsample aids in assessing the robustness of the results to potential violations of the exclusion restriction stemming from sequential interactions with the forum within a module.

Table 12: The effect of user engagement with course forums (IV estimates)

	Dependent variable:					
	$\sinh^{-1}$(Time spent on course material)			$\sinh^{-1}$(Performance)		
	(1)	(2)	(3)	(4)	(5)	(6)
Engagement with forum (β)	3.954***	0.843***	0.665***	4.982***	0.123***	0.120***
	(0.012)	(0.036)	(0.031)	(0.004)	(0.014)	(0.013)
Constant	4.108***			4.859***		
	(0.257)			(0.076)		
User FE		X	X		X	X
Course-Module-Week FE			X			X
Observations	25,916	25,916	25,916	25,916	25,916	25,916
R^2	−0.453	0.592	0.667	−0.085	0.813	0.846
Adjusted R^2	−0.453	0.443	0.544	−0.085	0.745	0.789

Note: *p<0.1; **p<0.05; ***p<0.01
All standard errors are clustered at user level.