The Effects of a Graphic Summary on the Reading of Causal Analytics Reports

ABSTRACT

Does a graphic summary help a reader to capture the story more accurately from a causal analytics report? In this experiment, I assess whether and the extent to which the use of a causal graph summarizing the findings result in a more accurate reading of a causal analytics report. 161 volunteers were recruited from Amazon Mechanical Turks over three weeks, out of which 148 completed the task with valid responses. Among these volunteers, 76 and 72 respondents were randomly assigned to each of two assignment groups – one with a reading task and a question and the other one with the same reading task and question but also a graphic summary. The randomization has led to two very similar groups in terms of education, working experiences, domain knowledge, and data science training, which may influence a reader's reading comprehension in advanced analytics reports. The findings suggest that a graphic summary helps to increase the reading accuracy by removing irrelevant information, but there is no evidence that the reading comprehension had any improvements with the presence of a graphic summary. The limitations and future extensions are discussed.

Keywords: Causal knowledge graph, Causal analytics, Visualization, Experiments

Jupyter Notebook:

https://github.com/victorchenberkeley/w241_finalproject/blob/main/w241_final_analysis.ipynb

INTRODUCTION

Causal Knowledge Graph (KG) is a graphic representation of domain expertise on the causes-and-effects hypotheses among real-world concepts such as financial returns, acquisitions, salary, etc. A causal KG cumulatively synthesizes both deductive knowledge such as economic and behavioral theories from the scientific literature and inductive knowledge such as opinions from domain experts. In areas like data science and artificial intelligence (AI), causal KG is increasingly recognized as a tool to improve the accessibility and explainability of domain knowledge and support causal inference in AI tasks (Jaimini & Sheth, 2022). Its application is emerging in data science, visualization, knowledge synthesis and management, and decision intelligence.

A notable effort is the GoPeaks initiative, jointly funded by Dell Technologies, the Institute of Management Accountants (IMA), and the National Science Foundation (NSF) (2021 I-Corps Grant and 2022-2023 Partnerships-for-Innovation Grant)[1] among others. It seeks to build an end-to-end solution from (1) synthesizing texts from domain expert documents (e.g., industry reports, academic publications) to causal graphs (GoPeaks, 2021) to (2) crowdsourcing causal knowledge into a graph database for query and management (GoPeaks, 2022a), and ultimately to (3) feeding causal KGs with meta-analysis evidence (GoPeaks, 2022b) into data science tools such as Python, R, Stata, and SAS. Numerous research efforts at GoPeaks have been recognized at major industry and academic events such as the Knowledge Graph Conference, NSF Open Knowledge Network (OKN) Innovation Sprint, and the Annual Causal Science Meeting. A recent proposal on causal decision intelligence co-developed by GoPeaks, NC State University,

[1] See an example on Artificial Intelligence System for Enterprise Performance Management that Integrates Causal Analytics and Human Expertise (NSF Award No. 214124), accessible at https://www.gopeaks.org/kg-brain-healthcare.

UCSF, and Texas A&M, among others, has been preselected to advance to full proposal submission for one of the 7 NSF National AI Institutes.

Currently, most of the existing GoPeaks tools seek to support causal inference analysts who focus on observational data, but extensions are being developed for experimentalists to improve their experimental designs.[2]

Although many advancements on the technology of causal KG were made over the last few years, it remains contested whether and to what extent such advancements may affect the efficacy of their intended users, e.g., data-driven decision-makers. As a new initiative, recently funded by NSF PFI-TT grant, I am starting to understand whether and how static and dynamic graphic graphs may improve the communication efficacy of advanced analytics research, readers' trust in the analytic findings, and ultimately the actionability of recommendations into products or policies.

The future of business relies on prescriptive insights from data analytics that recommend actions a company's executives should take. But in most organizations, it remains a challenge to close the gap between data-savvy analysts and the pressed-for-time, story-focused executives. Data visualization has been a powerful solution to close this gap by translating formulas into visual patterns, with a fast-growing global market value expected to reach $19.2 billion by 2027. The current focus of visualization is mostly on descriptive and predictive (forecasting) storytelling, with little effort to represent the causal insights, e.g., what actions and how they may affect targeted business outcomes. To my knowledge, there has been no research on the efficacy of causal visualization in data analytics research. Notably, we know little about whether readers

[2] I recently joined the Fidelity Investment to head its analytics efforts on experimental design and causal inference, including an initiative to combine potential outcome and causal structural model methods to improve the experimental platforms.

without advanced analytics training may benefit from a graphic summary. A **graphic summary** is a simplified visual illustration of the core ideas, theories, models, or findings from a textual data, capturing a necessary but insufficient subset of causal relationships that have been discussed or elicited in the text.

Thus, the research question in this experiment is: ***Would a graphic summary of the findings from complex causal analytics research lead to a more accurate reading of this research?***

THEORY AND HYPOTHESES

Humans use causality and contemplate interventions and counterfactuals in their understanding and decision-making (Pearl & Mackenzie, 2018). For important decisions, the human mind engages deliberate and slow reasoning of the complex causes-and-effects relationships among environmental cues, choice set, and predictive value of outcomes. A causal structure of the environment helps humans focus on a manageable subset of cues, thus effectively reducing the computational complexity (Garcia-Retamero & Hoffrage, 2006). In addition, a graphic representation of information may improve the understanding and judgment in human decisions, especially when such judgment requires spatial thinking, such as identifying connections, links, neighborhoods, and distance among different entities (Kirschenbaum & Arruda, 1994; Lurie & Mason, 2007). I argue that a causal structure among cues represents this spatial problem. Thus, a graphic representation of the causal structure may lead to a better understanding of the situation and thus greater confidence in the insights from data.

