De La Cruz, Marjorie; Abi-Hassan, Sahar; Denly, Michael

Working Paper

A Comment on "The PhD Pipeline Initiative Works: Evidence from a Randomized Intervention to Help Underrepresented Students Prepare for PhDs in Political Science"

I4R Discussion Paper Series, No. 152

Provided in Cooperation with:
The Institute for Replication (I4R)

INSTITUTE for REPLICATION

No. 152
I4R DISCUSSION PAPER SERIES

A Comment on "The PhD Pipeline Initiative Works: Evidence from a Randomized Intervention to Help Underrepresented Students Prepare for PhDs in Political Science"

Marjorie De La Cruz

Sahar Abi-Hassan

Michael Denly

September 2024

I4R DISCUSSION PAPER SERIES

I4R DP No. 152

A Comment on "The PhD Pipeline Initiative Works: Evidence from a Randomized Intervention to Help Underrepresented Students Prepare for PhDs in Political Science"

Marjorie De La Cruz[1], Sahar Abi-Hassan[2], Michael Denly[3]

[1]Florida International University, Miami/USA
[2]Northeastern University, Boston/USA
[3]Texas A&M University, College Station/USA

SEPTEMBER 2024

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

A Comment on "The PhD Pipeline Initiative Works: Evidence from a Randomized Intervention to Help Underrepresented Students Prepare for PhDs in Political Science"*

Marjorie De La Cruz† Sahar Abi-Hassan‡ Michael Denly§

July 20, 2024

Abstract

We provide a reproduction and replication of Brutger (2024), which examines the effects of the University of California, Berkeley's Pipeline Initiative in Political Science (PIPS) program on five self-reported outcomes related to interest and preparation towards pursuing graduate school. We are able to reproduce the author's results but do note some minor coding challenges. Our additional replication analysis confirms that the study's original results are robust to different model specifications. In future analysis of PIPS, we suggest that the author address our suggestions regarding the wording of the survey questions, sample selection, and statistical power. Overall, we commend the author on a good study of an important topic.

KEYWORDS: Diversity, Equity, Inclusion, DEI, Political Science, Graduate School, Program Evaluation, Randomized Controlled Trial.

JEL CODES: D63, M14, I24, I23.

*The authors declare no conflicts of interest. We thank Ryan Brutger and Derek Mikola for comments that helped improve the paper. Our errors are solely our responsibility.

†Florida International University: mdela168@fiu.edu

‡Corresponding author, Northeastern University: s.abi-hassan@northeastern.edu

§Texas A&M University: mdenly@tamu.edu

1 Study, Reproduction, and Replication Overview

The following reproduction and replication concerns Brutger (2024). It investigates the effects of a 2021-2022 program at the University of California, Berkeley, called the Pipeline Initiative in Political Science (PIPS). The latter is one of many academic pipeline programs that provide educational opportunities, guidance, and resources to qualified underrepresented and marginalized students (Byrd and Mason 2021). PIPS, in particular, is a one-semester program that aims to increase diversity, equity, and inclusion (DEI) in the student population of political science graduate programs, and in the long run, potentially diversify faculty and research outputs.

Brutger (2024) evaluates intermediate outcomes of PIPS using five student self-reported outcomes related to their interest and preparation towards pursuing graduate school. PIPS represents the first evaluation of a political science pipeline intervention involving a randomized controlled trial (RCT), which many scholars consider to be the gold-standard in program evaluation (e.g., Imbens and Rubin 2015).

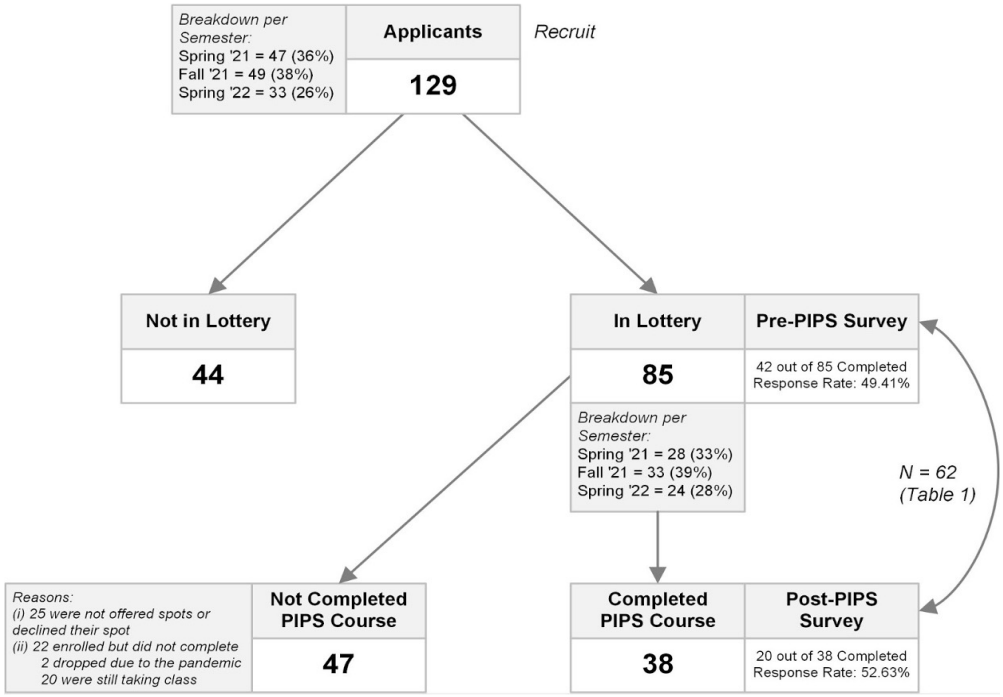
Brutger (2024) finds that PIPS increases graduate school preparation but not interest in applying for a PhD program. We are able to reproduce the same results as the author, although we do note some minor coding challenges. The author's results are robust to different model specifications that we introduce in our replication analysis. In future analysis of PIPS, we suggest that the author address our suggestions regarding the wording of the survey questions, sample selection, and statistical power. Overall, we commend the author on a good study of an important topic.

