Causal Inference Project

Shay A. Geller*
Ben-Gurion University
Beer Sheva, Israel
gelleral@post.bgu.ac.il

ACM Reference Format:
Shay A. Geller. 2020. Causal Inference Project. In *Proceedings of Introduction to Causal Inference - A Machine Learning Perspective.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Nowadays, modern learning occurs outside of traditional or even computer-enabled classroom. Massive Open Online Courses (MOOCs), online discussion forums, and other online learning environments present a new paradigm for learning[7]. Despite posing new challenges for educators, online learning environments have become an increasingly popular method for e-learning and distance learning. Online learning environments also play a role in traditional environments as alternate instructional paradigms, such as flipped classrooms and other forms of blended learning [14]. Despite their widespread use, it is widely acknowledged that learning with online platforms can be a cold and detached experience [8], with the high disengagement and dropout rates in MOOCs providing an exemplary case[16].

Computer-mediated discussion forums include the use of posts and hashtags to convey opinions, ideas, and emotions and are commonly used in social media platforms [10]. The increasing prevalence of online student forums in courses thus provides new opportunities: for students to signal their affective state to others using hashtags and for teachers to recognize and to respond to these signals in a way that supports students' learning.

While many works aims to identify students emotions in online environments, especially confusion or frustration [16, 17] to understand how it affects disengagement or dropout, we will ask a much simpler question: Do students that their post got answered by other students in the current lecture, tend to write more posts in the next lecture? Under this research question lays the hypothesis that students' engagement depends not only on the course material but also affected by the behavior of their peer students.

2 DATA

Our empirical methodology is based on the Nota Bene (NB) platform, an open-source collaborative online annotation tool for PDF, HTML, and video files.[1] The course content (e.g., textbook) is uploaded to the NB website by instructors. Students log on to read and annotate the text. NB is used by hundreds of university courses with more than 40, 000 registered student users. Students can see the relevant posts from the class while reading specific sections of content. The inline nature of annotation enabled by NB has been shown to promote student-to-student feedback, to increase the contribution rates to the forum, and thereby to encourage learning [9].

Table 1: Hashtags counts and their associated emojis

Hashtag	Count	Emoji	Hashtag	Count	Emoji
#confused	1,228	😖	#help	420	😵
#curious	6,595	🤔	#question	7,574	❓
#interested	9,237	🤓	#useful	10,866	😀
#idea	9,673	💡	#frustrated	65	😡

Our study focuses on Bis2A, a general biology course required for all life sciences majors, many social sciences majors and bioengineering students at a large public university in California. The course was offered in 2018 and enrolled 785 students. The course consisted of 25 lectures, and students in the course received reading assignments on the material before each lecture, which was uploaded to NB. Each reading assignment had approximately 3-4 days to complete. The students were required to provide three meaningful posts to the reading assignments in NB before each class. This encouraged active participation in forum discussions. Students received additional credit for including at least one hashtag in at least one of their posts per lecture.

The NB interface displays the hashtags graphically using relevant emoji symbols in students' posts, shown in Figure 1. NB allows both students and instructors to filter posts by hashtags, helping them navigate to specific posts of interest. Students were able to choose any of eight possible hashtags. Table 1 shows the frequency counts of the different hashtags used by the students in the course. In total, out of 58, 811 posts submitted by students in the Bis2A course, 40, 842 posts contained at least one hashtag. The total number of hashtags across all posts was 45, 658.

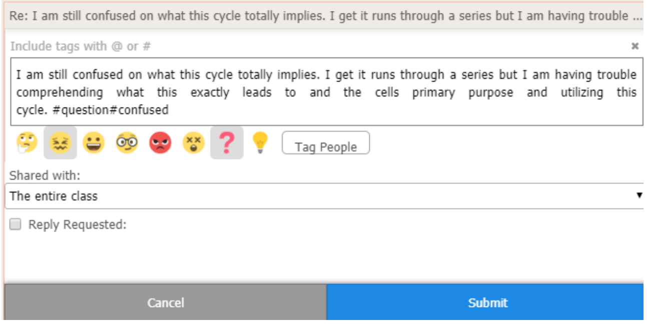
70% of the posts contain hashtags, well above the minimum requirement (1/3 of posts required a hashtag). This suggests that students perceive intrinsic value in the course from using hashtags, which reflects past work demonstrating the benefit of providing students with opportunities for self-assessment [13].