I argue a graphic summary can improve the reading accuracy of causal analytics research texts in two ways. First, it is a necessary subset of a text and thus helps the readers to screen out unnecessary or irrelevant information. It helps a reader focus on the key insights without being distracted by tangential information. Second, it translates cognitively taxing information (e.g., verbal logic) into sensual interactions. For thousands of years, stories have been communicated through sensually interactive signals, such as speech, sound, and visuals. Generally, storytelling from data involves visual media, such as illustrations, graphs, animations, and videos – in a movement named "communication-minded visualization" (Viegas & Wattenberg, 2006: 801). In recent years, visualization has been explored to aid a reader's understanding of advanced analytic models from verbal/textual data (Choudhry et al., 2020).

Thus, I propose:

H1: A graphic summary will increase the accuracy of reading causal analytics research.

H1a: A graphic summary will increase the accuracy of reading causal analytics research by removing irrelevant information.

H1b: A graphic summary will increase the accuracy of reading causal analytics research by enhancing the comprehension of the content.

EXPERIMENTAL DESIGN

Treatment

To test for reading accuracy, I designed a reading task of the key empirical findings from a replication I did based on Chernozhukov et al. (2018). Using a public database of 1991 low- and middle-income families, I estimated the mean difference in household net financial wealth between 401k eligible families and ineligible families, the coefficient of 401k eligibility on household net financial wealth in a linear regression without causal inference, as well as the causal effect estimation of 401k eligibility on household net financial wealth using debiased machine learning (DML) method. I summarized the background of the study and the three key findings. The accurate estimate is the causal effect after debiasing the confounding from the measures.

As illustrated in Appendix 1A, the control group's assignment is to read only the textual summary and select which number was the correct estimate of causal effect (in terms of 1991 dollar values) of 401k eligibility on a household's net financial wealth. As illustrated in Appendix 1B, a treatment combines textual and graphic summaries. The text is the same as that in the control group, but it was accompanied by a graphic summary on the same page. The graphic summary has reduced some irrelevant information by removing an incorrect answer from the three numbers in the text and visually illustrating the confounding and causal paths mentioned in the text. To differentiate the coefficient without clearing the confounding biases and the causal effect estimate after debiasing the measures, they were colored blue and red, respectively. To highlight that debiased measures are 'cleaner,' they pop up from the background of all confounding pathways.

Randomization Process

Volunteers were recruited from Amazon Mechanical Turk (MTurk) between March 1st, 2022, and April 15th, 2022. I restricted the selection of MTurk workers from English-speaking working professionals based in the US. After completing the task, recruited MTurk workers were directed from the MTurk platform to a Berkeley Qualtrics Survey, which asked for a unique survey code to validate their answers. Several questions were asked repeatedly under different framing to verify that a reader was carefully reading the questions before submitting an answer. The complete survey questions can be found in Appendix 2. As shown in Figure 1, on the Qualtrics platform, a randomizer was initiated after a respondent filled out background information and started accessing the reading task (textual summary only, or textual and graphic summary).

[Figure 1]

Comparison of Potential Outcomes

The primary potential outcome is ***whether a reader can identify this accurate estimate out of the three numbers in the text*** (Y1). To make the answer less obvious, the text had no direct indication of which number among the three is the correct estimate. To analyze the reading process, two secondary potential outcomes are ***the reading speed*** (Y2), measured as the number of seconds a reader took to complete the task, and ***whether or not a reader resorts to the original Chernozhukov et al. (2018) paper to help complete the task*** (Y3). Not to bias the results by introducing a thorough consultation with the original paper, each reader had to complete the reading task within 5 minutes.

Specifically, I compare Y1, Y2, and Y3 after completing the reading task between the following two groups of volunteers: (1) readers of the textual summary only (control group); (2) readers of the textual and graphic summaries (treatment group). In both groups, everything in the task was the same, except that the textual summary in the treatment group was accompanied with a graphic summary.

Consort Document

By the end of the MTurk recruitment window on April 15th, 2022, 161 volunteers were recruited (each would be awarded $0.75 for the task). Among these respondents, 151 had completed the tasks and entered the correct survey code. Two of these respondents' submissions were removed as inconsistent answers were given for the same question under different frames. Eventually, 148 respondents were valid in the sample; 76 were in the control group, and 72 were in the treatment group (see Figure 2).

[Figure 2]

Power Calculation

Following Serdar, Cihan, Yücel, and Serdar (2021) 's recent study on the best practice of sample size and statistical power, I estimate the minimum sample size $N_{min} = {Z_{\alpha/2}^2 \cdot s^2}/{d^2}$, where, s = 0.20 represents the standard deviation of Y1 obtained from the pilot study, d = 0.05 represents the minimum accuracy of the Y1 estimate of group mean difference, and $Z_{\alpha/2}$ =1.96, representing the Z-value for a confidence level of 95% for a two-tailed test (alpha=0.05).

Thus, the minimum $N_{min} \approx 61$, suggesting the threshold of a sample size to make statistical inference at the 95% confidence level and 0.05 accuracy. The actual number of valid respondents (N=148) in the sample has exceeded the threshold for statistical power, given the estimate's precision.

