Addressing Common Analytic Challenges to Randomized Experiments in MOOCs: Attrition and Zero-Inflation

Anne Lamb Harvard University
Jascha Smilack Harvard University
Andrew Ho Harvard University
Justin Reich Harvard University

ABSTRACT
Massive open online course (MOOC) platforms increasingly allow easily implemented randomized experiments. The heterogeneity of MOOC students, however, leads to two methodological obstacles in analyzing interventions to increase engagement. (1) Many MOOC participation metrics have distributions with substantial positive skew from highly active users as well as zero-inflation from high attrition. (2) High attrition means that in some experimental designs, most users assigned to the treatment never receive it; analyses that do not consider attrition result in "intent-to-treat" (ITT) estimates that underestimate the true effects of interventions. We address these challenges in analyzing an intervention to improve forum participation in the 2014 JusticeX course offered on the edX MOOC platform. We compare the results of four ITT models (OLS, logistic, quantile, and zero-inflated negative binomial regressions) and three "treatment-on-treated" (TOT) models (Wald estimator, 2SLS with a second stage logistic model, and instrumental variables quantile regression). A combination of logistic, quantile, and zero-inflated negative binomial regressions provide the most comprehensive description of the ITT effects. TOT methods then adjust the ITT underestimates. Substantively, we demonstrate that self-assessment questions about forum participation encourage more students to engage in forums and increases the participation of already active students.

Categories and Subject Descriptors
K.3.1 Distance Learning.

Keywords
MOOCs; A/B testing; randomized controlled trials; attrition; zero-inflation; treatment-on-treated; engagement

INTRODUCTION

While the first years of massive open online course (MOOC) research have been characterized by post-hoc analysis of observational data [4, 6], MOOC platforms are beginning to fulfill their promise of enabling easily implemented randomized experiments [2, 8]. Researchers and course developers can randomly assign students to encounter different course materials or questions. Because MOOCs generate detailed data from thousands of users around the world, the effects of these interventions can be evaluated on a wide variety of outcome measures, including student performance, persistence, and participation. Along with the benefits of these abundant data, experiments conducted in MOOCs present researchers with challenges not often encountered in traditional education settings or even other online venues.

This paper was accepted to the 2015 ACM Learning@Scale Conference.

Student heterogeneity is a distinctive feature of MOOCs. Some students register for a MOOC to browse the materials, while others fully commit to completing all course activities[13]. Many students stop participation soon after registration, while a small number of students are extremely active [6]. This variation presents two challenges to analyzing MOOC experiments. First, the diversity of MOOC students leads to unusual distributions of their behavior. Measures of participation and engagement are often operationalized as counts of activities, such as page views or problem attempts. These counts are zero-inflated by the many browsers who stop out early and right-skewed by the activity of the most devoted students. These unusual distributions make it difficult to perform statistical tests and make appropriate inferences about the intervention.

Second, high levels of attrition lead to complexities in determining whether a student has been "exposed" to a treatment. Social scientists often distinguish between analyses of the intent-to-treat (ITT) and treatment-on-treated (TOT) [3]. The former analyzes differences between persons *assigned* to control or experimental conditions, while the latter analyzes differences between persons who *experience* a treatment and those in the control condition. Since attrition from MOOCs is high, particularly at the beginning of a course, intent-to-treat studies may underestimate the effect of an intervention. To determine the "true" effect of the intervention, additional analytic effort is necessary to determine whether or not a participant actually received the treatment and correct for underestimation.

We explore these two issues—analyzing zero-inflated, skewed participation metrics and estimating treatment-on-treated effects—in the context of an experimental intervention to increase discussion forum participation in an edX MOOC, JusticeX.

BACKGROUND ON JUSTICEX

Michael Sandel's Justice course at Harvard University introduces students to theories of justice, and encourages them to critically examine their own views on moral and political controversies. One of Harvard's most popular residential courses for undergraduates, it was offered as a HarvardX course for the first time in 2013 [14].

JusticeX (the online version of Justice) includes 24 content chapters, each of which consists of key readings and a 20-30 minute lecture video. After the readings and video, participants encounter a discussion prompt that poses a challenging moral dilemma without a clear answer. Example prompts include: *Would it be right, as a last resort, for the police to use torture to extract information from a suspected terrorist about the location of a bomb*? *How involved should government be in legislating morality*? Participants are invited to view arguments for both sides and then discuss these issues in the forums.

Fostering debate and dialogue among participants is critical to achieving the goals of the course. As is typical of MOOC courses, however, only a small proportion of all enrollees ever participate in the forums [9]. In the 2013 version of JusticeX, 16% of registrants posted at least once to the discussion forum. Although this was one of the highest levels of forum participation across Harvard and MIT MOOCs at the time, one of the main goals of the 2014 run of the course was to improve rates of forum participation.

Interventions to Improve Forum Participation

To encourage increased dialogue and engagement among participants, the instructional team implemented three interventions in the Spring 2014 version of JusticeX. The first treatment consisted of a *self-test participation check*. Following each of the 24 lectures, participants could complete a three-question, non-credit quiz. In the first six lectures, those who were assigned to the first treatment received an additional question: "Have you participated in the discussion forum?" The participants could select "No" or "Yes, I've upvoted 2 or more good comments and posted at least 1 response." This non-credit question served as a reminder for students and set norms for desired levels of participation.

The second treatment consisted of *discussion priming*, where students could access discussion summaries from the 2013 course. For the third treatment, the JusticeX team sent *discussion preview emails*, which previewed the course content for the coming week and summarized the topics that would be addressed in the discussion forums.

