Do Supplemental Online Resources Improve Learning? Evidence From a Randomized Controlled Trial in an Undergraduate Quantitative Methods Course

Marco Alcocer[*] Leonardo Falabella[†] Alexandra Lange[‡] Nicholas Smith[§]

03/06/2021

1 Motivation and Research Questions

According to a 2017-18 survey by the American Political Science Association (APSA), 82.4% of U.S. political science departments offered at least one course in undergraduate research methods as part of their core curriculum.[1] Moreover, there are increasing calls to include methods training in undergraduate political science curricula by APSA working groups.[2]

The Department of Political Science at University of California, San Diego (UCSD) requires that all of its undergraduate majors take an introductory political methods course (POLI 30 Political Inquiry), which introduces students to fundamental tools of political inquiry, including research design, causal inference, and basic statistical methods. Yet, this course has the highest withdrawal and fail rate of any course in the department. As a group of PhD students who have taught discussion sections and assessed undergraduate work for this course, we have seen that many students struggle due to the fact that they enter the course with varying levels of preparation, course material is technical in nature and the course moves at a fast pace, with the introduction of new concepts relying on mastery of concepts previously introduced. Moreover, many students arrive in the course with preconceptions about the course's difficulty and lack confidence in their ability to master statistical concepts. These challenges are compounded by high student-to-TA ratios (68 students per TA) and a 10-week quarter system.

To address some of these challenges, we developed a series of supplemental online instructional modules to support student learning in the course. These online modules were designed to target concepts students have historically found most challenging in the course (rather than comprehensively cover all course material) and consist of short videos addressing these concepts, explanatory text, and practice quizzes designed to reinforce student understanding.

This study is designed to assess the impact of these online modules on student learning in the course. Specifically, we hope to answer the following questions: does the availability and use of our online supplementary educational modules improve student comprehension of course material? If so, to what extent? Moreover, using student demographic data, we hope to gain insight into which students benefit most from the availability

[*]University of California, San Diego. m2alcoce@ucsd.edu.
[†]University of California, San Diego. lfalabel@ucsd.edu.
[‡]University of California, San Diego. awoordru@ucsd.edu.
[§]University of California, San Diego. nms001@ucsd.edu.

[1]Davis, Megan, Erin McGrath, and Betsy Super. 2019. "2017-18 APSA Departmental Survey: Enrollments and Curriculums." American Political Science Association, May 21, https://www.apsanet.org/Portals/54/APSA%20Files/Data%20Reports/Enrollment%20Data/APSA%20Departmental%20Survey_Enrollment%20and%20Curriculum_FINAL.pdf?ver=2019-05-21-113745-243

[2]Smith, Steven Rathgeb and Meghan McConaughey. 2021. "Rethinking the Undergraduate Political Science Major: The Wahlke Report Revisited." *PS Political Science & Politics* 54 (2): 358-362.

and use of this content, and to what extent modules might address achievement, opportunity, and equity gaps among different demographics.

2 Pre-Analysis Plan

In this document we pre-register a set of primary analyses aimed at answering the research questions presented above. Specifically, we focus on an experiment designed to test the effect of the supplementary materials on student learning. We also wish to understand student attitudes and opinions about the supplementary material and course content.[3] Our pre-analysis plan specifies the hypotheses and research design of the experimental portion of the study.

In this pre-analysis plan we pre-register one main hypothesis and two secondary hypotheses. We commit to reporting all pre-registered tests in the paper, and any deviation from the pre-registered plan will be carefully itemized and explained. Depending on data quality, we also intend to conduct exploratory tests in order to explore heterogeneity. However, any supplementary test will be clearly stated in the paper as exploratory and not pre-registered so that the reader is fully aware of this.

3 Research Design

We employ a within-subject design. Students enrolled in POLI 30 Political Inquiry[4] during spring quarter 2021 were randomly assigned into six treatment groups. There were 216 total students enrolled in the course during the first week of the quarter, and treatment groups ranged in size from 32 to 39. Not every student will be a subject in our study because some students may drop the course or fail to take the midterm exam. Students are also allowed to opt out of the study at any time. At the time of writing the pre-analysis plan, we have not reviewed which students have dropped, not taken the exam, or opted-out of the study.