ANALYSIS

Data and Measurement

The valid sample (N=148) represents a relatively diverse group of working professionals. Overall, the respondents mostly had completed postgraduate education, studied in non-STEM or non-business/economics/finance areas, worked for more than five years, enrolled in 401(k) with personal or discretionary investment experiences, but had no formal training related to causal analytics.

Specifically, as illustrated in Figures 3A – 3F. Specifically, the sample represents the following background covariates that may affect the potential outcomes.

[Figures 3A to 3F]

- Education: 45 with less than undergraduate education, 78 with undergraduate education, 16 with a master's education, and 8 with more than a master's education.
- Area(s) of study: 34 with education in business, economics, and finance, 3 with education in statistics or data science (1 of which also had education in business, economics, and finance), 34 in other STEM (1 of which also had education in business, economics, and finance), and 78 in other area(s).
- Professional experience: 132 with more than 5 years of work experience, 5 with 4-5 years, 7 with 1-3 years, and 4 with no work experience (< 1 year).
- Enrollment in 401(k): 84 enrolled and 64 not enrolled.
- Investment background: 83 primarily handled their investments themselves, 37 had investments by professional investors, and 28 had no investments.
- Data science training on causal analytics: 22 had data science training related to experiments and causal inference, whereas 126 did not.

Now let me analyze the potential outcomes – reading accuracy, reading speed, and whether to check external references. Among 148 respondents, 40 scored the accurate answer ($8000-$9000 causal effect), representing 27% of the sample. Each respondent took 108 seconds (or less than 2 minutes) to complete the reading task. In addition, 45 respondents reported that they checked the references in the list to help them make a judgment. However, given a 5-minute reading time limit, their efforts to check the references were very unlikely to impact their reading comprehension.

Covariate Balance Check

Before estimating the causal effects, I probed into the efficacy of randomization by comparing the background covariates between the treatment and the control groups. Specifically, the mean difference in a covariate k is calculated as

$$\text{Mean difference in } X_k = \text{Mean } X_k \text{ (Treatment Group) - Mean } X_k \text{ (Control Group)}$$

[Table 1 and Figures 3A to 3F]

As Figures 3A to 3F present, the covariates are highly similar between the two assignment groups. Specifically, as reported in Table 1, in terms of **education** (score of 0 for less than undergraduate, 1 for undergraduate, 2 for master's, and 3 for more than master's education), the mean difference in education score is 0.37 (the p-value of t-test for mean equality is 0.004). Although the mean difference is statistically significant, the size of the difference is relatively small (0.37 out of a scale of 3) (see Figure 3A for a detailed distribution).

In terms of **study areas**, respondents trained in business, economics, and finance may have considerably more specialized knowledge about the subject of the reading. Those trained in statistics and data science may know more about the methodology of the reading. In either case, a respondent was given a score of 1 for the relevance of their study areas. Otherwise, a respondent was given a score of 0. As reported in Table 1, the mean difference in the relevance of study areas is 0.04 (the p-value of t-test for mean equality is 0.57) (see Figure 3B for a detailed distribution). Thus, there is no evidence to suggest that the two assignment groups are statistically different in their study areas.

In terms of **professional experience**, respondents with less than one-year work experience were given a score of 0; those with 1-3 years of work experience were given a score of 1; those with 4-5 years of work experience were given a score of 2; and those with more than 5 years of work experience were given a score of 3. As reported in Table 1, the mean difference in the score is 0.08 (the p-value of t-test for mean equality is 0.44) (see Figure 3C for a detailed distribution). Thus, there is no evidence to suggest that the two assignment groups are statistically different in their work experiences.

In terms of **401(k) enrollment**, respondents enrolled in 401(l) were given a score of 1, and 0 otherwise. As reported in Table 1, the mean difference in the score is 0.11 (the p-value of t-test for mean equality is 0.17) (see Figure 3D for a detailed distribution). Thus, there is no evidence to suggest that the two assignment groups are statistically different in their 401(k) enrollment.

Regarding **investment background**, respondents with no investments were given a score of 0; those with investments handled by professional investors were given a score of 0.5, and those handling their investments were given a score of 1. As reported in Table 1, the mean difference in the score is 0.11 (the p-value of t-test for mean equality is 0.08) (see Figure 3E for a detailed distribution). While the difference is statistically significant at the 0.05 level, the actual difference is tiny, only 0.11 out of a scale of 1. In fact, in both groups, most respondents handle their investments themselves.

Regarding **data science training related to experiments and causal inference**, respondents were given a score of 1 if their answer was "Yes" and 0 otherwise. As reported in Table 1, the mean difference in the score is 0.04 (the p-value of t-test for mean equality is 0.55) (see Figure 3F for a detailed distribution). Thus, there is no evidence that the two assignment groups were statistically different in causal analytics training.

Overall, the randomization was highly effective, splitting the sample of respondents into two highly similar readers in terms of their academic and professional backgrounds.

Average Treatment Effect (ATE)

Now we can estimate the average treatment effect (ATE) by calculating the mean difference between the two assignment groups in terms of potential outcomes. To test H1, I calculate the difference in the average reading accuracy of the two groups. Specifically, if respondents in either group correctly selected "$8000 - $9000" as the expected additional net financial wealth if a household were enrolled in 401(k), all else being equal, they were given a score of 1 and 0 otherwise.