After comprehensive analysis of all three treatments, we found evidence that only the first intervention led to increased participation in the forum discussions. To simplify our discussion of how to address common methodological challenges, in the remainder of this paper, we limit our focus to the first treatment, the participation check. We mention the other treatments for transparency.

Random Assignment

By default, the A/B testing framework in edX dynamically assigns users to experimental groups whenever the system "needs to know" their group identity. Assignment can happen when students load a page or navigate to a sub-section that includes randomized content, or when faculty generate a course grade report that is dependent upon a randomized problem. Thus, students can be assigned to a treatment condition *without encountering the treatment*.

In the case of these JusticeX experiments, since at least one of the interventions (preview emails) was not embedded within the platform, a "trick" was used to assign all students as early as possible. A blank, or "hidden", HTML unit was embedded at the bottom of the first course page that nearly every student would encounter. Upon navigating to this page, the platform assigned students to the experimental or control group for each of the three interventions.

The disadvantage of assigning students to a treatment group as early as possible was that many assigned to treatment soon stopped out. Assigning students early, in a context with high attrition, increases the likelihood that assigned students never experience the treatment and that resulting distributions are zero-inflated. As expected, students were assigned in approximately equal numbers to treatment and control conditions in all three treatments. To avoid possible treatment contamination, we focus only on the "control" subgroup that received no treatments, and on the subgroup that received only the self-test participation check. There were 2,399 students assigned to the control condition. Of these students, 44% were female, the mean age was 33, and 64% had a BA or higher. The treatment group with 2,378 students was statistically indistinguishable from the control group on demographic characteristics.

Operationalizing Engagement as "Forum Actions"

Like many indirect measures of engagement in online environments, we operationalize forum participation as a count variable. We count the number of times students conduct any of four forum actions: creating a new post, replying to an existing post, editing a post or reply, and upvoting a comment. Upvoting is a feature that allows users to "like" or agree with what another user has written, and all users can see how many votes a certain discussion post has received. We call the primary outcome of interest the number of "forum actions," an indicator of student participation. From this measure, we can infer whether experimental interventions increased student participation in the forums.

Ideally, the outcome measures for these experiments would also measure student learning, such as an assessment of student's moral reasoning ability. Since such complex domains are difficult to assess at scale, we are limited to examining how the intervention affected users' quantity of forum participation.

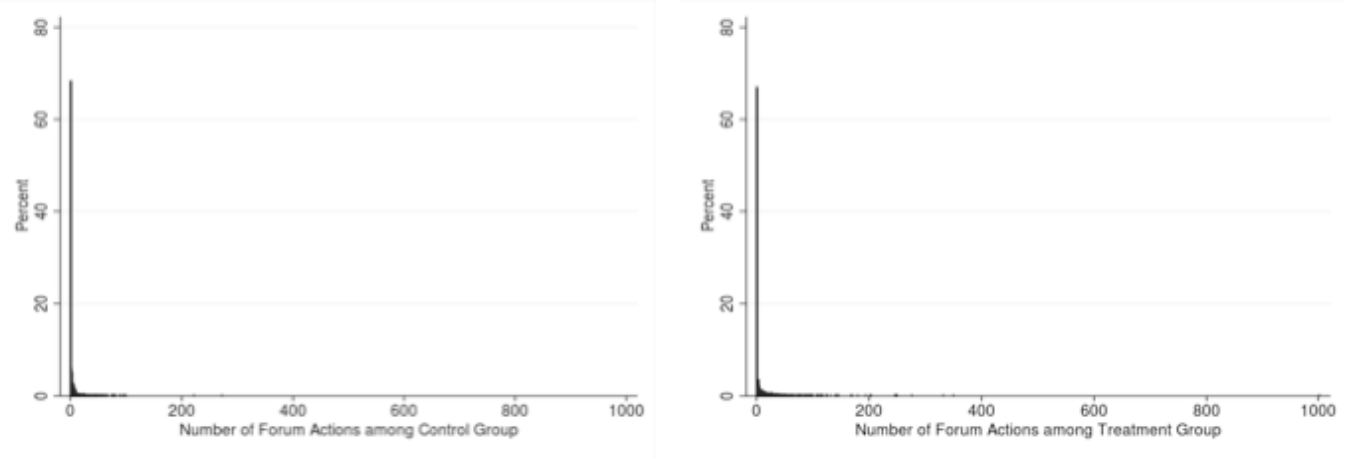

Figure 1: Distributions of forum actions for control group (*n*= 2399) and treatment group (*n*=2378) in 2014 JusticeX.

CHALLENGE #1: MODELING ZERO-INFLATED, SKEWED COUNT OUTCOMES

In this section, we compare a variety of approaches for estimating the effect of the self-test discussion check on forum actions, and we suggest advantages and disadvantages for each approach.

A visual inspection of the distributions of forum actions in the experimental and control conditions, shown in Figure 1, highlights the analytic challenges with our outcome variable. Due to a combination of high attrition and low overall forum participation, forum activity is zero-inflated relative to characteristic count distributions such as Poisson or negative binomial distributions. Moreover, right-hand (positive) skew results from the relatively small proportion of extremely active users in the course [7].

Summary statistics provide some indication of the intervention's effectiveness. The mean number of forum actions in the control group is 3.1 actions, compared to 7.3 actions in the experimental group. However, means of distributions can be a misleading indicator of central tendency in cases of zero-inflation or extreme values. Unlike in a Gaussian distribution, the mean does not represent a peak or any visually interpretable feature of the distribution. Percentiles can be more informative as summary statistics. In the control condition, participants from the 1st to the 68th percentile had 0 actions, those at the 75th percentile had 2 actions, and those at the 95th percentile has 16 actions. Comparatively, in the experimental group, while those at the median had 0 actions, those at the 65th, 75th, and 95th percentiles had 1, 4, and 39 actions respectively.