All treatment groups were given access to a Canvas page with supplemental teaching resources. Each treatment group had access to a total of four modules, two of which were common to all groups, and two of which varied by treatment group. The modules contain informational text, practice quizzes, and one or more instructional videos focused on a specific subset of course concepts introduced prior to the midterm exam. The content of the modules was designed in collaboration with POLI 30 instructors, previous graduate student teaching assistants for the course, and five undergraduate students who had previously taken the course. As mentioned above, modules focused on concepts that POLI 30 students have historically struggled with most.

Students were provided with access to their Canvas page during the first week of the course and were notified about materials both via email and in class. Students were also aware that the effectiveness of the resources was being tested, but were not aware of design details. Students were sent invitations to their Canvas page a few times in the weeks leading up to the midterm exam. It is important to note that students were encouraged, but not required, to use the supplementary materials, and no additional/extra credit points were awarded for their use.

Two modules were available for every treatment group: an introductory module explaining how to use the Canvas page and a module on research design, theories, and hypotheses. Each treatment group also had access to two additional modules selected from a set of four (Introduction to Variables, Confounding and Intervening Variables, Research Design, and Introduction to Inference). We chose this allocation of modules because the modules available to all students contain foundational knowledge about research design, and

[3] Analysis of student opinions about course materials will be based on observational survey data and will be exploratory and thus we do not include further discussion of it here.

[4] POLI 30 Political Inquiry is a mandatory, 10-week course on quantitative research design and statistics for the social sciences. Students were introduced to the study and the opt-out IRB form during week 1 of the quarter. Our sample consists of undergraduate students at UCSD who were enrolled in the course during Spring term of 2021, excluding those who opted out or did not take the midterm exam. All students who consented to participating in the experiment were assigned to treatment, including students that enrolled late. We will present two sets of results, including and excluding students that enrolled late. At the time of writing the pre-analysis plan, we have not reviewed which students have dropped, not taken the exam, or opted out of the study.

Table 1: Access to modules by treatment group.

	Treatment Group					
Module	**1**	**2**	**3**	**4**	**5**	**6**
Introduction	x	x	x	x	x	x
Research Questions, Theories, and Hypotheses	x	x	x	x	x	x
Introduction to Variables	x	x	x			
Confounding and Intervening Variables	x			x	x	
Research Design		x		x		x
Introduction to Inference			x		x	x

Table 2: Overlap between content in modules and questions.

	Q2			Q3		Q4				Q5			
Module	**b**	**c**	**d**	**a**	**b**	**a**	**b**	**c**	**d**	**a**	**b**	**c**	**d**
Introduction to Variables	x	x	x										
Confounding and Intervening Variables				x	x								
Research Design				x						x	x	x	x
Introduction to Inference						x	x	x	x				
Research Design		x		x		x							
Introduction to Inference			x		x	x							

therefore crucial to the comprehension of subsequent modules. Table 1 shows a complete description of access to modules by treatment group.

Five weeks into the 10-week quarter course, students took a midterm exam worth 20% of the course grade. The midterm exam (included in the Appendix) consisted of 5 questions, which included a total of 18 subquestions. In our study, we are interested in the performance of students on 4 questions, which included 13 subquestions: questions 2b-2d and questions 3-5. Question 1 asked about skills that the learning modules do not teach, and question 2a asked about content taught in the “Research Questions, Theories, and Hypotheses” module, which was available to all six treatment groups. Consequently, these questions were dropped from our study.

By limiting ourselves to the set of 13 subquestions in our study, we limit our analysis to questions that focused on content covered in the supplemental instructional materials that were available to some treatment groups and not others. Table 2 shows a complete description of how modules and questions overlap.

In this study, the unit of observation is student-subquestion. Treated units include those subquestions answered by students that had access to a relevant supplemental module. Untreated units include student answers to subquestions covered by supplemental instructional material *not* available to them. Thus, each student also acts as their own control group. For example, students in treatment group 3 had access to the “Introduction to Variables” and “Introduction to Inference” modules. Therefore, all answers by students in group 3 to subquestions 2d and 4a are treated units of observation, whereas all answers by students in group 3 to subquestions 3b and 5a are non-treated units of observation.

Importantly, as noted above, the contents of all modules were supplementary. The knowledge and skills necessary to answer each question were available through lectures by the instructor, discussion sessions with the TAs, and the course textbook. Our goal with this research design is to test if access to supplementary videos, instructional text, and quizzes improve learning outcomes.