[Figure 5]

Specifically, as illustrated in Figure 5, 26 out of 72 respondents correctly identified the causal effect estimates in the treatment group, compared to 19 out of 76 respondents in the control group. As reported in Table 1, the results translate into a mean score of 0.1974 (less than 0.3333 if all answers were random guesses) in the control group and 0.3472 (slightly more than 0.3333 if all answers were random guesses) in the treatment group. Thus, the point estimate of ATE is approximately 0.15. Given the similarity of the variance of the score in both groups (0.23 and 0.16 in the treatment and the control groups respectively), the t-test for mean equality reports a p-value of 0.04. Thus, if the sharp null hypothesis were true (i.e., the difference in potential outcomes for all respondents under the two reading groups was strictly zero), there is only a 4% chance that we would observe the ATE at the level of 0.15. In other words, we reject the sharp null hypothesis under the frequentist approach.

Therefore, the findings support H1.

But was the finding driven simply by an easier random guess (after removing an irrelevant answer in the graphic summary) (H1a) or through a greater comprehension (H1b)? To answer this question, I gave a closer look at the potential outcomes. I conducted a proportion Z-test against a random guess. Specifically, I tested whether the proportion of accurate answers was statistically significantly different from a proportion if everybody in the group made a random guess out of all the choices.

In the control group, if everybody were making a random guess out of three choices given to them, we would have observed 25 correct answers. In fact, we only observed 15, and the proportion Z-test reported a p-value of 0.065, suggesting that the observed answers were statistically significantly *worse* than a random guess. This is likely because the wording in the text used such key phrase as “all else being equal” when referring to a linear regression that has not corrected the confounding biases. The respondents may mistakenly focus on this key phrase to identify causal effects. This was why proportionally more respondents in the control group selected the linear regression coefficient (Almost $6000) as the answer for causal effect – which was incorrect.

In the treatment group, however, if everybody were making a random guess out of three choices (ignoring the graphic summary), we would have observed 24 correct answers. We observed 25. A proportion Z-test reported a p-value of 0.86, suggesting this observed proportion of correct answers was not statistically significantly different from a random guess out of three. Furthermore, if the 57 readers who selected either number annotated on the graphic summary, we would have observed 19 correct answers. We observed 25. A proportion Z-test for the 57 readers reported a p-value of 0.25, suggesting that the selection out of two numbers in the graphic summary was not statistically significantly different from a random guess.

In summary, while H1 was supported, the evidence does not lend strong support that the readers significantly improved their reading comprehension with a graphic summary – their answers were likely driven by an easier random guess (one out of two, instead of one out of three).

As a supplementary analysis, as Table 1 reports, the reading process in terms of reading speed and whether checking references does not seem to be different between the two assignment groups.

Subgroup Analysis

[Table 2]

Next, I explored how the ATE varies within subsamples based on the covariate values. As previously discussed, a minimum of 61 observations is required to make statistical inference at the 95% confidence level with a 0.05 accuracy from the true ATE. As Table 2 reports, the point estimate of ATE remains consistently positive for all subsamples. However, there were not enough observations to have statistical power in many subsamples. For subsamples that met or exceeded the minimum sample size of 61, the following groups reported a statistically significant and positive ATE: Readers who had no formal causal analytics training, enrolled in 401(k), handling their investments themselves, did not study anything on business, economics, finance, statistics, or data science, and had worked for more than five years. Further analysis in a larger-sample experiment is required to test whether the ATE remains statistically significant.

Regressions

Finally, given the binary nature of the reading accuracy (1 if correct, 0 otherwise), I conducted two logistic regressions to estimate the treatment coefficient on the reading accuracy. In model 1, I only included the intercept and the treatment. In model 2, I included all the covariates. The correlation matrix (Appendix 3) suggests no exposure to severe multi-collinearity biases. In both models, heteroskedasticity-corrected standard errors were reported.

[Tables 3 and 4]