These simple comparisons suggest that the experimental intervention increased forum actions, but the nature of the distributional shift is unclear. Did it reduce the number of people with zero posts, without affecting the activity of frequent posters? Did it only engage the most active posters, without shifting the behavior of non-posters? Did it evenly shift the distribution to the right? In this section, we present four different analytic approaches to address these questions: OLS, logistic regression, quantile regression, and count regression. In these analyses, we include all those assigned to the experimental condition regardless of actual exposure to the treatment, an intent-to-treat analysis.

Ordinary Least Squares

Ordinary Least Squares (OLS) regression is familiar to many analysts, simple to conduct, and an appropriate exploratory step in analyzing MOOC participation metrics. The risk of OLS regression in this context is that it assumes the residuals will be normally distributed. While we can expect OLS to give a good linear approximation of the average treatment effect given our large sample size [3], a difference in means is not an ideal summary of treatment effects with a skewed distribution. Furthermore, the common interpretation of parameter estimates—that the effect of the intervention is of a similar magnitude at all levels of the outcome—is incorrect here. Finally, statistical inferences about the difference in means in the population will be incorrect, although this is generally not problematic when sample sizes are as large as they are MOOCs.

One common approach to dealing with non-normal distributions in an OLS framework is to transform them, although this typically does not work with count variables [11]. We briefly demonstrate the futility of this approach in this context. We begin by taking a series of transformations, including a log transformation, square and cubic roots, and a variety of inverse transformations. While the log and cubic root transformations (shown in Figure 2) succeeded in stretching out the values, the resulting distributions still appear far from normal.

Despite knowing that the assumption of normality would be violated, we persist with OLS using the non-transformed outcome. Even with the violation of normality, the OLS estimates are likely to be consistent given our large sample size. We perform a simple bivariate regression of the

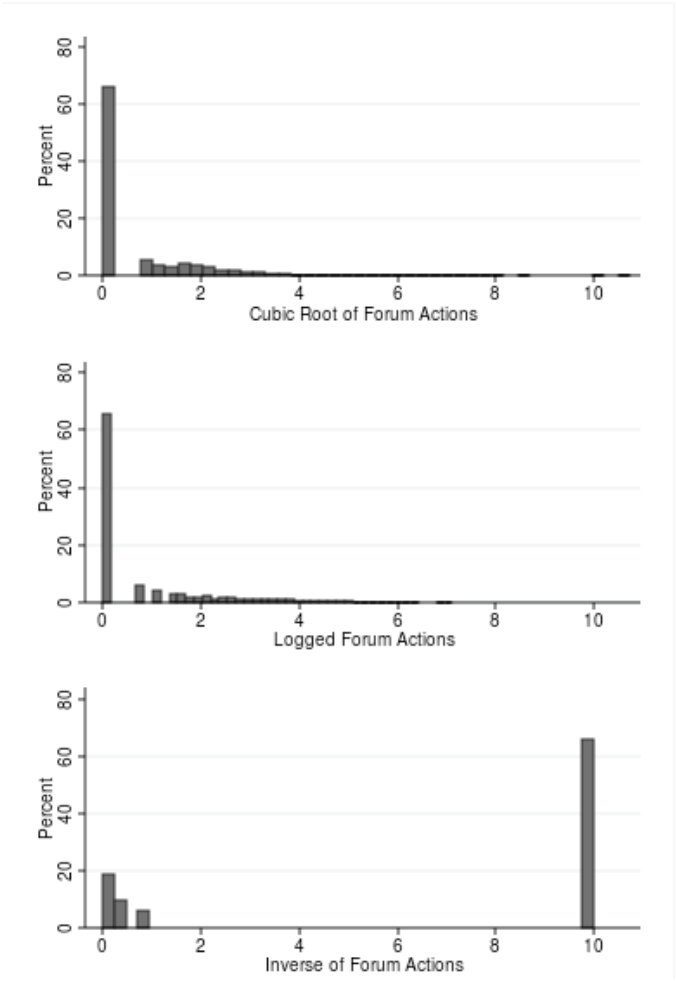

Figure 2: Cubic root, log, and inverse trasformations of forum actions

number of forum actions on a dichotomous indicator of whether the individual was assigned to the treatment. From this simple model, we obtain our first estimate of the average treatment effect: on average, users assigned to the treatment group had about 4.2 more forum actions relative to those in the control group (p<.001). Participants in the control condition have an average of 3.1 forum actions and those assigned to treatment have an average of 7.3 actions. In the two-sample case, the model will always recover the means exactly, and indeed we find that the OLS estimates are identical to the conditional means of the sample. Although the means are not the most interpretable indicator of central tendency in the outcomes, and the statistical inferences are suspect, the model accurately summarizes the mean treatment effect.

Logistic Regression

As an alternative expression of the effect of the treatment, we dichotomize the dependent variable, distinguishing only between zero actions and one or more actions. We can then fit the dichotomous outcome data using a logistic regression model. Dichotomization sacrifices all distinctions between nonzero counts of actions, however, if the purpose of the treatment is merely to increase the likelihood of any action, and there is no interest in distinctions between one action and thousands of actions, logistic regression on the dichotomous outcome is appropriate. Due to the dichotomization, this approach cannot estimate the additional effect of the treatment among those who already had at least one action.