Figure2 shows the number of posts in each lecture and day, where the colors represent different lectures. In each lecture, the

Introduction to Causal Inference - A Machine Learning Perspective, March, 2020, Course from Tel Aviv University

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] http://nb.mit.edu/



Re: I am still confused on what this cycle totally implies. I get it runs through a series but I am having trouble ...

Include tags with @ or #

I am still confused on what this cycle totally implies. I get it runs through a series but I am having trouble comprehending what this exactly leads to and the cells primary purpose and utilizing this cycle. #question#confused

Figure 1: Nota Bene GUI including post with hashtag and relevant emoji.

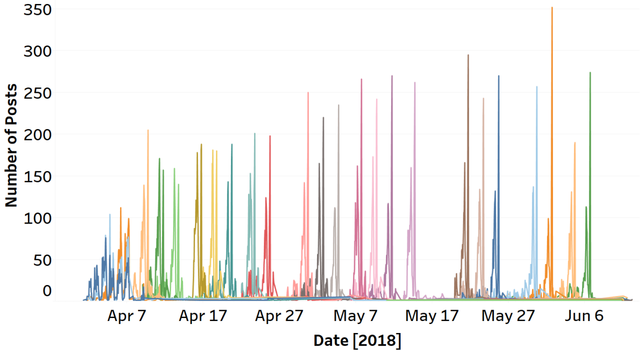


Figure 2: Distribution of posts over time. Each colors are different reading assignments and the picks are usually their due date.

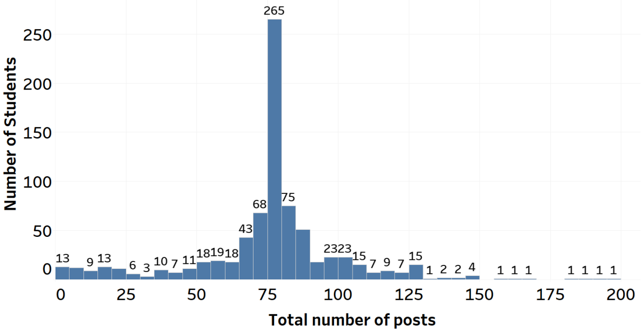


Figure 3: Histogram of total number of posts per student over the entire semester

pick usually represents the due date of the reading assignment. We can see that most of the posts are close to the due date and then rapidly decrease. This made us believe that some of the posts were written to get the course credit. We chose to remove them from our analysis because the students who wrote these posts probably knew that they have a minor chance to be answered.

In figure 3 shows a histogram of the total number of posts per student over the entire semester. We can see that the mean is around

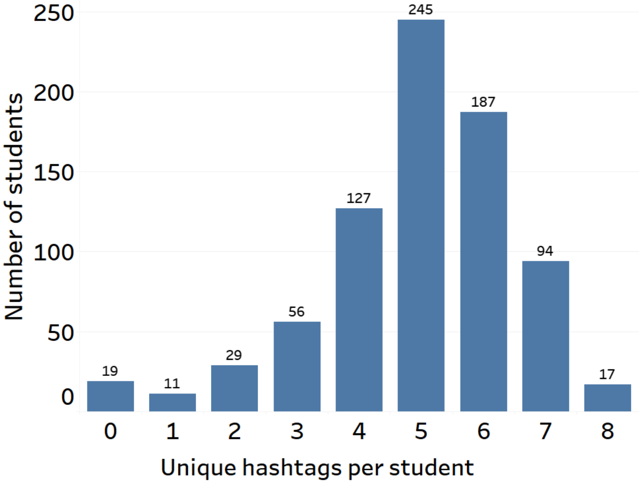


Figure 4: Histogram of unique hashtag use by students

75 posts, which is three posts per lecture, exactly as requested to get course credit. The histogram is nicely normally distributed, suggesting that some students contributed more than minimum requirements, but some less. In this work, we will test if the students who contributed more than minimal requirement did it because their posts got answered by other students.

Prior research on the NB platform found out that when there are too many participants in a discussion, it becomes saturated with annotations, and there is nothing left to say [9]. In larger sections, students might be adding comments to existing threads rather than starting their threads due to this saturation effect. They found out that the optimal section size is around 40 students. It is important to note that these sections existed only in the context of the online annotation forum and did not affect the amount of time students spent in class. So student's posts can be answered only by peer students within the same section.