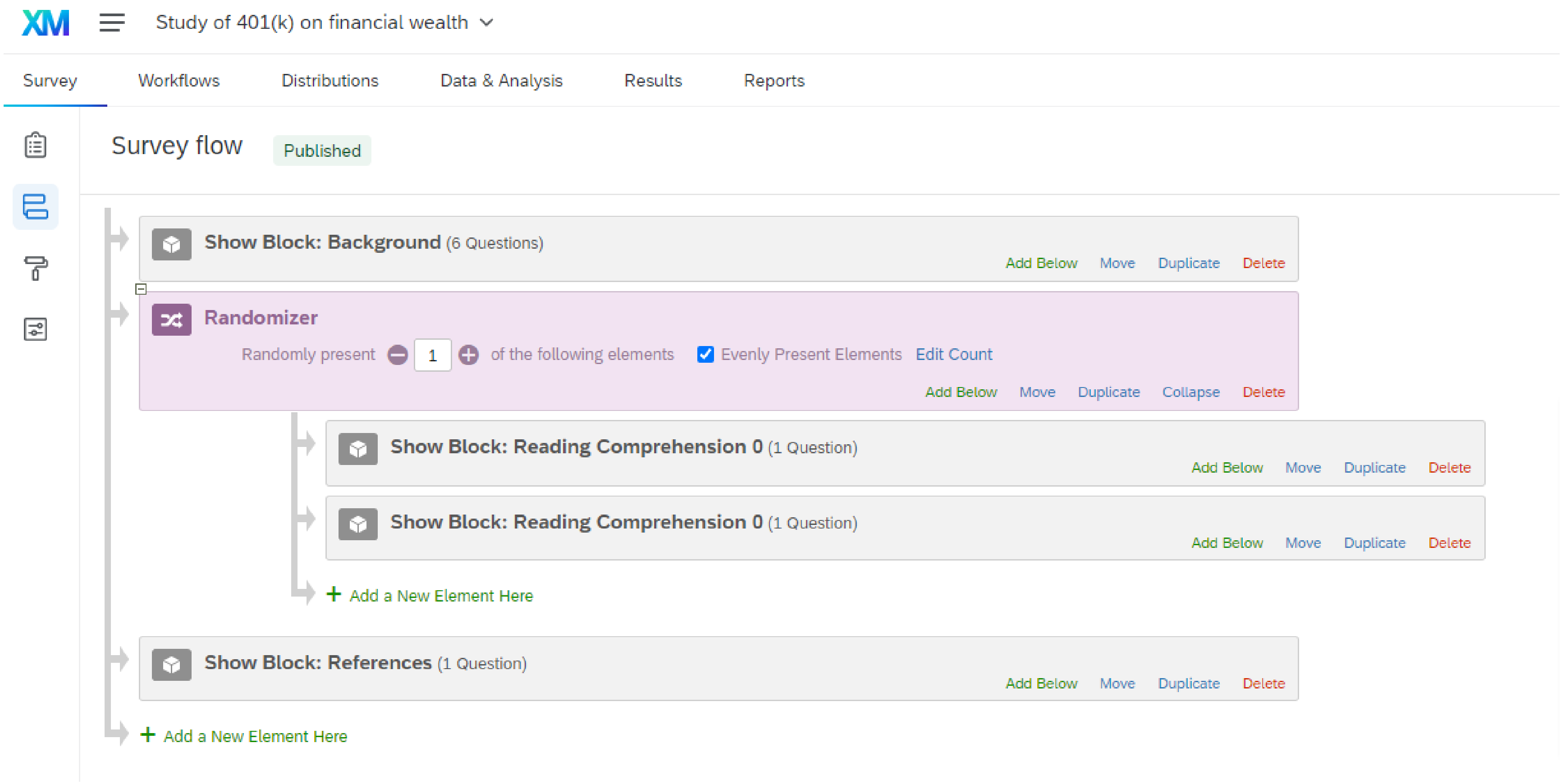
As Table 3 reports, when only the treatment and the intercept were included in the model, the treatment had a 0.77 coefficient (p-value = 0.042) on the binary variable of the outcome. It is statistically significant at the 95% confidence level. Table 4 reports the results after adding all the covariates. The model appears to have improved from model 1, with a greater Pseudo R-squared (suggesting a better goodness-of-fit with the data) and a larger Log-Likelihood. As Table 4 reports, the coefficient of the treatment is 0.79 (p-value = 0.042), which remains similar as well as statistically significant. This coefficient translates to an odds ratio of $e^{0.79} \approx 2.72$ for a reader in the treatment group to score the correct answer compared to a control group reader.

Conclusion and Discussion

This experiment seeks to analyze the causal effects of a graphic summary on the reading accuracy of causal analytics research. The experiment suggests that the reading accuracy improved when a graphic summary was accompanied by the texts, where treatment is a graphic summary that has reduced some unnecessary information from the texts and differentiates two different methods used in the research. However, such improvement seems primarily driven by an easier random guess after removing irrelevant information, and no evidence is found for improved comprehension of the texts. Readers' reading process also appeared to be similar whether or not there was a graphic summary, measured as the reading speed and whether checking external references or not.

This study is a preliminary analysis to start a series of more advanced experiments on whether and how causal KGs may affect the communication efficacy, trust, and actionability of advanced causal analytics research. It has various limitations that should be addressed in future efforts. First, the sample size currently lacks the power to make accurate statistical inferences when the sample is divided into sub-samples. Future studies should seek to increase the sample size. Second, it is valuable to use block-specific randomization by different covariates. Due to the small-time window and sample size, I did not do this. Third, more varieties of topics, research subjects, and analytics research methods should be engaged in the experiments to explore the external validity when the domain changes. Finally, future experiments should target a more accurate representation of the intended users (e.g., data-driven decision-makers).

Figure 1. Randomization



Survey flow Published

Show Block: Background (6 Questions)

Randomizer

Randomly present 1 of the following elements ☑ Evenly Present Elements Edit Count

Show Block: Reading Comprehension 0 (1 Question)

Show Block: Reading Comprehension 0 (1 Question)

Show Block: References (1 Question)