The logit model estimates the log-odds of having had one or more forum actions conditional on assignment to treatment. On average, being assigned to the treatment group increases the log-odds of having a forum action by 0.284 (p<.001). We estimate that students in the control condition have a 31.8 percent probability of having at least one forum action, and students in the treatment condition have a 38.2 percent probability of having one action, a difference of 6.4 percentage points. Again, the two-sample setup ensures that the model recovers the observed logits and corresponding probabilities exactly. The p-values are slightly different than OLS given the loss of information in the outcome and the alternative model specification, however, the sample size is sufficiently large to render the difference moot.

Quantile Regression

We use quantile regression to begin to explore the distribution effects of the treatment. Quantile regression allows us to determine how an intervention changes the shape of an entire distribution – whether it spreads out or is compressed, or whether it shifts upward or downward. We use the conditional quantile function, $Q_t(Y_i|X_i)$, to describe the relationship between our predictor, assignment to treatment, and the outcome at different points along the conditional distribution of forum actions. At a specified quantile, the estimator works by asymmetrically weighting over- and under-predictions [3].

The single dichotomous predictor again allows observed conditional quantiles to be fit perfectly by the model at each quantile (Figure 3). At the median, users in both groups have zero forum actions. The number of forum actions diverges by the 65th percentile, when those in the treatment group have 1 forum action relative to the control group that has 0 actions.

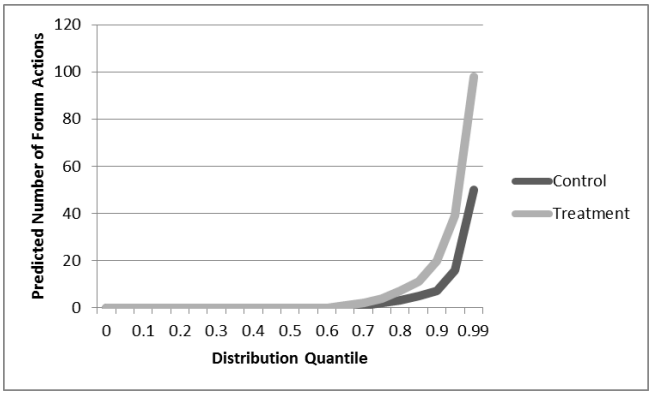

Figure 3. Predicted number of forum actions by quantile of forum action by experimental condition.

This discrepancy diverges even more substantially beyond the 75th percentile. At the 80th percentile, those in the treatment have approximately 4 more forum actions, close to the average treatment effect. By the 90th percentile, those in the treatment have 13 more forum actions, and by the 99th percentile, they have almost 50 more forum actions. From the 65th percentile and onward, those assigned to treatment have nearly double the number of forum actions as those in the control group.

Zero-Inflated Negative Binomial

Thus far, we have contrasted OLS (which expresses the average treatment effect), logistic regression (which fits differences in probabilities of having at least one forum activity), and quantile regression (which captures the effect of the intervention on the different quantiles of the distribution). In this section, we fit a model that seems appropriate for the abnormal shape of the outcome distribution. To do so, we turn to the family of regression models for count data. We choose a zero-inflated negative binomial (ZINB) model, because over half of the observations in our sample have zero values and the data are over-dispersed [12]. In the absence of zero-inflation, we could use an ordinary negative binomial.

The ZINB models zero-inflated, over-dispersed count data using two separate processes. First, it assumes the existence of "structural" zeros, and models the probability, for the entire sample, of having such a zero. A structural zero exists when it is not possible for the user to have a value other than zero. For instance, approximately 35% of users attrite before encountering either the discussion forum or the first self-assessment: they could not experience treatment or have an outcome. (Although we can observe users who have an actual zero value, the model does not allow us to distinguish between those with structural zeros and those with zero counts.) The ZINB then uses a negative binomial regression to model the count data among those without structural zeros. These processes together produce two estimates: the probability of having a structural zero and the predicted count of forum actions. We show the results of this model in Table 1. In column 1b, we estimate that the differences in structural zeros are not statistically significant. In column 1A, we estimate that the treatment causes a .0826 increase in the log of the counts. When we exponentiate this estimate, we obtain an incidence rate ratio of 2.3. Counts can be thought of as resulting from a process

	(1a) Forum Actions	(1b) Structural Zero
Treatment	0.826***	-0.103
	(0.1281)	(0.1927)
Constant	1.425***	-1.113
	(0.1469)	(0.5790)
Ln Alpha	1.775***	
	(0.2114)	
Observations	4,937	4,937

Robust standard errors in parentheses
*** p<0.001, ** p<0.01, * p<0.05

Table 1: Zero-inflated negative binomial regression of the number of forum actions on assignment to treatment.

with a rate of activity, in this case the rate at which a student completes forum actions over the duration of the course. We can interpret the incidence rate ratio as indicating that the participation self-check intervention increases the rate of forum activity by a factor of 2.3.

We display the predicted probabilities of structural zeros and the mean predicted counts resulting from the ZINB model in Table 2. In the first column, we show the predicted probability of structural zeros, though the difference is not statistically significant. Of those without structural zeros, users in the treatment have, on average, about 4.2 more forum actions than those in the treatment, recovering the sample means consistently with OLS results. We also compare differences in the predicted probabilities of counts along the distribution. Users in the treatment have higher fitted probabilities of having each corresponding count than those in the control, consistent with our findings that the treatment yields distribution effects not captured by the average effect. If the data are indeed generated by the ZINB process, this approach would be more parsimonious than quantile regression, where the latter would be encourage interpretation of chance differences between quantiles of the treatment and control distributions.