1.1 Program Eligibility and Selection

To be eligible for admission into the PIPS program for the time period examined, students needed to meet a number of criteria, including: 1) at least being in the second year of their undergraduate studies; 2) having an interest in learning more

about pursuing a PhD; 3) be a first-generation college student from a i) historically minority group, ii) an underrepresented group, or iii) a low-income background; and 4) meet a 3.5 Grade Point Average (GPA) cutoff.[1] To recruit such a diverse body of applicants, program organizers promoted PIPS through e-mails to students majoring in social sciences and reached out to groups of undergraduates in various programs, such as Berkeley's Undergraduate Researchers of Color.

Figure 1: Original Experimental Design



Notes: We created the above figure on the basis of Brutger (2024, Supplementary Appendix, Tables 1-2 and Figure 1). PIPS refers to "Pipeline Initiative in Political Science." All applicants in the lottery were eligible to complete the Pre-PIPS survey, while all students who completed the PIPS course were eligible to complete the post-PIPS survey. Brutger (2024) notes that the number of applicants declined in Spring 2022 most probably due to Fconflicts in the students' schedules. The author hopes to overcome this flaw in recruiting applicants in future cohorts. Two students dropped the PIPS course due to extenuating circumstances during the COVID-19 pandemic. Only students in Spring and Fall of 2021 completed the course at the time of writing. Because students from Spring 2022 were still taking the PIPS course, they were included in the "Not Completed" category. Spots per semester are limited to 20 seats.

Students apply to PIPS through a brief online application. Given that the program only offers 20 seats per semester, to offer a better learning environment, students who

[1]The GPA cutoff was later abandoned to include motivated students with lower GPAs, instead asking the students to explain additional contexts that would help in evaluating their application and assessing diversity. However, all data in the paper come from when the 3.5 GPA cutoff was in place, so requirements did not change across the cohorts in the analysis.

make it past the application process are entered into a lottery. The idea behind the lottery is that it provides all admitted applicants an equal probability of being selected into the program. Before starting the program, each student in the lottery is asked to complete an anonymous survey that Brutger (2024) calls the "pre-PIPS survey". The students who earned a spot through the lottery and completed the program were later provided with another survey called the "post-PIPS survey". Figure 1 provides a full explanation of the recruitment and selection process.[2]

1.2 Variable Operationalization and Methods

Brutger (2024) uses observations from both the Pre- and Post-PIPS surveys to create the main explanatory dummy or dichotomous variable, `PIPS`. This variable compares responses from students who completed PIPS (coded as 1) and students who passed the admissions criteria and were selected in the lottery to participate in PIPS, but were not enrolled into the program (coded as 0).[3] Thereafter, Brutger (2024) tests the effects of the PIPS through self-reported outcomes using ordinary least squares (OLS) and describes the main results on pages 385-386.

As shown in Table 1, the five self-reported outcomes of interest gauging student opinions are: 1) Interest in a PhD program; 2) overall preparation to apply to graduate school; 3) preparation for the personal statement (PS); 4) preparation for the statement of purpose (SOP); and 5) preparation for seeking strong letters of recommendation (LORs).[4] The survey question for the first outcome, PhD Interest, asks "How likely are you to pursue a PhD in political science or other field?" and reports whether

[2]PIPS was launched at a large public institution with a PhD program (i.e., UC Berkeley), which made it easier for graduate student engagement. According to the author, other pipeline programs are resource intensive and difficult to scale up, whereas PIPS is a relatively low-cost program to the department. From the student's perspective, PIPS is a great opportunity, as it allows for course credit and academic progress.

[3]There were two free-response questions. The first in the pre-PIPS survey asked, "What would you most want to learn from PIPS?" The second one in the post-PIPS survey asked, "What were the most valuable lessons or takeaways from participating in PIPS?"

[4]Brutger (2024) wishes to expand data collection efforts to assess the effects of PIPS on long-term outcomes, such as admission to graduate school.

Table 1: Summary Statistics

	No. Obs	Mean	SD	Min	Max
Enrolled in PIPS	62	0.323	0.471	0	1
PhD Interest	62	0.903	0.298	0	1
Prepared to Apply	58	0.552	0.502	0	1
Prepared PS	58	0.638	0.485	0	1
Prepared SOP	58	0.379	0.489	0	1
Prepared LORs	58	0.552	0.502	0	1
Four-Category Outcome Variables					
PhD Interest	62	3.274	0.682	1	4
Prepared to Apply	58	2.517	0.941	1	4
Prepared PS	58	2.690	0.977	1	4
Prepared SOP	58	2.293	0.973	1	4
Prepared LORs	58	2.552	1.079	1	4
Demographic Control Variables					
Male	62	0.355	0.482	0	1
female	62	0.500	0.504	0	1
Non-Binary	62	0.048	0.216	0	1
Native American	62	0.016	0.127	0	1
Asian	62	0.210	0.410	0	1
Black	62	0.065	0.248	0	1
Hispanic	62	0.403	0.495	0	1
Middle Eastern	62	0.129	0.338	0	1
Islander	62	0	0	0	0
White	62	0.258	0.441	0	1
First Generation	57	0.667	0.476	0	1

the respondents are likely to apply to doctoral programs, on scale of with outcomes 'very likely', 'somewhat likely', 'somewhat unlikely', and 'very unlikely'. The other four dependant variables concern preparation for graduate school applications, with the original question as following: "If you choose to apply to graduate school, how prepared are you to: [(2) complete the application, (3) write a personal or diversity statement, (4) write a research statement, (5) have strong letters of recommendation."