Figure 2. Flow Diagram Tracking the Sample Observations over the Experiment

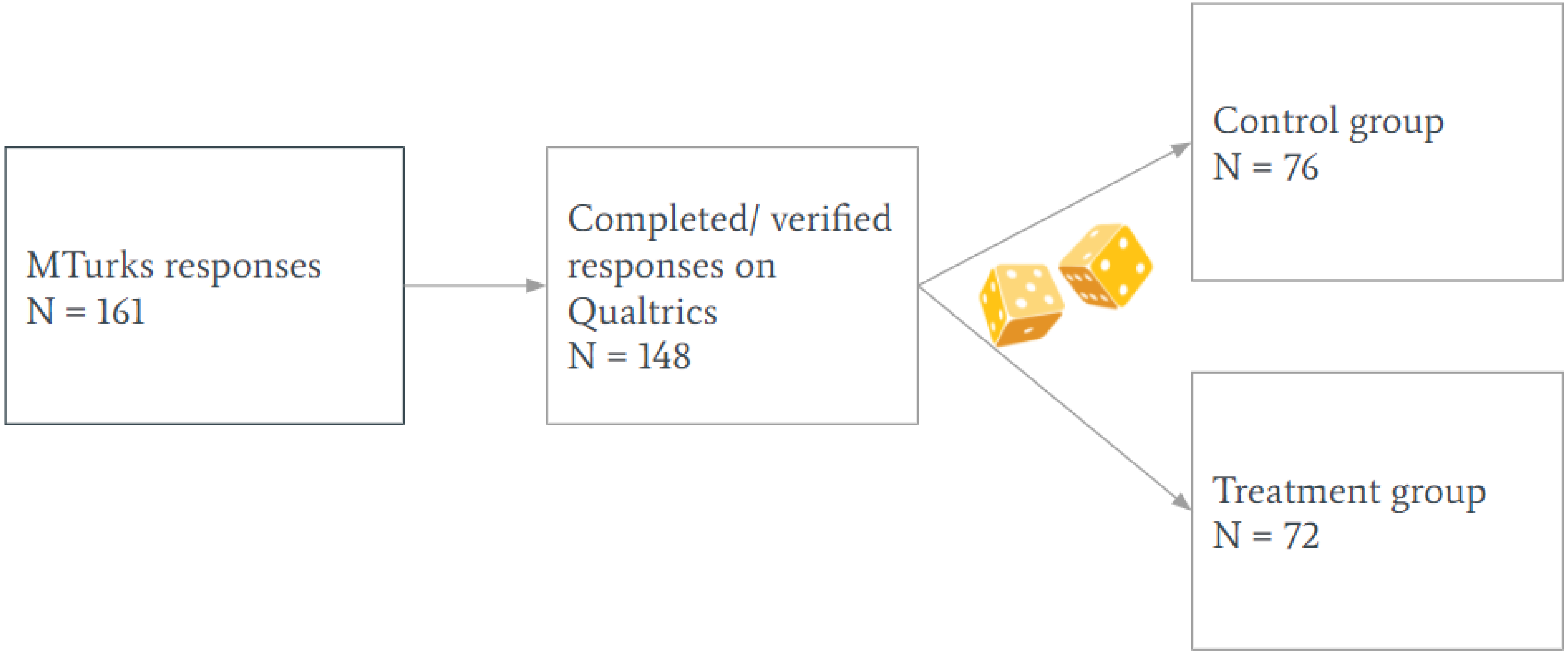


Figure 3A. Sample Distribution by Education

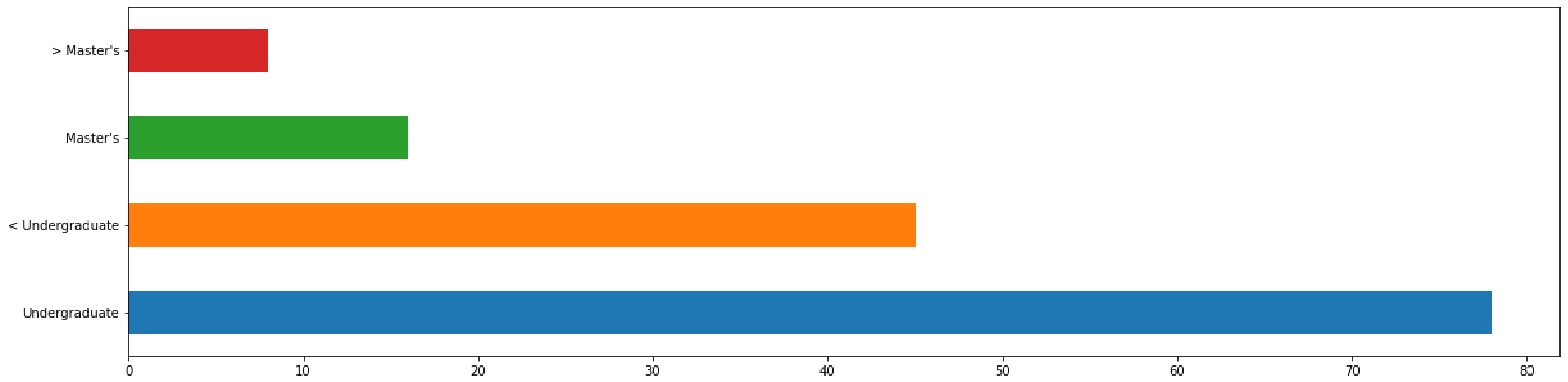


Figure 3B. Sample Distribution by Study Area

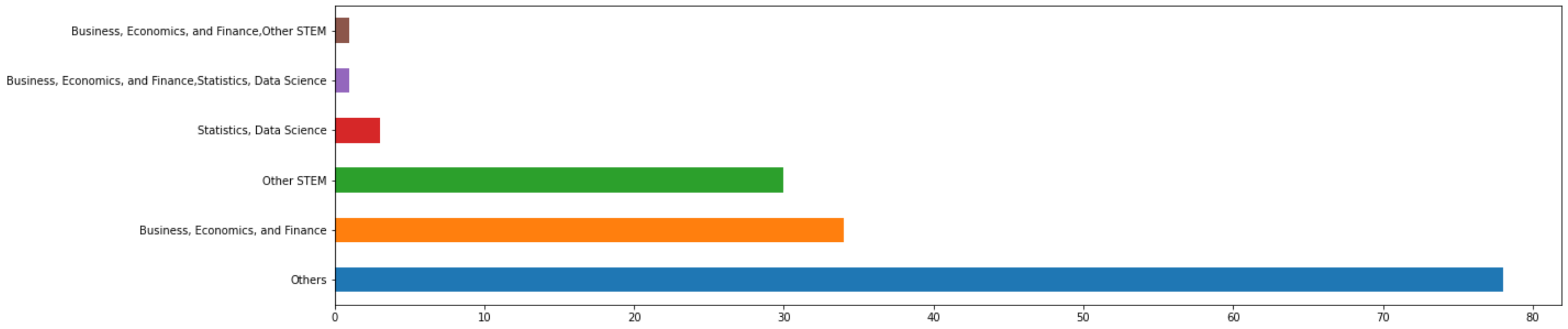


Figure 3C. Sample Distribution by Professional Experience