	Probability of Structural Zero	Predicted Mean Forum Actions	Pr ($Y \geq 3$) 75th Pct.	Pr ($Y \geq 7$) 90th Pct.	Pr ($Y \geq 50$) 99th Pct.
Control	25%	3.1	25%	13%	1%
Treatment	23%	7.3	32%	20%	4%

Table 2: Predicted probability of a degenerate zero, predicted mean forum actions, and predicted probabilities of the number of forum actions greater than or equal to the 75th, 90th, and 99th percentile of the distribution of forum actions for the control group, by treatment status, based on the ZINB model results.

Key Outcome Estimate	Interpretation	Advantages	Disadvantages
OLS			
β_{OLS}=4.2	The mean number of forum actions was 4.2 actions higher for the treatment group versus the control.	• Simple • Unbiased and consistent estimator of the average treatment effect	• Normality assumption violated • May be misinterpreted as estimated effect for all treated individuals
Logistic Regression			
Pr(Y|T=1) = 38.2% PR(Y|T=0) = 31.8%	Students in the treatment group have a 6.4 percentage-point higher probability of having at least one forum action compared to the control group	• Assumptions of the model are met	• Dichotomizing the outcome leads to a loss of information • Cannot estimate distribution effects
Quantile Regression			
$Q_Y(.75\|T=1) = 2$ $Q_Y(.95\|T=1) = 23$ $Q_Y(.99\|T=1) = 48$	At the 75th percentile, those assigned to treatment have 2 additional forum actions compared to those in the control. Above the 65th percentile, users in the treatment group in each quantile had at least double the forum actions as users in the control group.	• More comprehensive summary of distributional treatment effects • Appealing visualization of differences across the distribution	• No single summary statistics • Can only describe difference in each quantile
Zero-Inflated Negative Binomial			
β_{NB}=0.826 IRR= 2.3	No differences in structural zeros between treatment and control. The log of the counts of forum actions is 0.826 higher for users assigned to the treatment group relative to the control group. Being assigned to the treatment increases the rate of forum activity by a factor of 2.3.	• Designed to address zero-inflated, skewed distributions of count variables • Provides multiple avenues by which the intervention influences the outcome	• More complicated to fit and test • Less familiar to many researchers and practitioners • Parameter estimates more difficult to interpret and explain

Table 3: Summary of modeling strategies for the effect of assignment to treatment on forum actions in JusticeX

Summary of Intent-to-Treat Results

In Table 3, we summarize the pros and cons of the four methods we use for analyzing the intent-to-treat effects. OLS is simple to use and interpret, but suffers from a poor fit. Moreover, the average treatment effect fails to capture the full impact of the treatment. Dichotomizing the outcome and employing a logistic regression addresses issues with model fit, but again, provides only a limited explanation of the treatment effects. The quantile regression provides a more comprehensive picture of the distributional effects, showing that among those who actually use the forum, the treatment increases activity. However, as a series of separate regression models, it resists a simple summary statistic. Finally, the ZINB provides insight into both average effects and distribution effects. It shows that the treatment increases the rate of posting among those who do visit the forum discussion. We find a combination of these estimates useful in summarizing the ITT effects.

Assignment to treatment for the participation checks increased forum actions. On average, users assigned to the treatment had 4.2 more forum actions than users in the control group. We predict a 6.4 percentage point increase in in the probability of having any forum activity for those assigned to the treatment group over the control group. Additionally, assignment to treatment increases forum activity for those at the upper end of the distribution. Beginning at the 65th percentile of the outcome distribution, in which the control group had 0 actions relative to 1 action among the treatment students, students in the experimental condition had at least twice the forum actions at each respective quantile. We estimate that assignment to treatment increases the rate of forum actions by a factor of 2.3.

CHALLENGE #2: IDENTIFYING AND MODELING TREATMENT ON TREATED

In the previous section, we compared all students assigned to the self-test participation check treatment with an equivalent control group randomly assigned to no intervention. Since we know that many students attrite from MOOCs before engaging with any of the course content [6], we harbor an intuition that our intent-to-treat analyses underestimates the effect of the intervention on students who *actually encountered the intervention*.

The edX data stores provide a simple database of all those assigned to a treatment condition, but identifying students who actually received the intervention proves more complicated. It requires at least two steps: operationalizing the definition of "receiving the treatment" and then extracting evidence of receiving treatment from the tracking logs. For the self-test participation check intervention, we could operationalize receipt of treatment in several possible ways: simply viewing the page with the self-test questions, answering at least one question on the page, or answering the actual discussion self-check question. The first two options are more capacious, and potentially include students who navigated to the page but did not actually see the self-check question. The latter operationalization addresses this problem, but potentially omits students who saw the question and chose not to answer it, perhaps because they had not yet participated in the forum. Nevertheless, we chose this latter option, defining receipt of treatment as answering the self-check question. We acknowledge that alternative definitions may be defensible, and they could provide helpful sensitivity analyses.

We can further complicate the issue of treatment-on-treated by considering the "dosage" of the treatment. The self-test discussion check appeared after each of the first six lectures, so students could have been exposed to the treatment as many as six separate times. Of those assigned to the treatment group, 36% answered at least one discussion self-check, 10% answered at least three, and 1% answered all six. Thus, we could define receipt of treatment broadly as having encountered at least one discussion self-check, or we can identify the linear effect of multiple treatments by measuring different "doses" of treatment.