Although the survey design for the five dependant variables is on a scale, these measures were dichotomized for the OLS regressions in the main Table 2. For the PhD Interest outcomes the author combines the survey answers of 'somewhat likely' and 'very likely' under 1 (interested) and the answers of 'somewhat unlikely' and 'very unlikely' under 0 (not interested). For the other four outcomes of interest, the author combines the survey answers of 'somewhat prepared' and 'very prepared' under 1 (pre-

pared) and the answers of 'not very prepared' and 'not prepared at all' under 0 (not prepared). Table 1 reports summary statistics for these five dependent variables – in their binary and ordered scale form – and all demographic control variables.

1.3 Study Results

The results from Brutger (2024, Table 1, 386) suggest that the PIPS program led to a 48.9% increase in the number of students who were prepared to apply to graduate school, potentially increasing their chances of receiving admission. To justify this quantitative result, the author suggests that qualitative analyses of students' materials showed considerable improvement over time. For example, at the beginning of the program, personal statements tended to have many weaknesses. By the end of the program, though, the final versions of the materials exhibited considerable improvements, reflecting better academic writing, clearer goals, and an increased sense of belonging in academia.[5]

In Table 3 in the the Appendix, Brutger (2024, 8-10) examines the results with the four-category dependent variable, where 1 = very unlikely/not prepared at all; 2 = somewhat unlikely/not very prepared; 3 = somewhat likely/somewhat prepared; and 4 = very likely/very prepared. While the results from Brutger (2024, Table 1, 386) using the dichotomous outcome suggest that the program does not significantly increase interest in PhD in Table 1, results with the four-category outcome in Appendix Table 3 more clearly show that the program has no effect on the likelihood of applying to graduate school. Nevertheless, all other outcomes are consistent with Brutger (2024, Table 1, 386), showing that PIPS effectively improves students' preparation towards

[5]The actual PIPS 2021 program consisted of seven sessions. The fourth session discussed personal statements, giving the assignment of drafting one and reviewing other students' statements. These two components were required to pass the class. The fifth session gave feedback and advice on the statements, assigning the opportunity to revise the personal statement. Session 6 of the course introduced preparing writing samples and the Statement of Purpose, also known as the Research Statement, giving this as an optional assignment. The final section discussed application materials and the next steps after the PIPS program.

graduate school. Appendix Table 4 further shows that the results are robust when controlling for demographics. Finally, Appendix Table 5 narrows down observations to those students who enrolled in PIPS, comparing their Pre-PIPS and Post-PIPS surveys—i.e., thereby excluding other respondents who were in the lottery but did not enroll in the program. The results in Appendix Table 5 are consistent with those of the main Table 1 (Brutger 2024, 386), and the estimates are even larger, pointing to the potential benefits of the program.[6]

2 Computational Reproducibility

Replication files, including the main `.csv` file with the data, the `.rtf` codebook, and R code script are available in the *Journal of Politics* (JOP) Dataverse at the following link: https://doi.org/10.7910/DVN/UPLZAK. A JOP replication analyst previously replicated the author's results successfully.

Table 2: The Author's Main Results (Table 1 in the Paper)

	(1) PhD Interest	(2) Prepared to Apply	(3) Prepared Personal Statement	(4) Prepared SOP	(5) Prepared LORs
PIPS	-0.005	0.489***	0.525***	0.417***	0.408***
	(0.082)	(0.128)	(0.120)	(0.129)	(0.133)
	[0.954]	[0.000]	[0.000]	[0.002]	[0.003]
(Intercept)	0.905	0.400	0.475	0.250	0.425
	(0.046)	(0.071)	(0.067)	(0.072)	(0.074)
	[0.000]	[0.000]	[0.000]	[0.001]	[0.000]
No. Obs	62	58	58	58	58

Notes: OLS model. Standard errors are in parentheses. P-values are in brackets. Significant at the ***[1%] **[5%] *[10%] level.

While we are able to reproduce the same results as the author, we are unable to output the same tables from R's `stargazer` package (Hlavac 2022). The author's

[6]See reproduction of Brutger (2024) Appendix results in Appendices C, D & E of this manuscript.

code threw an error when attempting to generate Table 1 in Brutger (2024). We attempted several fixes, such as removing nay NA's from the original, inspecting the structure of the models, as well as ensuring consistency across the models. We were able to generate tables for each model separately, but not together as provided in the replication code. To address this issue, we used instead the `modelsummary` package from Arel-Bundock (2022) to reproduce Table 1 from the original paper (see Table 2, above). We reproduced the remainder of the tables in the paper, Tables 3-5 in the Appendix, using Stata (See results in the Appendix). Table A7 in the Appendix provides a full overview of the article's computational reproducibility.

2.1 Pre-Analysis Plan

The author did not register a pre-analysis plan. However, we wrote a pre-analysis plan for this replication paper.[7]

3 Robustness Reproduction and Replication

In this replication, we undertake both computational and robustness reproducibility, using the author's analysis data. As part of robustness reproducibility, we provide balance checks, alternative model specifications that are consistent with the study's data-generating process and the analysis data provided by the author, test for heterogeneous treatment effects, power analysis, and a discussion of sample selection and external validity.