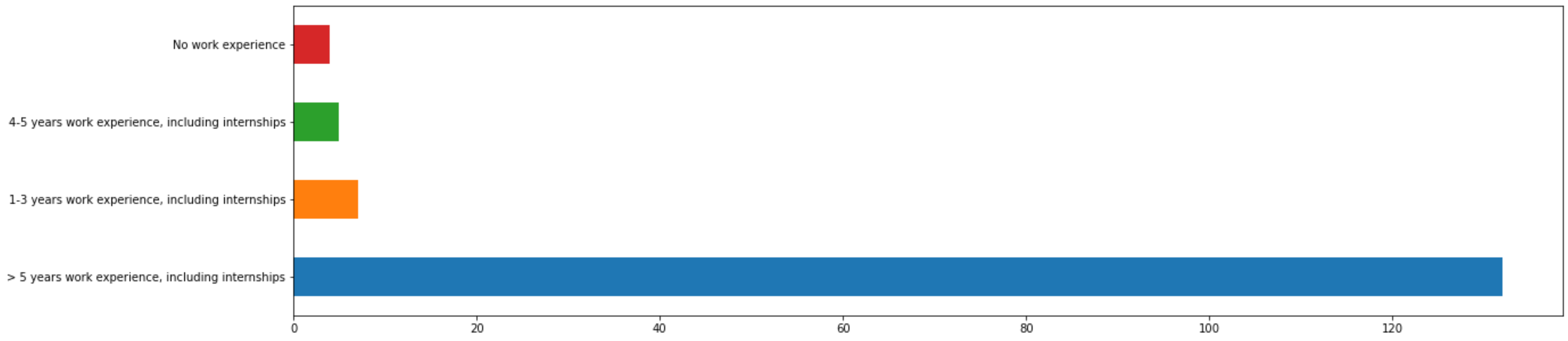


Figure 3D. Sample Distribution by Enrollment in 401(k)

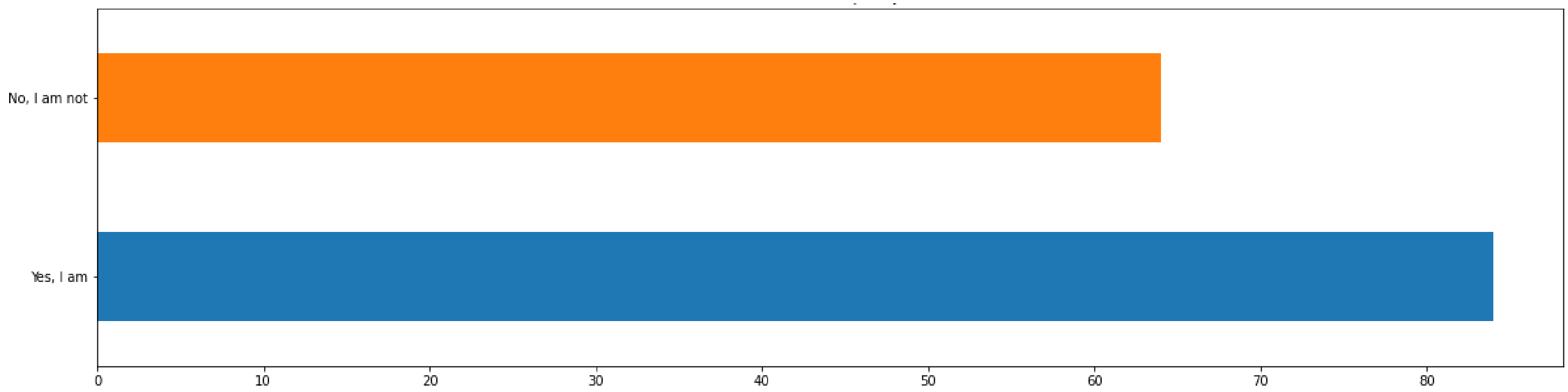


Figure 3E. Sample Distribution by Investment Background

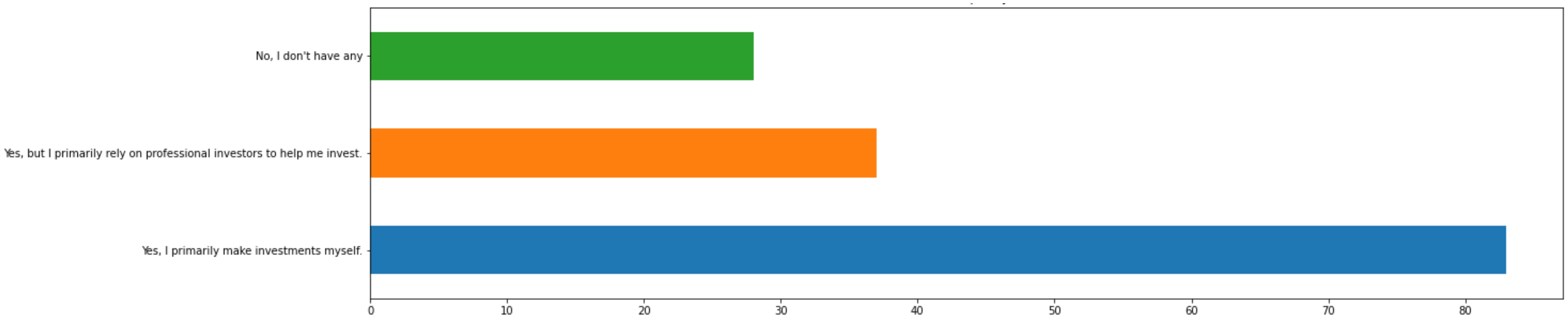


Figure 3F. Sample Distribution by Data Science Training focusing on Causal Analytics