With these decisions in mind, we demonstrate three approaches for adjusting the intent-to-treat effects to estimate the treatment-on-treated (TOT) effects. All approaches attempt to correct the underestimation of treatment effects inherent in the intent-to-treat analyses presented in the prior section. First, we present a simple Wald estimate of the TOT, by dosage of treatment received. Doing so adjusts the average treatment effects from the intent-to-treat analysis. Second, we perform a manual two-stage least-squares estimate where we use a logistic regression in the second stage. Finally, we attempt to quantify the distribution effects of the TOT using an instrumental variables quantile regression.

	Some (Answered 1 or More Prompts)	Half (Answered 3 or More Prompts)	All (Answered 6 or More Prompts)
Wald Estimator	11.7	42.7	312.9
Conditional Sample Means	14.2	27.3	37.4

Table 4: Wald estimate of treatment on treated using the primary outcome, number of forum actions.

One strategy we cannot use is simply restricting our comparison to those who received the treatment and those who were assigned to the control group. Those who receive the treatment (our "compliers") likely differ along key unobservable characteristics related to our outcome relative to those who do not comply, thus eliminating the benefit of randomization. This is what fundamentally motivates our TOT analyses, and because of this, unlike our intent-to-treat estimates, our TOT estimates are not so easily compared to observational sample data.

Wald Estimator

The Wald estimator provides the simplest approach to average TOT estimates in the case of one predictor variable and one instrument. The Wald estimator is simply a ratio of the difference in mean forum actions between the control and assignment to treatment groups to the difference in proportion of users receiving the treatment (which in this case equals the proportion of those receiving the treatment from the assignment to treatment group). It rescales the OLS estimate of the average treatment effect by the proportion of those who actually received the treatment. In Table 4, we present the Wald estimator corresponding to the three different "dosages" of treatment received compared to the conditional sample means. While the Wald estimator for those receiving *some* of the treatment seems plausible, the estimator does not appear realistic for those receiving *half* or *all* of the treatment. The Wald estimator becomes less reliable when compliance rates are so low.

Two-Stage Least Squares Logistic Regression

In two-stage least squares (2SLS), the first stage regresses an endogenous predictor variable on an exogenous instrumental variable. The predicted values are then substituted into the structural equation of the second stage in order to predict the relationship between that endogenous variable and the outcome of interest using only the variation in the endogenous variable induced by the exogenous instrument. While typical 2SLS uses linear regression in both stages, we modify this approach to use a second stage logistic regression. By doing so, we aim to adjust our ITT estimates from the logistic regression in order to identify the effects on users receiving the treatment. Essentially, our approach amounts to using the probability of receiving

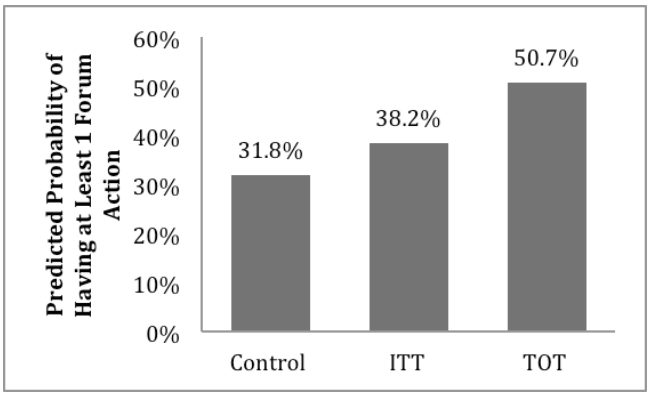

Figure 4: Predicted probabilities of having at least one forum action.

treatment as the predictor of having had at least one forum action in the logistic regression.

The results of this modified 2SLS approach reveal that, as expected, the predicted probability of having at least one forum action is higher for those induced to receive the treatment compared to the general sample of those assigned to treatment. In Figure 4, we compare the ITT and TOT estimates of the predicted probability of having forum activity. Those who actually answer the self-check question have a nearly 51% probability of having at least one forum action, which is almost 20 percentage-points higher than the probability for the control group. It is also substantially higher than the original ITT estimate. In fact, when comparing this final estimate to the conditional mean from the sample, it seems on target. Of individuals who answered the self-check, two-thirds had at least one forum action.

Instrumental Variable Quantile Regression

We first used quantile regression to determine how the distribution of forum actions changed as a result of assignment to treatment. Here, we employ an instrumental variables quantile regression to adjust the distribution effects relative to those who actually received the treatment. The instrumental variables approach allows us to incorporate a binary endogenous regressor indicating exposure to treatment into the quantile regression model [1]. Since, as with regular instrumental variables estimates, the IV quantile regression gives us the estimates for the compliers only, we expect the estimates to increase relative to the intent-to-treat estimates.

The results of the IV quantile regression confirm that those who receive the treatment are more active on the forum than those in the control. The effects of the treatment-on-treated begin to show by the 40th percentile, and the difference in predicted forum actions between the treated and the control group grows rapidly toward the higher end of the distribution. At the median, the compliers have, on average, 3 more forum actions than the control group. By the 80th percentile, the difference in forum actions has risen to 15, by the 95th percentile, the compliers have about 43

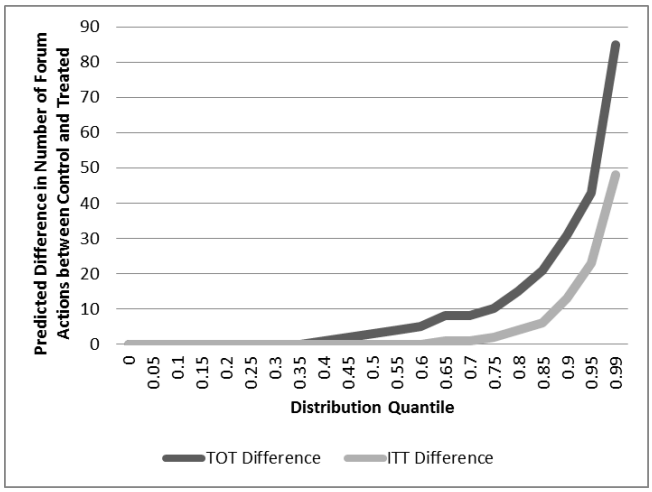

Figure 5: Difference between treatment and control conditions in the number of forum actions by quantile.

more actions than the control group, and at the 99th percentile the difference is 85 actions.