The main advantage of our research design is that it reduces errors associated with differences in skill and effort between students. The within-subject design is equivalent to a block-randomized experiment in which each participant is a block. In our study, each student took a midterm exam in which they had access to treatment for some questions, and not for others. Holding constant the differences in skill and effort between students, as well as differences in the levels of difficulty between subquestions, we test if students perform better at questions to which they had access to treatment. This research design also has the added advantage of fairness: no particular module is "better" than another, and the groups of questions pertaining to each module were equally weighted in the exam.

4 Hypotheses

The main hypothesis is straightforward:

H1: The supplemental online materials will improve student learning.

In addition, we plan on evaluating exploratory hypotheses, depending on data availability, because there are reasons to believe that the treatment may have differential effects across students.

First, we believe that the treatment might have differential effects across students with different levels of academic achievement. On the one hand, it is reasonable to believe that our supplemental materials could be most useful for and have the greatest impact on the students who struggle the most academically. The materials are designed to introduce concepts intuitively and at a basic level. This could mean that students who usually struggle to comprehend the material are the students who most benefit from the treatment. On the other hand, it could also be the case that students who would have performed better without the treatment are also the students who are more likely to take advantage of our materials. If this is the case, it could be that these materials cause overperforming students to master the course concepts and perform even better than they would have without them, resulting in even larger achievement gaps between higher and lower performing students.

H2: The supplemental online materials will have a greater effect on students with lower academic performance than students with higher academic performance.

Second, we are interested in exploring whether these materials have differential effects on students from historically underrepresented groups. It could be the case that these students do not respond to the white faculty[5] like other students. For example, they might be less willing to attend office hours, email the professor, attend TA office hours, or engage and participate during class time, leading them to underperform. If this is the case, these online materials might provide a lower-barrier for these students, especially since two of the four video presenters are Latino and one is female. They also do not require direct contact with the professor or TA, making them more accessible to these students.

H3: The supplemental online materials will have a greater effect on minority students than white non-hispanic students.

5 Variables and Measurement

The dependent variable is student learning as measured by their standardized grades on individual questions and subquestions on the midterm exam. We designed each exam question to be similar in difficulty and include material covered by its corresponding module. We then created a detailed rubric for each question to assure consistent grading.

Each exam question was graded by a different Teaching Assistant for the course. In other words, each TA graded one question per exam, and graded the same question for all students. All TAs were PhD students in the Political Science Department and had previously graded for Political Science courses at UCSD. To ensure consistency, we trained each TA on the grading rubric. This also allows us to use exam question fixed effects to capture any inter-TA grading differences. TAs were not aware of treatment assignment when grading. They also had no access to the online modules.

The treatment will be a dummy variable denoting whether student i received module m. That is, each student has four treatment values, one per module, two of which are 0 and two are 1. After the course is completed, we will receive student-level information, including academic information such as GPA, year in college, and transfer status, as well as personal information including gender and race.

[5] All four faculty members who teach this course are white males.

6 Indentification Strategy

We estimate the effect of the instructional materials with an intent-to-treat analysis. The estimand is the intent-to-treat effect (ITT). We will use OLS linear regression as the estimator.

Since we also have data on compliance, we know which students accessed the modules to which they were assigned, which means we can also estimate the local average treatment effect (LATE).[6] The LATE is the average treatment effect for compliers. In this study this is especially useful because it is an encouragement design in which we expect non-compliance to be prevalent. We use an instrumental variables approach to estimate the LATE.

6.1 Regression Analysis

We use regression analysis to test the main hypothesis, H1. This allows us to incorporate covariates. The main model follows the following form:

$$Y_{iq} = \alpha + \beta Treatment_{iq} + \phi X_i + \gamma_i + \lambda_q + \epsilon_{iq}$$

Where q denotes each module/exam question and i denotes each student. In the model β is the causal coefficient of interest. Y_{iq} are the points for each exam question, $Treatment_{im}$ is the treatment indicator, X_i is a matrix with student-specific covariates, γ_i are student fixed effects and λ_q are module fixed effects. The exam question fixed effects should absorb any differences in grading across exam questions. The inclusion of student fixed effects makes this model a within-student design. Finally, because our treatment is assigned at the student level, we cluster the standard errors at the student level for all models.