2.1 Assumptions and Preprocessing

In this work, we are testing a causal effect in a case study experiment. Our data has a few characteristics that need to be addressed before applying the causal analysis. We will explain how our data can be approximated as a pseudo-randomized control trial (RCT) while following a pseudo target trial protocol. We will present the assumptions and the preprocessing required to address them. Moreover, we will state when our data was insufficient or missing for these assumptions so our results will be taken with caution.

2.1.1 Eligibility criteria. Eligible students for the trial are students that this is their first enrolment in the course. They also must be students that wrote at least one post in each reading assignment to show some participation activity. Also, we excluded posts that were written less than 5 hours before the due date because these posts had less chance to be answered and are only written to get the credit points. Our data do not contain whether it is the first enrolment of the student, but we were able to filter posts according to the due date and students according to participation in every lecture. This results with 247 students out of 779 initial students and 21,722 posts out of 58,811 initial posts.

2.1.2 Treatment strategies. We consider a treated student in lecture i as a student that other students answered at least one of his posts in lecture i. Any type of answer(short \ long \ hashtag) is fine. The treatment will be applied to each lecture.

The first part of the Stable Unit Treatment Value Assumption (SUTVA), *No Interference*, states that treatment for a specific person does not affect the outcome of any other person. We will assume that the student which his post got answered, do not affect the number of posts other students generated.

The second part of SUTVA, *Consistency* - states that the treatment should be well defined. We defined the treatment as any type of answer by another student.

2.1.3 Randomized assignment. In optimal condition, at the beginning of the semester, students will be separated into two groups: (1) Control - Students that their posts will be naturally replied (or not) by other students. (2) *Treatment* - Students that at least one of their posts in each lecture will be replied by another student (randomly). In our case, after each lecture, the control group will be the students that none of their posts answered, and the treatment group are all the other students.

2.1.4 Start, end, follow-up. The start of the experiment will be at the beginning of each reading assignment and will end up to 5 hours before the due date (similar reason as we described in the Eligibility criteria part). We will then test their number of posts in the following lecture. We will conduct these experiments over the entire semester. This results in 24 experiments in a semester, one for each lecture minus one, naturally because the first assignment will not have an outcome part. In our setting, the same student can be in the control and treatment group in different lectures.

2.1.5 Outcomes. For each student, we will count the number of posts he wrote at the next reading assignment.

2.1.6 Causal effect. We are interested in the average treatment effect of the treated (ATT). We will elaborate more on it next section.

2.1.7 Analysis plan. The causal analysis methods we use are propensity score matching(PSM), inverse propensity score weighting (IPSW), S-learner, and T-learner techniques. We will elaborate on them more in the method section.

In the next section, after presenting some background terms, we will state and explain another assumption that enables our causal analysis, like the strong ignorability assumption.

3 POTENTIAL OUTCOMES FRAMEWORK AND CAUSAL EFFECTS

In the potential outcomes framework, there are two possible treatments (e.g., active treatment vs. control treatment) and an outcome. Given a sample of subjects and a treatment, each subject has a pair of potential outcomes: $Y(0)$ and $Y(1)$, the outcomes under the control treatment, and the active treatment, respectively.

However, each subject receives only one of the control treatment or the active treatment. Let T_i be an indicator variable of subject i denoting the treatment received ($T_i = 0$ for control treatment vs. $T_i = 1$ for active treatment). Thus, only one outcome, $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$, is observed for each subject: the outcome under the actual treatment received.

For each subject, the effect of treatment is defined to be $Y_i(1) - Y_i(0)$. The average treatment effect (ATE) is defined to be $E[Y_i(1) - Y_i(0)]$ [5]. The ATE is the average effect, at the population level, of moving an entire population from untreated to treated. A related measure of treatment effect is the *average treatment effect for the treated*(ATT) [5]. The ATT is defined as $E[Y(1) - Y(0)|T = 1]$. The ATT is the average effect of treatment on those subjects who ultimately received the treatment. In a randomized control trial, these two measures of treatment effects coincide because, due to randomization, the treated population will not, on average, differ systematically from the overall population.

Naturally, in our study, not all posts will be answered. It would be unrealistically to estimate the effect of the treatment (answers to posts) if it were applied to all students. So in our experiment, ATT is of greater importance than ATE.

4 CAUSAL INFERENCE METHODS

In this section, we will describe the methods we used to asses the causal effect.

4.1 Propensity Score Matching

Propensity score (PS) is a method of reducing bias in treatment effect estimation. At its most basic form, PS is defined for binary treatment as:

$$\pi(X_i) = P(T_i = 1|X_i) \tag{1}$$

where $\pi(X_i)$ is the propensity score of subject i, T_i is the treatment, in our case, whether at least one of the student's posts got replied, and X_i represents the covariates (we will elaborate on them in the next section).