3.1 Balance Checks

Table 1 provides full summary statistics for the variables used in the study which can be found in the Dataverse `.csv` file. Pairwise t-tests in Table 3 show that control/Pre-PIPS (n = 42) and treatment/Post-PIPS (n = 20) groups do not vary significantly according

[7] It is accessible at OSF. We do not deviate from that pre-analysis plan.

Table 3: Balance Check of Treatment and Control Groups in the Surveys

	(1) 0 = Pre-PIPS		(2) 1 = Post-PIPS		(1)-(2) Pairwise t-test	
Variable	No. Obs	Mean(SE)	No. Obs	Mean(SE)	No. Obs	Mean difference
Male	42	0.310	20	0.450	62	-0.140
		(0.072)		(0.114)		
Female	42	0.548	20	0.400	62	0.148
		(0.078)		(0.112)		
Non-binary	42	0.048	20	0.050	62	-0.002
		(0.033)		(0.050)		
Native American	42	0.000	20	0.050	62	-0.050
		(0.000)		(0.050)		
Asian	42	0.238	20	0.150	62	0.088
		(0.067)		(0.082)		
Black	42	0.095	20	0.000	62	0.095
		(0.046)		(0.000)		
Hispanic	42	0.405	20	0.400	62	0.005
		(0.077)		(0.112)		
Middle Eastern	42	0.119	20	0.150	62	-0.031
		(0.051)		(0.082)		
White	42	0.214	20	0.350	62	-0.136
		(0.064)		(0.109)		
First-Generation	39	0.641	18	0.722	57	-0.081
		(0.078)		(0.109)		

Notes: Standard errors are in parentheses. Significant at the ***[1%] **[5%] *[10%] level. All mean differences lack stars, as none are statistically significant at the above levels.

to the control variables. We also test some of the additional demographics included in the Dataverse file and find no significant differences, suggesting no randomization failure (see Table 10).

Table 4 further breaks down Pre- and Post-PIPS survey responses by demographic characteristics. On the one hand, the program is successful in recruiting first-generation students, who represent more than half of the observations. On the other hand, the current treatment group does not include any black respondents due to the lottery setup, although the Pre-PIPS survey includes four black students. The other demographics are more evenly distributed, which speaks to the alignment of the results to the author's original intentions as well as the benefits of randomization.

Table 4: Distribution of Demographic Characteristics in Treatment and Control Groups in the Surveys

Variable	Pre-PIPS	Post-PIPS	Total
First-Generation	25	13	38
Female	23	8	31
Male	13	9	22
Other Gender/Unspecified	7	3	10
White	9	7	16
Asian	10	3	13
Black	4	0	4
Hispanic	17	8	25
Middle Eastern	5	3	8
Native American	0	1	1
Other Race/Unspecified	6	6	12

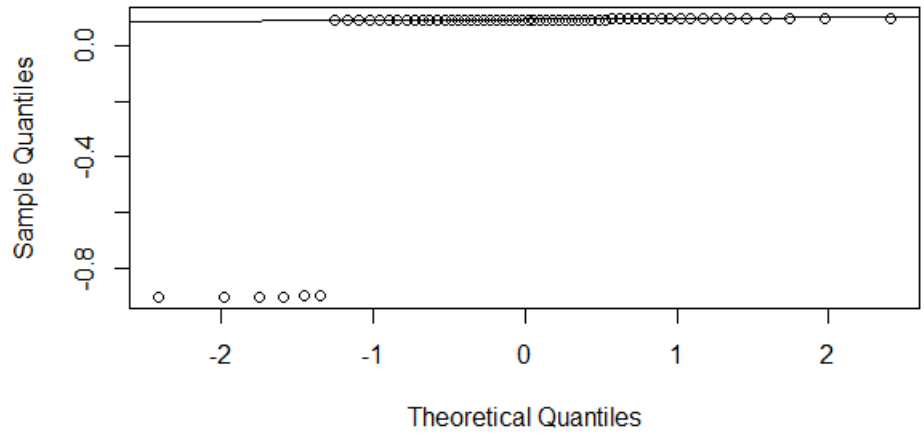
Note: Some students belong to more than one racial category.

3.2 Alternative Specifications: Binary and Ordered Logistic Regressions

The literature suggests that using OLS is generally the best method for randomized experiments with binary dependent variables (Gomila 2021), but we introduce robustness tests with logistic and ordered logistic regressions for a few reasons. First, the original version of the dependent variable obtained from the PIPS survey has four categories (see Table 1). Second, while the author uses a dichotomized version of that 4-category dependent variable in the main analysis, the quantile-quantile (QQ) plot in Figure 2 suggests that the OLS residuals in Table 1, Column 1 of Brutger (2024) are not normally distributed, and normality of residuals is essential for OLS. Third, in such cases with non-normality of residuals, it is generally preferable to use logistic regression given that it does not require normality of residuals. Although the quantities produced by logistic and ordered logistic regressions output are different than the more interpretable OLS outputs, they provide another window into the robustness of the results.

The results of the logistic regressions without covariates in Table 5 and the models with covariates in Table 6 are consistent with the author's original findings, with a few

Figure 2: Q-Q Plot of the Residuals from Table 1, Column 1 in Brutger (2024)



caveats. The model testing for Preparedness to write a Personal Statement (Model 3) does not generate an effect due to perfect collinearity. In addition, we observe a decrease in the statistical significance of the models testing the outcomes of Preparedness to Apply (Model 2), Preparedness to write a Statement of Purpose (Model 4), and Preparedness to obtain Letters of Recommendation (Model 5). This remains the case when we control for the variables `Male`, `White`, and `First Generation`.[8] Similarly to the original study, all control variables are not statistically significant at any level, suggesting that the randomization worked.