For the sake of comparison, in Figure 5 we plot the difference in predicted number of forum posts by quantile for (1) those assigned to treatment (ITT) versus the control, and (2) those induced to receive treatment (TOT) versus the control. As expected, the intent-to-treat effects underestimated the true effects of the treatment-on-the-treated. The entire curve shifts upward when correcting for the bias of the ITT estimates. While the ITT estimates indicated no difference in forum participation by participants all the way to the 60th percentile, the TOT estimates imply that users in the 40th percentile and greater have increased forum activity.

Summary of Treatment-on-Treated Results

Many JusticeX participants, due to attrition, our narrow definition of "receipt of treatment," or preferences for non-linear course progression, did not encounter the self-check question prompt intervention. While our intent-to-treat analyses provided strong evidence that assignment to treatment did have an effect on forum activity, we were also interested in quantifying the additional effect the intervention may have had on those users who actually received the treatment. Therefore, we used three methods to adjust our ITT estimates to uncover the impact of TOT.

The Wald estimator adjusted the average treatment effect by dividing the OLS estimate by the proportion of compliers. This simple adjustment implies that the ITT estimate should be increased by approximately a factor of 3 to uncover the effect of the intervention on those who received the treatment, or that those who received the treatment had, on average, approximately 12 more forum actions than those in the control. We were less certain of our Wald estimates when assessing higher doses of the treatment. Our manual 2SLS with a logistic second stage regression allowed us to adjust our original logistic results,

again testing whether the intervention encouraged additional students to become active on the forum. We find that while approximately one-third of users in the control are predicted to have at least one forum action, over half of those who receive treatment are predicted to do so. Finally, by employing the IV quantile regression, we notice that the trend predicted by the ITT estimate is magnified among the compliers; the treatment shifts the curve even higher for those who receive the treatment.

CONCLUSION

In analyzing the results of a randomized experiment in the MOOC JusticeX, we encountered two challenges that will likely impact other researchers analyzing MOOC experiments. First, because our outcome of interest was a participation metric—fundamentally a count variable—it had substantial right-hand skew and zero-inflation. Second, due to high rates of attrition at the beginning of the course, many assigned to treatment stopped out before receiving treatment. Precisely estimating the effects of interventions requires accounting for this "non-compliance."

To address the first issue, we use four approaches to estimate both the average treatment effects and the distribution effects of the participation self-check. While the naïve OLS estimates appear to adequately estimate the difference in means between the control group and those assigned to treatment, on the whole, the average treatment effect is an unsatisfactory statistic for quantifying the breadth of the treatment effects. Rather, we prefer to describe the effect of the treatment using a variety of summary statistics. Logistic regression with a dichotomized outcome variable uncovers the effect of assignment to treatment on getting students to participate at least once. We estimate that assignment to treatment increases the probability of participating in the forums by six percentage-points. We also found value in using an approach that could model distributional effects, in this case, the ZINB. This allows us to summarize the effect of the treatment on those who did participate in the forum*:* we estimate that the treatment increased the rate of student forum actions by a factor of 2.3. As a substantive conclusion, it appears that including self-assessment questions about forum participation is a low-cost, low-burden, and effective way of getting more students to participate more often in course discussion forums.

In addressing the second challenge of estimating treatment-on-treated effects, adjusting the treatment effects for compliance proves important given the high rates of attrition and inactivity. A simple Wald estimator provides an easy to implement adjustment, but may not work for situations with very low compliance, and we interpret the estimator with caution. Still, we believe that a simple intent-to-treat OLS analysis may substantially underestimate the effect of the treatment on those who actually answered the discussion prompt. With the Wald, we estimate that, on average, students who received treatment had 12 more forum actions than students in the control group. Again, this should not be interpreted as an average effect at any point in the distribution.

Our 2SLS logistic regression further emphasizes the importance of estimating TOT effects. We estimate that those who received the intervention had a 20 percentage-point higher probability of having a forum action than the control group. The IV quantile regression confirms the initial distribution effects, and shows that among those induced to receive treatment by random assignment, the magnitude of effects increase at each quantile along the distribution, though the IV quantile regression resists a simple summary statistic. In interpreting the TOT effects, our estimates likely represent an "upper bound," since we narrowly define receipt of treatment. These findings further reinforce the substantive conclusion that participation self-checks could be a valuable tool for increasing engagement.

One method that we would have liked to use to estimate the treatment-on-treated effects is an IV ZINB. Theoretically, this would be similar to running a two-stage least-squares model with the zero-inflated negative binomial as the second stage. Currently, an analogous program exists for doing this with a Poisson model [10] but has not yet been modified for use with the ZINB. Software innovations in this regard would be a useful contribution, as manual implementation will produce inaccurate standard errors.