By Rosenbaum and Rubin [12], PS is the most basic function of covariates that has the balancing property, which means treatment assignment is independent of covariates given the propensity score. This, of course, requires all confounding variables to be known as well as the existence of a real choice between treatment and control for each patient at the time of treatment selection, both critical criteria for what Rosenbaum and Rubin [12] called the assumption of strong ignorability treatment assignment. We will describe our covariates in the next section and will clarify why we can assume strong ignorability in our experiment.

To produce the least biased propensity score model, it is important to not only include covariates that are correlated with treatment but also those correlated with outcome. Doing so would decrease the precision of the treatment effect estimate [3] variables whose removal result in insignificant changes in estimated treatment effect and an increase in precision are seen as unlikely confounders and can be safely removed from propensity model [4].

The average treatment effect can be computed from propensity score estimates using iterated expectations.

$$\mu = E[Y(1)-Y(0)] = E_{\pi(X_i)}[E[Y(1)|\pi(X_i)]-E[Y(0)|\pi(X_i)]] \tag{2}$$

where $E_{\pi(X_i)}$ is the expectation with respect to the distribution of $\pi(X_i)$ in the entire population.

x	Age	Age of the student
x	Sex	The sex of the student
x	Demographic	The origin country of the student
x	First Language	Student's mother tongue
✓	Author ID	The ID of the student
✓	Lecture number	The lecture's number
✓	#Posts	Number of posts in the current lecture
✓	#First-Posts	Number of posts that are the first post in the thread in the current lecture
✓	#Reply-posts	Number of posts that are a reply posts the current lecture
✓	#Students-in-section	Number of students in the section
✓	Verbosity	Average length of the posts the students wrote in the current lecture
✓	#Hashtags	Number of hashtags the students used in the current lecture

Table 2: Covariates list. x represents missing covariate, and ✓ represent observed covariate

The propensity scores produced can be used to find a conditional estimate of treatment effects given propensity score π, over the distribution of π. This can be best accomplished through matching between treatment and control patients, stratification, or using the PS directly as a covariate in the regression. Matching protects against misspecification of the propensity model but can significantly reduce sample size.

There are many techniques to perform propensity score matching (PSM) [2], and we chose the greedy nearest neighbor. Greedy nearest neighbor matching selects a treated subject and then selects as a matched control subject, the untreated subject whose propensity score is closest to that of the treated subject (if multiple untreated subjects are equally close to the treated subject, one of these untreated subjects is selected at random). So, to calculate the ATT using PSM, for each sample i with $T_i = 1$, we will find the nearest neighbour j according to propensity score, where $T_j = 0$, and then will calculate $S(i) = Y_i - Y_j$. The mean of all these $S(i)$s is the ATT.

4.2 Inverse Probability of Treatment Weighting

Inverse probability weighting (IPW) is a statistical technique for calculating statistics standardized to a pseudo-population different from that in which the data was collected.

Inverse probability of treatment weighting (IPTW) using the propensity score uses weights based on the propensity score to create a synthetic sample in which the distribution of measured baseline covariates is independent of treatment assignment. The use of IPTW is similar to the use of survey sampling weights that are used to weight survey samples so that they are representative of specific populations [11].

The inverse probability weighted (IPW) method is to weigh the treated and comparator(control) observations to make them representative of the population of interest. To estimate ATT, the weight for a treated subject is taken as one, and the weight for a comparator subject is defined as $\frac{\pi(X_i)}{1-\pi(X_i)}$ [5]. Suppose there are n subjects in the sample. Denote X_i, T_i, and Y_i, respectively, as the observed covariates, treatment assignment, and outcome for the ith subject $i = 1, ..., n$. According to [1], the IPTW, which is the IPW estimator for ATT is defined as

$$\hat{\mu}_{ATT,IPW} = \frac{\sum_{i=1}^{n} T_i Y_i}{\sum_{i=1}^{n} T_i} - \frac{\sum_{i=1}^{n}(1 - T_i)Y_i\pi(X_i)/(1 - \pi(X_i)}{\sum_{i=1}^{n}(1 - T_i)\pi(X_i)/(1 - \pi(X_i))} \quad (3)$$

In the IPTW, we obtained a substantial inverse propensity score, so we clipped the values to be in $[-10, 10]$ [6].