Because we have outcome measures at the subquestion-student level, we also estimate the regression using subquestion-student level data. This estimates the effect of the treatment on more granular concepts. The regression model is:

$$Y_{is} = \alpha + \beta Treatment_{iq} + \phi X_i + \gamma_i + \lambda_s + \epsilon_{is}$$

Where s denotes each exam subquestion and i denotes each student. β is the causal coefficient of interest. Y_{is} are the points for each exam subquestion, $Treatment_{iq}$ is the treatment indicator, X_i is a matrix with student-specific covariates, γ_i are student fixed effects and λ_q are subquestion fixed effects. The exam subquestion fixed effects should absorb any differences in grading across exam questions as well as factors that affect each subquestion the same across students. The inclusion of student fixed effects makes this model a within student design.

6.1.1 Randomization Inference

We then calculate how likely it is that we would estimate a coefficient at least as large as the one we observe if the sharp null hypothesis of no effect for any unit were true. Since our main hypothesis follows the form H1: H0 $= 0$ and HA ≤ 0, we conduct one-tailed tests for H1 using randomization inference with equation 1 and 2. We run 5,000 permutations[7] of possible treatment assignments with each student being assigned to two control and two treatment groups, and save the treatment coefficient, β_p. We then calculate how likely it is that we observe the main coefficient of interest given the distribution of β_p. We select a significance level of 0.05. This estimation strategy tests the sharp null of no treatment effect for any unit, $H0_s : \tau_i = Y_i(0) - Y_i(1) = 0$.

6.2 Instrumental Variable

To estimate the LATE for H1, we use an instrumental variables (IV) approach, where we regress the outcome on the treatment, using treatment assignment as an instrument. The first stage is then:

[6] Also known as the complier average causal effect (CACE).
[7] If 5,000 permutations is too computationally burdensome we will decrease the number of permutations.

$$Treated_{im} = \alpha_1 + \beta_1 Treatment_{im} + \phi_1 X_i + \gamma_i + \lambda_m + \epsilon_{im}$$

Where $Treatment_{im}$ denotes treatment assignment and $Treated_{im}$ denotes whether a student received the treatment.

In the second stage we regressing the outcome on the fitted values from the first stage:

$$Y_{im} = \alpha_2 + \beta_2 \hat{T}reated_{im} + \phi_2 X_i + \gamma_i + \lambda_m + e_{im}$$

The coefficient of interest is β_2, which estimates the LATE.

6.3 Exploratory Hypotheses

We use linear regression to test the exploratory hypotheses, H2-H3:

$$Y_{im} = \alpha + \beta Treatment_{im} \times StudentChar_i + \phi X_i + \gamma_i + \lambda_m + \epsilon_{im}$$

Where $StudentChar_i$ denotes the student-specific characteristic that we are interested in (gender, race, GPA). The interaction between this variable and the treatment assignment will allow us to estimate the heterogeneous effects of interest.

6.4 Standard Errors

Because our unit of observation is student-question, but the treatment is randomized at the student level, we cluster standard errors at the student-level for all regression analyses. This includes models 1-5.

6.5 Significance Test

Since our main hypothesis follows the form H1: H0 $= 0$ and HA ≤ 0, we conduct one-tailed tests for H1. For the exploratory hypotheses we conduct two-tailed tests.

6.6 Missing Data

Missing data on the dependent variable will be left as is, and as such respondents who have missing data are excluded from the analysis. If covariate values are missing, we will recode the missing values to the overall mean.[8]

6.7 Further Exploratory Analysis

Depending on the composition of the data on student characteristics, we will want to conduct further exploratory analyses on the effects of the intervention on specific populations (e.g., gender, transfer status). However, since we do not know which, or whether these will be possible, we only pre-register the main primary hypothesis. Nevertheless, any analysis beyond the main hypothesis that we conduct and include in the paper will be clearly stated as being exploratory and not pre-registered.

[8] Lin, Winston, Donald P. Green, and Alexander Coppock. 2016. "Standard operating procedures for Don Green's lab at Columbia." http://alexandercoppock.com/Green-Lab-SOP/Green_Lab_SOP.pdf.

6.8 Treatment Overlap and Intensity

One consideration is that the learning modules were designed with student learning in mind, not an experiment. This means that the modules sometimes build on each other, quickly recap concepts from other modules, or refer to content from other modules. For the rollout of this experiment, we tried to limit the amount of content overlap by not publishing certain modules or editing some module content so that each module covered different materials.