We have elided two other important challenges to analyzing MOOC experiments in this paper. First, it may be possible for students skipping back and forth [5] through a course to encounter an opportunity to generate an outcome variable before exposure to treatment. For instance, some JusticeX students may have participated in the forums before experiencing the treatment intervention. This would indicate that we have overestimated the effect of treatment. Given our early treatment assignment and use of an aggregate outcome metric, we think the risk is small here, but it could be significant in other studies. Another important challenge that we did not address in this paper is dealing with multiple factor experiments and interactions. We limited our samples to only those exposed to one of three treatments and a true control group exposed to no treatments, but we could analyze students exposed to multiple treatments to explore possible interactions. When factorial designs intersect with issues of treatment compliance and complex outcome distributions, the analytical challenges multiply. Yet, as the online learning environment appears poised to host many different studies simultaneously, addressing interactions among MOOC experiments is a critical frontier for research.

Zero-inflated and skewed outcomes, high attrition after assignment to treatment groups, difficulty in defining and measuring exposure to treatment, difficulty in defining dosage of treatment, unclear timing between outcome and treatment, and factorial experimental designs all conspire to make the analysis of MOOC experiments more

complicated—in most circumstances—than running a simple t-test.

Our work to address a few common methodological challenges – zero-inflated, skewed outcome distributions and identification of treatment-on-treated effects—gives rise to a number of implications for MOOC designers who wish to implement randomized trials. First, researchers should consider tradeoffs in when they assign students to treatment conditions. Assigning all students immediately to treatment groups upon course entry allows for simple intent-to-treat analysis, but complicates treatment-on-treated analysis. Dynamically assigning students to treatment groups upon encountering a possible treatment condition might simplify treatment-on-treated analysis, but could complicate defining intent-to-treat analyses and lead to exposure to treatment after the outcome of interest. Researchers should also be careful in their analysis of experiments where actual treatment is difficult to measure, such as email-based experiments. These experiments allow flexible interactions with participants outside the confines of a platform, but it can be difficult to measure compliance with treatments, since many platforms mask whether a person receives or opens an email. When designing experimental interventions, researchers should also carefully consider how exposure to treatment will be defined. In our situation, having users actually answer the self-assessment problem provided one clear way, but not the only way. We encourage researchers to perform sensitivity analyses to better understand how defining this concept affects the results.

The introduction of easily implemented A/B tests in MOOC platforms greatly expands the possibility of conducting causal research in online spaces. Researchers using these platforms must be attentive to the peculiarity of MOOC outcomes and consider carefully how best to analyze experiments and present results in ways that will be accessible to both researchers and course developers.

REFERENCES

[1] Abadie, A., Angrist, J. and Imbens, G. Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings. *Econometrica*, *70*(1). 91-117. DOI=10.1111/1468-0262.00270.

[2] Anderson, A., Huttenlocher, D., Kleinberg, J. and Leskovec, J. Engaging with Massive Online Courses. in Proceedings of the 2014 International World Wide Web Conference, (Seoul, Korea, 2014), 687-698.

[3] Angrist, J. D. *Mostly harmless econometrics : an empiricist's companion.* Princeton University Press, Princeton, 2009.

[4] Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D. and Seaton, D. Studying learning in the worldwide classroom: Research into EdX's first MOOC. *Research & Practice in Assessment*, 8(2013), 13-25.

[5] Guo, P. J. and Reinecke, K. Demographic Differences in How Students Navigate Through MOOCs. in *Proceedings of the First ACM Conference on Learning @ Scale Conference.* (Atlanta, GA, 2014), 21-30.

[6] Ho, A. D., Reich, J., Nesterko, S. O., Seaton, D. T., Mullaney, T., Waldo, J. and Chuang, I. *HarvardX and MITx: The First Year of Open Online Courses, Fall 2012-Summer 2013.* HarvardX & MITx Working Paper No. 1. , 2014. Retrieved from SSRN: http://ssrn.com/abstract=2381263

[7] Huang, J., Dasgupta, A., Ghosh, A., Manning, J. and Sanders, M. Superposter Behavior in MOOC Forums. in *Proceedings of the First ACM Conference on Learning @ Scale Conference.* (Atlanta, GA, 2014), 117-126.

[8] Kizilcec, R. F., Schneider, E., Cohen, G. and McFarland, D. Encouraging Forum Participation in Online Courses with Collectivist, Individualist, and Neutral Motivational Framings. *eLearning Papers, 37*(2014), 13-22.

[9] Manning, J. and Sanders, M. How Widely Used Are MOOC Forums? A First Look. *Signals: Thoughts on Online Learning*, (2013). Retrieved from https://signalblog.stanford.edu/how-widely-used-are-mooc-forums-a-first-look/

[10] Nichols, A. IVPOIS: Stata module to estimate an instrumental variables Poisson regression via GMM. *Statistical Software Components*, (2008).

[11] O'Hara, R. B. and Kotze, D. J. Do not log-transform count data. *Methods in Ecology and Evolution*, *1*(2). 118-122. DOI=10.1111/j.2041-210X.2010.00021.x.

[12] Rabe-Hesketh, S. and Skrondal, A. *Multilevel and longitudinal modeling using Stata.* Stata Press Publication, College Station, Tex., 2008.

[13] Reich, J. MOOC Completion and Retention in the Context of Student Intent. *EDUCAUSE Review Online*, (Forthcoming).

[14] Reich, J., Nesterko, S. O., Seaton, D. T., Mullaney, T., Waldo, J., Chuang, I. and Ho, A. D. *JusticeX: Spring 2013 Course Report.* HarvardX Working Paper No. 4. , 2014. Retrieved from SSRN: http://ssrn.com/abstract=2382248