Table 5: Logistic Regression of the Effects of PIPS on Interest and Seld-assessed Preparation for Graduate School

	(1) PhD Interest	(2) Prepared to Apply	(3) Prepared Personal Statement	(4) Prepared SOP	(5) Prepared LORs
PIPS	-0.054	2.485**	0.000	1.792**	1.912**
	(0.912)	(0.816)	(.)	(0.619)	(0.708)
	[0.953]	[0.002]	[.]	[0.004]	[0.007]
No. Obs	62	58	40	58	58

Notes: OLS model. Standard errors are in parentheses. P-values are in brackets. Significant at the ***[1%] **[5%] *[10%] level.

The survey design for the questions capturing students' self-reported outcomes of

[8] Brutger (2024) does something similar in the Appendix Models.

Table 6: Logistic Regression of the Effects of PIPS on Interest and Preparation for Graduate School with Controls

	(1) PhD Interest	(2) Prepared to Apply	(3) Prepared Personal Statement	(4) Prepared SOP	(5) Prepared LORs
PIPS	-0.792	2.526**	0.000	1.724**	1.816*
	(1.016)	(0.844)	(.)	(0.638)	(0.719)
	[0.436]	[0.003]	[.]	[0.007]	[0.012]
Male	0.000	0.437	0.811	-0.315	0.284
	(.)	(0.662)	(0.755)	(0.653)	(0.624)
	[.]	[0.509]	[0.283]	[0.630]	[0.649]
White	0.047	-0.118	1.535	0.552	0.055
	(1.239)	(0.718)	(0.899)	(0.678)	(0.678)
	[0.970]	[0.870]	[0.088]	[0.415]	[0.935]
First Generation	0.071	-0.824	-0.614	0.467	0.053
	(1.015)	(0.667)	(0.739)	(0.659)	(0.622)
	[0.944]	[0.217]	[0.406]	[0.479]	[0.932]
No. Obs	35	57	39	57	57

Notes: Standard errors are in parentheses. P-values are in brackets. Significant at the ***[1%] **[5%] *[10%] level.

the program is organized in four categories going from 'very unlikely' to 'very likely', and 'not prepared at all' to 'very prepared'. In this light, as well as the aforementioned non-normality of residuals, we also reproduce the analysis using an ordered logistic regression. The results in Table 7 and 8 are consistent with the study's original findings. The first model testing interest is applying for a PhD in political science or another field lacks statistical significance. However, we see an improvement in the results in comparison to the logistic regressions presented in Table 5 & Table 6. The effect of PIPS on preparedness to write a personal statement is significant at the 1% level (Model 3 in Table 7). The remainder of the effects in Models (1), (2), (4), and (5) in Table 7 exhibit similar statistical significance as the logit model. This also remains the case when we control for demographics in Table 8. Finally, we perform a Brant test on all ordered logistic regressions reported in Tables 7 and 8 and find that the proportional odds assumption is not violated in any of the models, suggesting that the ordered logit

models contain valuable information.

Table 7: Ordered Logistic Regression for the Effects of PIPS on Interest and Self-assessed Preparation for Graduate School

	(1) PhD Interest	(2) Prepared to Apply	(3) Prepared Personal Statement	(4) Prepared SOP	(5) Prepared LORs
PIPS	0.062	2.716***	3.548***	1.760**	1.599**
	(0.525)	(0.721)	(0.844)	(0.562)	(0.0549)
	[0.906]	[0.000]	[0.000]	[0.002]	[0.004]
No. Obs	62	58	58	58	58

Notes: Standard errors are in parentheses. P-values are in brackets. Significant at the ***[1%] **[5%] *[10%] level.

Table 8: Ordered Logistic Regression for the Effects of PIPS with Controls

	(1) PhD Interest	(2) Prepared to Apply	(3) Prepared Personal Statement	(4) Prepared SOP	(5) Prepared LORs
PIPS	-0.235	2.684***	3.449***	1.720**	1.539**
	(0.574)	(0.729)	(0.852)	(0.573)	(0.564)
	[0.683]	[0.000]	[0.000]	[0.003]	[0.006]
Male	0.647	0.238	0.226	0.495	0.469
	(0.558)	(0.531)	(0.543)	(0.512)	(0.551)
	[0.247]	[0.654]	[0.677]	[0.334]	[0.394]
White	0.203	-0.186	0.724	0.050	-0.260
	(0.603)	(0.589)	(0.601)	(0.554)	(0.568)
	[0.737]	[0.752]	[0.228]	[0.928]	[0.647]
First Generation	0.355	0.115	0.069	0.446	-0.196
	(0.565)	(0.535)	(0.545)	(0.519)	(0.540)
	[0.530]	[0.830]	[0.899]	[0.391]	[0.717]
No. Obs	57	57	57	57	57

Notes: Standard errors are in parentheses. P-values are in brackets. Significant at the ***[1%] **[5%] *[10%] level.

The above analysis demonstrates that while ordinary least squares (OLS) regression is a widely accepted method for analyzing randomized experiments with binary dependent variables, our logistic and ordered logistic regressions provide valuable robustness checks, particularly in the presence of non-normal residuals. The consistent findings across logistic and ordered logistic models, despite some minor variations in

statistical significance, underscore the reliability of the original study's results. Therefore, these robustness checks not only validate the initial OLS findings but also enrich the interpretation of the results, ensuring that the conclusions drawn are both robust and reflective of the study's design.