However, there are two concepts that are covered by more than one module and that are tested on the exam:

1. *Dependent and independent variables* are quickly covered by a module all students received, thoroughly covered by the module on variables that some students were assigned, and discussed in the confounding and intervening variables module some students were assigned to.
2. *Confounding variables* are thoroughly covered by the confounding and intervening variables module some students were assigned to, and quickly recapped in the research design module.

This means that some students were exposed to stronger treatments than others for these two concepts.

The exam questions and the parts covered by each module are included in the Appendix. To account for this overlap, we take three different approaches.

1. Run models as is without accounting for this overlap.
2. Exclude the parts of the questions that are covered by more than one module (Question 2 part A, and Question 3 part A), reweight the other parts of the exam questions to total 20 points, and re-run models. This will give us the treatment effect on outcomes that are only covered by one module.
3. Rather than having a dummy variable for the treatment, have a numerical variable ranging from 0-3 denoting the number of modules a student is exposed to per exam question. This would tell us the degree of the effect of being assigned to more than one module per concept.
   a. For the question on dependent and independent variables (Question 2 part A,), students that were not assigned the confounding and intervening variables module or the introduction to variables module would get assigned a 1. For students assigned one of these modules, they would get assigned a 2. For students that were assigned both of these modules, they would be assigned a 3.
   b. For the question on intervening variables (Question 3 part A), students not assigned the confounding and intervening variables module or the research design module would get assigned a 0, students assigned one of these modules would be assigned a 1, and students assigned both modules would be assigned a 2.

7 Compliance

We used an encouragement design: students were invited to engage with the supplementary online materials but were not required to. We anticipate that some students may have chosen not to interact with the modules they were assigned. It is also possible that students shared instructional materials with fellow classmates.

We will measure compliance in two ways:

- One-sided noncompliance can be evaluated using Canvas activity reports. Canvas provides information such as the number of times students visited a page, watched a video, and completed a practice quiz.
- The possibility of two-sided noncompliance will be evaluated using a post-outcome survey in which students are asked whether they interacted with modules to which they were not assigned.

8 Appendix

8.1 Midterm Exam

1. The star of the UCSD basketball team has played well lately. In the last five games, she has point totals of:

22, 30, 18, 16, 24

Answer the following questions. Show your work in order to get full credit.

   a. What type of variable is "point total," as described above: nominal, ordinal, or interval/continuous? Why?
   b. What is the median point total (as a number)?
   c. What is the mean point total (as a number)?
   d. What is the variance of this sample of point totals (as a number)?

2. Consider the following hypothesis: "Candidates tend to spend more money in close elections."

   a. What is the dependent variable? What is the independent variable?
   b. What is the unit of analysis?
   c. How would you measure the closeness of elections? Justify why you believe this is a good measure.
   d. What type of variable would your measurement strategy produce (e.g. ordinal, nominal, interval, or ratio?). Explain why, using the definition of this type of variable.

3. Consider the following hypothesis: "Famines are less likely to happen in countries where there is press freedom."

   a. Can you think of a confounding variable in this hypothesis? Please justify your answer. In your justification, be sure to demonstrate that you understand the concept of confounding variables by identifying the criteria of confounding variables and explaining how your confounding variable meets each criterion.
   b. Can you think of an intervening variable in this hypothesis? Please justify your answer. In your justification, be sure to demonstrate that you understand the concept of intervening variables by identifying the criteria of intervening variables and explaining how your intervening variable meets each criterion.

4. Suppose we want to study whether UCSD students approve of the cost of parking on campus. We set up a table by the entrance of two parking structures on campus and ask students who are walking in what they think about the cost of parking.

   a. What is the population in this study?
   b. Is this a probability or a non-probability sample? Please justify your answer. In your justification, be sure to demonstrate that you know the difference between probability and non-probability samples.
   c. When can we make inferences about populations from samples?
   d. Could we make inferences about UCSD student approval of parking costs using the sampling procedure described in the prompt? Why or why not?

5. Suppose you want to answer the following research question: "Do people become more receptive to being vaccinated if they watch a scientist explain how vaccines work?"

   a. Describe an experiment that would enable us to answer this research question (3-4 sentences).
   b. In this experiment, what would be the treatment?
   c. What would be the treatment group? And the control group?
   d. How would this experiment mitigate our concerns about confounds?