4.3 S-Learner and T-Learner

Heterogeneous treatment effect estimation via learning objectives can be implemented using any method that is framed as a loss minimization problem, such as boosting decision trees or other models. In this section, we focus on simulation experiments using the S-Learner and T-Learner. The S-Learner fits a single model for $f(X_i, T_i) = E[Y|X = X_i, T = T_i]$. T-Learner fits the functions $f_k(X_i) = E[Y|X = X_i, T = k]$ separately for $k \in \{0, 1\}$. Calculating the ATT in S-Learner and T-Learner settings is just the mean over $f(X_i, 1) - f(X_i, 0)$ or the mean over $f_1(X_i) - f_0(X_i)$ respectively over every sample i with $T_i = 1$.

After experimenting with different classifiers, best and most stable results were achieved with Linear Regression with L1 regularization (also referred to as Lasso).

5 CONFOUNDERS AND COVARIATES

Each row in our dataset represents the behavior of a single student in a specific lecture. Table 2 presents possible confounders. The rows with leading x represent missing confounders, and leading ✓ represents observed confounder. The observed confounders are used as covariates in the models we described in section 4. This results in a total of 305 covariates.

Due to privacy and time issues, we were not able to obtain some of the confounders. We hope to get them soon. Therefore, our results should be interpreted accordingly.

6 RESULTS

As a sanity test, we first checked the mean number of posts per lecture in the control and treatment groups. There are 3255 rows in the control group and 2793 rows in the treatment group. We found out that the students in the control group contributed 3.26(±0.10) posts on average in each lecture, and the treatment group contributed 3.647(±0.19) posts. This result is statistically significant at $p < 0.0001$. This is a piece of evidence that the treatment group contributed more posts than the control group. Now we will test whether it is because the posts of the treatment group were answered by other students.

For a baseline model that estimates the causal effect, we used the matching method from Causal Inference Python Package [15]. This package was used to infer causal relations in other works as well[18]. It addition to this baseline, we implemented in code[2] the models we described in section 4.

In table 3, we can see the ATT, the standard error, and confidence intervals for each method over all the lectures and all students. For example, by the T-Learner model, students that their posts got answered, tend to write 0.17 more comments in the next lecture. Some causal models find this number higher and some lower, but all methods returned positive ATT scores, and the confidence intervals do not contain 0, which suggests that they all find enough evidence for a positive causal effect.

Moreover, most of the results are close to the baseline model, except for the IPTW. This might be due to a biased propensity score model. We can see that our propensity score matching model (*PS Matching* in the table) is not similar to the baseline model. This can be explained by the implementation difference of the matching technique or the calculation of the propensity score.

In total, even though we see that there is a positive causal relationship between students that their posts got answered and their number of posts in the next lecture, it is not clinically significant, meaning, the relation is too weak to mean something useful in the educational context.

Another experiment we ran is grouping the data by lecture, and aggregating the results over lectures. For example, the data of lecture 3 contains rows from lectures 1,2, and 3. This way, we can evaluate an accumulated effect over the lectures. In Figure 5 we can see accumulated ATT and confidence intervals over lectures. We used the baseline model to get the ATT score. The figure contains results starting from lecture 3. This is due to too few control units compare to the number of features in the data of the first two lectures.

We can see that the ATT is positive over all of the lectures, and the confidence intervals narrowing down as the semester proceed, and the causal model gets more evidence. We can see that after lecture 12, all confidence intervals do not include zeros, and the ATT stabilized around 0.13. Lecture 25 result is similar to the result in the baseline model from table 3 because both of them contain all the data. In general, the results suggest that the causal effect is pretty much positive and constant over all of the lectures, with some exception of high variance at the beginning of the semester, which decreases over time.

7 DISCUSSION

In this work, we estimated the effect of reply to students' posts on their engagement. We found weak evidence that students that at least one of their posts got answered in one lecture will contribute more comments in the next lecture. Moreover, we found that the causal effect remains positive and constant over lectures.