3.3 Heterogeneous Treatment Effects

Due to the inability of randomizing the semester in which students join the PIPS program, we analyze whether the semester when a students joins the PIPS impacts the results. In a typical observational study, we would have clustered the standard errors by semester. However, because the study has unit-level random assignment, clustered standard errors are inappropriate (Abadie et al. 2023). Accordingly, we test for the impact of the `Semester` by interacting it with the binary treatment on enrollment in PIPS (`PIPS`). As Table 9 shows, the coefficient of `PIPS` $\times$ `Semester` does not show statistically significant results, suggesting that the cohort does not affect the results. Per Brambor et al. (2006), we do not interpret the constitutive terms.

Although the results here seem to confirm those of the author, it is worth noting the difficult-to-follow coding of the `Semester` variable.[9] On that score, it is not a binary variable capturing whether the students participated in the program in one semester or another. Instead, Brutger (2024) codes `Semester` as 0 for students in the control group who were not selected, not distinguishing by semester; 1 for the students who participated during the fall 2021 semester; and 2 for the students who participated in the spring 2022 semester. Thus, the author's coding of the semester variable prevents very precise analysis of heterogeneous treatment effects.

[9] Initially, the author refers to the variable as `Enroll_Sem`.

Table 9: Heterogeneous Treatment Effects by Semester

	(1) PhD Interest	(2) Prepared to Apply	(3) Prepared Pesonal Statement	(4) Prepared SOP	(5) Prepared LORs
(Intercept)	0.880***	0.435***	0.522***	0.261***	0.609***
	(0.059)	(0.094)	(0.089)	(0.095)	(0.091)
	[0.000]	[0.000]	[0.000]	[0.008]	[0.000]
PIPS	-0.380	0.065	0.478	-0.061	-0.009
	(0.234)	(0.363)	(0.343)	(0.364)	(0.351)
	[0.109]	[0.858]	[0.169]	[0.868]	[0.980]
Semester	0.031	-0.041	-0.055	-0.013	-0.216***
	(0.047)	(0.072)	(0.068)	(0.072)	(0.070)
	[0.513]	[0.574]	[0.424]	[0.861]	[0.003]
PIPS × Semester	0.219	0.291	0.055	0.313	0.366
	(0.143)	(0.226)	(0.214)	(0.227)	(0.219)
	[0.130]	[0.204]	[0.798]	[0.174]	[0.101]
No. Obs	62	58	58	58	58

Notes: OLS model. Standard errors are in parentheses. P-values are in brackets. Significant at the ***[1%] **[5%] *[10%] level.

3.4 Power Analysis

Given that the sample has at most 62 observations, we assess the statistical power of the tests performed. To do so, we take into account that that the treatment group has 20 observations and the control group has 42 observations. Then, we calculate the statistical power of the specification with the most observations (Table 1, Column 1) using a two-sided t-test with unequal group sizes and 0.05 as the significance level.[10] The results using the uncorrected Cohen's d based on the article results suggest that Brutger (2024) has 0.0503 chance of detecting a true effect if one exists. With Hedges' correction for the small sample size, the results do not improve and are the same to the fourth decimal point. To give Brutger (2024) a further chance at obtaining statistical power, we use the upper and lower bounds of the Cohen's d confidence intervals $[-0.5209786, 0.5522710]$. For the lower bound the power is 0.47, and for the upper bound it is 0.51. Given that the typical benchmark for power in experimental

[10] Appendix 8 provides the statistics that we use for the calculation.

designs is 0.8, it is clear that the author's conclusions about the program working exceed what the sample size in the data allow him to infer.[11]

4 Sample Selection and External Validity

We note limitations with the PIPS program selection process that introduces sample selection challenges and external validity questions. As Figure 1 details, Brutger (2024) excludes the 44 students from analysis who applied for the program but were not included in the lottery after the admissions process. Of course, experimental designs aim to compare treatment with control, and those 44 students were not part of the lottery. Plus, our balance test suggests no randomization failure (see Table 3), so we are not concerned with the study's internal validity but its external validity. Notably, by excluding these 44 students, Brutger (2024) is making an already selected sample of highly-achieving students who are both admitted to the elite UC Berkeley and are interested in a PhD program even more selected. Accordingly, the sample undoubtedly constitutes what case study scholars call an extreme case (see Gerring 2017), so we have concerns regarding whether the results could apply beyond other top-10 to top-20 US universities. To convince the reader otherwise for future analyses of the program, the author could provide descriptive statistics on students who are not selected for the lottery and more clearly define the target population statistically. This way, students who are selected for the program and are in the sample can be compared to a clear target population beyond elite Berkeley students, such as through balance tests involving education data that one can draw from a census or the US Department of Education.

[11] On that score, it is worth repeating that the above analysis pertains the highest possible sample size in the paper (62), not the sample with 4 missing values.

5 Execution of the Program, Survey, and Codebook

The above discussion of external validity and the exclusion of students from the program lead us to questions regarding program design, such as:

- Did the program advertisement not mention the 3.5 GPA threshold?
- If there is going to be a lottery anyway, why not mention the selection criteria right away?
- Alternatively, did these 44 students apply despite knowing that they did not meet criteria and were then denied admissions to the lottery?
- If students did meet the criteria, then why not include these students in the lottery?

We also note some wording challenges with the survey and how it likely impacts the results. Brutger (2024, 383) argues in the abstract that the "program resulted in a 48.9 percentage point increase in the number of students who felt prepared to apply to PhD programs." However, the survey in the paper's Appendix only mentions "graduate school" without explicitly mentioning a "PhD" for the preparation questions. Based on the question wording, it is entirely possible that the students are interpreting "graduate school" to mean a Master's degree, Juris Doctor (JD), or another professional program, not just a PhD. If so, that implies a different conclusion than the one that Brutger (2024) advances.