Model	ATT	S.e	95% Conf. int.	
Baseline	0.129	0.044	0.044	0.215
IPTW	0.293	0.005	0.281	0.305
PS Matching	0.060	0.02	0.019	0.101
S-Learner	0.057	0.000	0.057	0.057
T-Learner	0.170	0.004	0.163	0.179

Table 3: ATT, standard error, and confidence intervals of each model

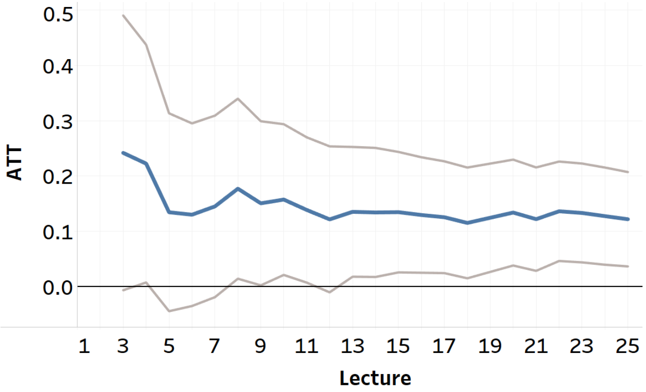


Figure 5: Accumulated ATT (blue) and confidence intervals (gray) over lectures

Even though all causal models in Table 3 suggested a positive effect, the effect is not clinically significant. Together with missing confounders in our data, we conclude that there is not enough evidence for a strong causal relationship, and more experiments should be done to claim such a relation.

Another thing to keep in mind is that one of our assumptions was that a student which his post got answered, do not affect the number of posts other students generate, meaning the No Interference assumption of SUTVA. This might not be the case because a student that his posts did not attract any replies may be jealous of students that their posts did attract replies and might be affected by it (for positive or negative).

REFERENCES

[1] Younathan Abdia, KB Kulasekera, Somnath Datta, Maxwell Boakye, and Maiying Kong. Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. *Biometrical Journal*, 59(5):967–985, 2017.

[2] Peter C Austin. A comparison of 12 algorithms for matching on the propensity score. *Statistics in medicine*, 33(6):1057–1069, 2014.

[3] M Alan Brookhart, Sebastian Schneeweiss, Kenneth J Rothman, Robert J Glynn, Jerry Avorn, and Til Stürmer. Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156, 2006.

[4] Holly A Hill and David G Kleinbaum. Bias in observational studies. *Wiley StatsRef: Statistics Reference Online*, 2014.

[5] Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

[6] Fan Li, Laine E Thomas, and Fan Li. Addressing extreme propensity scores via the overlap weights. *American journal of epidemiology*, 188(1):250–257, 2019.

[7] Tharindu Rekha Liyanagunawardena, Andrew Alexandar Adams, and Shirley Ann Williams. Moocs: A systematic study of the published literature 2008-2012. *The International Review of Research in Open and Distributed Learning*, 14(3):202–227, 2013.

[2] https://github.com/shaygeller/Causal-Inference-Course-Project.git

[8] Joanne M McInnerney and Tim S Roberts. Online learning: Social interaction and the creation of a sense of community. *Journal of Educational Technology & Society*, 7(3):73–81, 2004.

[9] Kelly Miller, Sacha Zyto, David Karger, Junehee Yoo, and Eric Mazur. Ana@articlerosenbaum1983central, title=The central role of the propensity score in observational studies for causal effects, author=Rosenbaum, Paul R and Rubin, Donald B, journal=Biometrika, volume=70, number=1, pages=41–55, year=1983, publisher=Oxford University Press lysis of student engagement in an online annotation system in the context of a flipped introductory physics class. *Physical Review Physics Education Research*, 12(2):020143, 2016.

[10] Saif M Mohammad and Svetlana Kiritchenko. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326, 2015.

[11] Stephen L Morgan and Jennifer J Todd. 6. a diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology*, 38(1):231–282, 2008.

[12] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[13] John A Ross. The reliability, validity, and utility of self-assessment. 2006.

[14] Cathy Sandeen. Integrating moocs into traditional higher education: The emerging “mooc 3.0” era. *Change: The magazine of higher learning*, 45(6):34–39, 2013.

[15] L. Wong. Causal inference in python. *[Online] Available: https://github.com/laurencium/causalinference*.

[16] Diyi Yang, Miaomiao Wen, Iris Howley, Robert Kraut, and Carolyn Rose. Exploring the effect of confusion in discussion forums of massive open online courses. In *Proceedings of the second (2015) ACM conference on learning@ scale*, pages 121–130, 2015.

[17] Amy X Zhang, Michele Igo, Marc Facciotti, and David Karger. Using student annotated hashtags and emojis to collect nuanced affective states. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, pages 319–322. ACM, 2017.

[18] Yuanda Zhu, Hang Wu, and May D Wang. Feature exploration and causal inference on mortality of epilepsy patients using insurance claims data. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 1–4. IEEE, 2019.