Furthermore, due to the way in which the experiment is set up, it is impossible to match a student's Pre- and Post-PIPS survey answers and assess within-person effects or changes. Accordingly, we suggest that in the future the author tracks a student's Pre- and Post-PIPS survey answers and then anonymize the data.

More broadly, beyond the lack of a pre-analysis plan, the data structure and codebook are difficult to follow. Table 10 shows the distribution of students who took the

Table 10: Distribution of Students Who Took the Surveys by Semester

Already Enrolled	Not Offered Spots	Spring 2021	Fall 2021	Total
No	25	0	17	42 (Pre-PIPS Obs.)
Yes	0	8	12	20 (Post-PIPS Obs.)
Total	25	8	29	62 (All Obs.)

Notes: According to the author's codebook, `Enrolled` is a binary variable, indicating whether the student enrolled and completed the PIPS program when they took the survey. The distribution per semester, `Enrolled_Sem`, contains three categories: the first category is for those who did not receive spots or who had not yet enrolled in PIPS (0), which we call "Not Offered Spots"; the second category is for those enrolled in Spring 2021 (1); and the third category is for those enrolled in Fall 2021 but had not yet completed PIPS (2).

surveys by semester found in the Dataverse `.csv` file. While we commend the author for allowing students the ability to complete the survey anonymously, the data in the replication file suggest to us that none of the students enrolled in Spring 2021 who completed the Post-PIPS survey also completed the Pre-PIPS survey. On top of that, it is hard to follow why the author writes that the second category of `Enrolled_Sem` is for those enrolled in the Fall 2021 but that had not yet completed PIPS. Based on our review of the data, the `Enrolled` variable shows that some students did indeed complete PIPS. We thus recommend that the author update the codebook to reflect the actual content of the data and/or provide additional clarity.

6 Conclusion

This paper provides a reproduction and replication of Brutger (2024). For the reproduction, we are able to reproduce the results from the author's original analysis, although we do note some minor coding challenges. For the replication, our additional analyses lead us to conclude that the limited sample size and design limitations do not enable the author to reach such certain and broad conclusions that the program "works". Particularly concerning are the wording of the survey questions underpinning the analysis, sample selection issues, lack of statistical power, and the null result on student interest in a PhD. With respect to the latter, our view is that for the program to fully "work", it must not only help students with preparation but encourage them to

apply as well. Indeed, given that the Pipeline Initiative in Political Science (PIPS) is clearly a program with a noble goal, we hope that the author can address our concerns to improve the program and its analysis in the future.

By the same token, we would like to be clear that the conclusion regarding application preparation is likely true: it is hard to imagine going through all of the trouble that Brutger (2024) did for PIPS to have no impact on students' preparation for *graduate school* applications. In technical terms, PIPS is a very strong treatment. In our view, it is thus likely that with a greater sample size, the author's conclusions regarding preparation will be more statistically robust for UC Berkeley students—though not necessarily students at less elite institutions.

Finally, we commend Brutger (2024) for starting PIPS during the difficult pandemic times and finding a system that accomodates both the students and the institution. Notably, PIPS does not impose additional burdens on the academic system but improves the educational experience for students facing additional uncertainties and barriers in education. Qualified underrepresented and marginalized individuals, including first-generation students, can greatly benefit from additional guidance and resources steering them toward higher educational achievements, such as a PhD (Byrd and Mason 2021). Although two students dropped PIPS due to unforeseen circumstances in the COVID-19 era, the fact that the program continued speaks to its commitment and likely success beyond the short-term. Bravo!

References

Abadie, A., Athey, S., Imbens, G. W. and Wooldridge, J. M.: 2023, When Should You Adjust Standard Errors for Clustering?, *Quarterly Journal of Economics* **138**(1), 1–35.

Arel-Bundock, V.: 2022, modelsummary: Data and Model Summaries in R, *Journal of Statistical Software* **103**(1), 1–22.

Brambor, T., Clark, W. R. and Golder, M.: 2006, Understanding Interaction Models: Improving Empirical Analyses, *Political Analysis* **14**(1), 63–82.

Brutger, R.: 2024, The PhD Pipeline Initiative Works: Evidence from a Randomized Intervention to Help Underrepresented Students Prepare for PhDs in Political Science, *Journal of Politics* **86**(1), 383–387.

Byrd, C. and Mason, R.: 2021, *Academic Pipeline Programs*, Lever Press.

Gerring, J.: 2017, *Case Study Research: Principles and Practices*, second edn, Cambridge University Press, Cambridge.

Gomila, R.: 2021, Logistic or Linear? Estimating Causal Effects of Experimental Treatments on Binary Outcomes Using Regression Analysis, *Journal of Experimental Psychology: General* **150**(4), 700–709.

Hlavac, M.: 2022, stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.3.

Imbens, G. W. and Rubin, D. B.: 2015, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press, New York.

7 APPENDIX A: Computational Reproducibility Summary

Table A7: Replication Package Contents and Reproducibility

Replication Package Item	Fully	Partial	No
Raw data provided			✓
Analysis data provided	✓		
Cleaning code provided			✓
Analysis code provided	✓		
Reproducible from raw data			✓
Reproducible from analysis data	✓		

Note: This table summarizes the replication package contents contained in Brutger (2024).

8 APPENDIX B: Power Analysis Calculations

Statistics used for the power analysis:

- **Mean for the treatment group (PhD $\bar{\text{I}}$nterest$_{\text{treatment}}$)**:

  $PhD\ \bar{I}nterest_{\text{treatment}} = \frac{1}{n_{\text{treatment}}} \sum_{i=1}^{n_{\text{treatment}}} \text{PhD Interest}_{\text{treatment},i}$

- **Mean for the control group (PhD $\bar{\text{I}}$nterest$_{\text{control}}$)**:

  $\text{PhD } \bar{\text{I}}\text{nterest}_{\text{control}} = \frac{1}{n_{\text{control}}} \sum_{i=1}^{n_{\text{control}}} \text{PhD Interest}_{\text{control},i}$

- **Standard deviation for the treatment group ($sd_{\text{treatment}}$)**:

  $sd_{\text{treatment}} = \sqrt{\frac{1}{n_{\text{treatment}}-1} \sum_{i=1}^{n_{\text{treatment}}} (\text{PhD Interest}_{\text{treatment},i} - \text{PhD } \bar{\text{I}}\text{nterest}_{\text{treatment}})^2}$

- **Standard deviation for the control group (sd_{control})**:

  $sd_{\text{control}} = \sqrt{\frac{1}{n_{\text{control}}-1} \sum_{i=1}^{n_{\text{control}}} (\text{PhD Interest}_{\text{control},i} - \text{PhD } \bar{\text{I}}\text{nterest}_{\text{control}})^2}$

- **Pooled standard deviation (sd_{pooled})**:

  $sd_{\text{pooled}} = \sqrt{\frac{(n_{\text{treatment}}-1)sd^2_{\text{treatment}} + (n_{\text{control}}-1)sd^2_{\text{control}}}{n_{\text{treatment}}+n_{\text{control}}-2}}$

- **Effect size (Cohen's d)**:

  $d = \frac{\text{PhD } \bar{\text{I}}\text{nterest}_{\text{treatment}} - \text{PhD } \bar{\text{I}}\text{nterest}_{\text{control}}}{sd_{\text{pooled}}}$

- **Hedges' g (corrected Cohen's d for the small sample size)**:

  $g = d \times \left(1 - \frac{3}{4(n_{\text{treatment}}+n_{\text{control}})-9}\right)$

9 APPENDIX C: Reproduction of Appendix Section 4: Results with Four-Point Outcome Measure, Table 3: Effect of PIPS on Interest and Preparation for Graduate School

Table A9: Four-Point Outcome Linear Regression Results

PhD_Interest	(1) PhD to Apply	(2) Prepared Personal Statement	(3) Prepared SOP	(4) Prepared LORs	(5) Prepared
PIPS	0.038	1.022***	1.256***	0.864**	0.892**
	(0.187)	(0.232)	(0.224)	(0.254)	(0.285)
	[0.839]	[0.000]	[0.000]	[0.001]	[0.003]
(Intercept)	3.262***	2.200***	2.300***	2.025***	2.275***
	(0.106)	(0.129)	(0.125)	(0.124)	(0.159)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
No. Obs	62	58	58	58	58

Notes: OLS model. Standard errors are in parentheses. P-values are in brackets. Significant at the ***[1%] **[5%] *[10%] level.

10 APPENDIX D: Reproduction of Appendix Section 5: Results Controlling for Demographics, Table 4: Effect of PIPS on Interest and Preparation for Graduate School

Table A10: Linear Regression Results with Controls

PhD_Interest	(1) PhD to Apply	(2) Prepared Personal Statement	(3) Prepared SOP	(4) Prepared LORs	(5) Prepared
PIPS	-0.061	0.482***	0.471***	0.395**	0.385**
	(0.083)	(0.133)	(0.121)	(0.135)	(0.140)
	[0.469]	[0.001]	[0.000]	[0.005]	[0.008]
Male	0.150	0.081	0.104	-0.062	0.060
	(0.081)	(0.131)	(0.119)	(0.133)	(0.138)
	[0.074]	[0.539]	[0.384]	[0.645]	[0.667]
White	0.009	-0.022	0.203	0.112	0.009
	(0.088)	(0.140)	(0.127)	(0.143)	(0.148)
	[0.920]	[0.875]	[0.117]	[0.436]	[0.950]
First Gen	0.004	-0.159	-0.098	0.091	0.012
	(0.044)	(0.132)	(0.119)	(0.134)	(0.138)
	[0.961]	[0.231]	[0.413]	[0.497]	[0.933]
(Intercept)	0.869***	0.491***	0.469***	0.193	0.406**
	(0.074)	(0.118)	(0.107)	(0.120)	(0.125)
	[0.000]	[0.000]	[0.000]	[0.115]	[0.002]
No. Obs	57	57	57	57	57

Notes: OLS model. Standard errors are in parentheses. P-values are in brackets. Significant at the ***[1%] **[5%] *[10%] level.

11 APPENDIX E: Reproduction of Appendix Section 6: Results limited to those enrolled in PIPS, Table 5: Effect of PIPS among those Admitted to the 2021 Program

Table A11: Linear Regression Results Limited to Enrolled in 2021

PhD_Interest	(1) PhD to Apply	(2) Prepared Personal Statement	(3) Prepared SOP	(4) Prepared LORs	(5) Prepared
PIPS	-0.041	0.536***	0.588***	0.431**	0.657***
	(0.092)	(0.140)	(0.119)	(0.156)	(0.131)
	[0.658]	[0.001]	[0.000]	[0.009]	[0.000]
(Intercept)	0.941***	0.353**	0.412***	0.235*	0.176
	(0.068)	(0.101)	(0.086)	(0.112)	(0.094)
	[0.000]	[0.001]	[0.000]	[0.044]	[0.070]
No. Obs	37	35	35	35	35

Notes: OLS model. Standard errors are in parentheses. P-values are in brackets. Significant at the ***[1%] **[5%] *[10%] level.