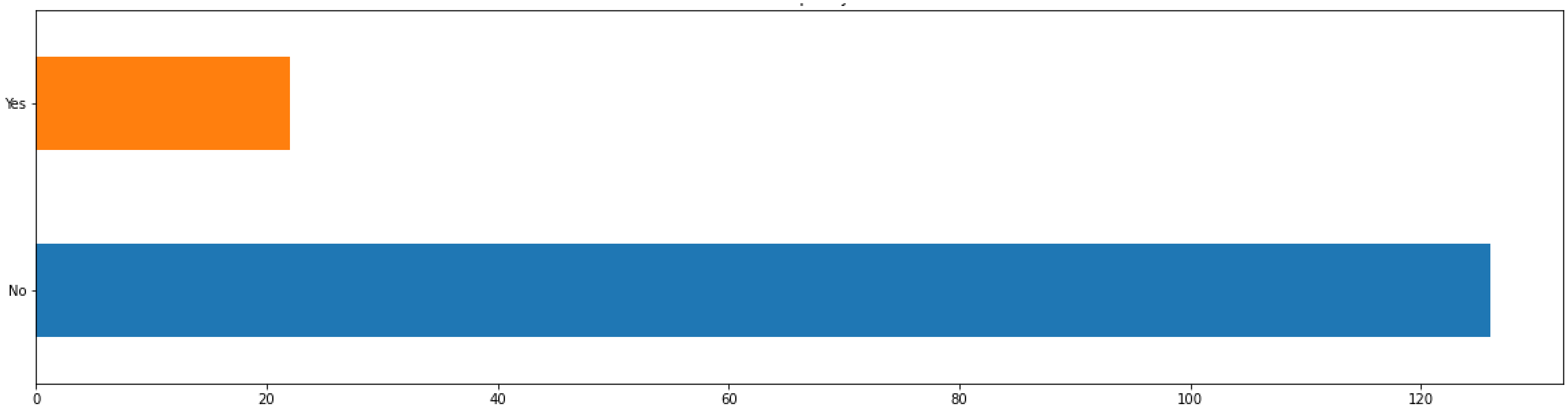


Figure 4A. Sample Distribution by Education and Assignment Group

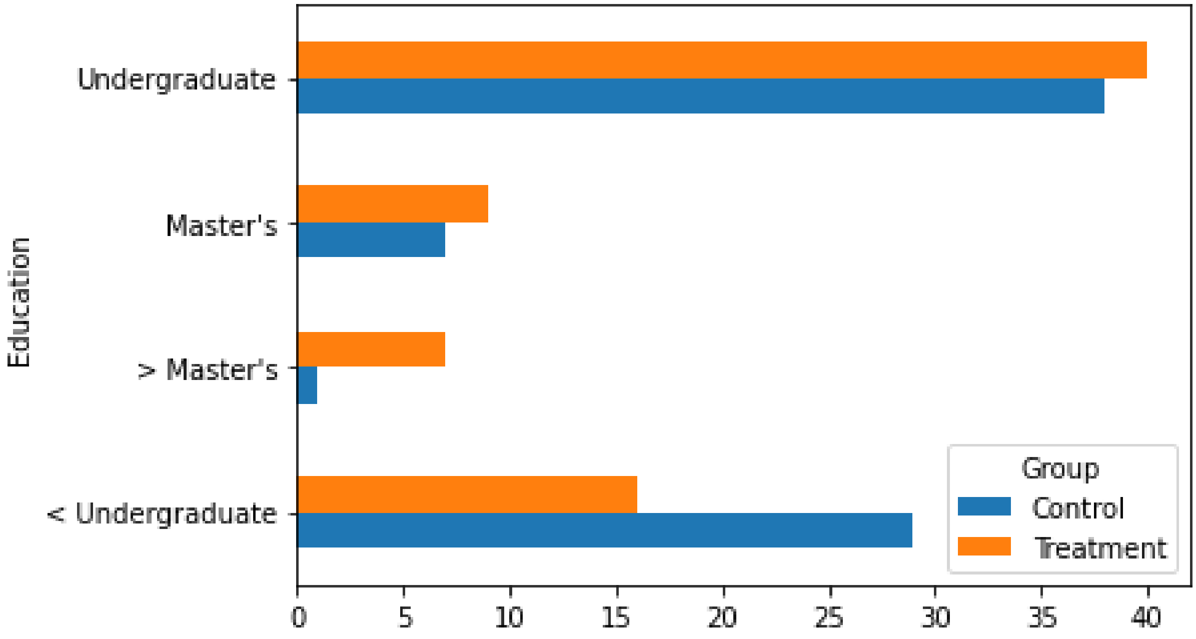


Figure 4B. Sample Distribution by Professional Experience and Assignment Group

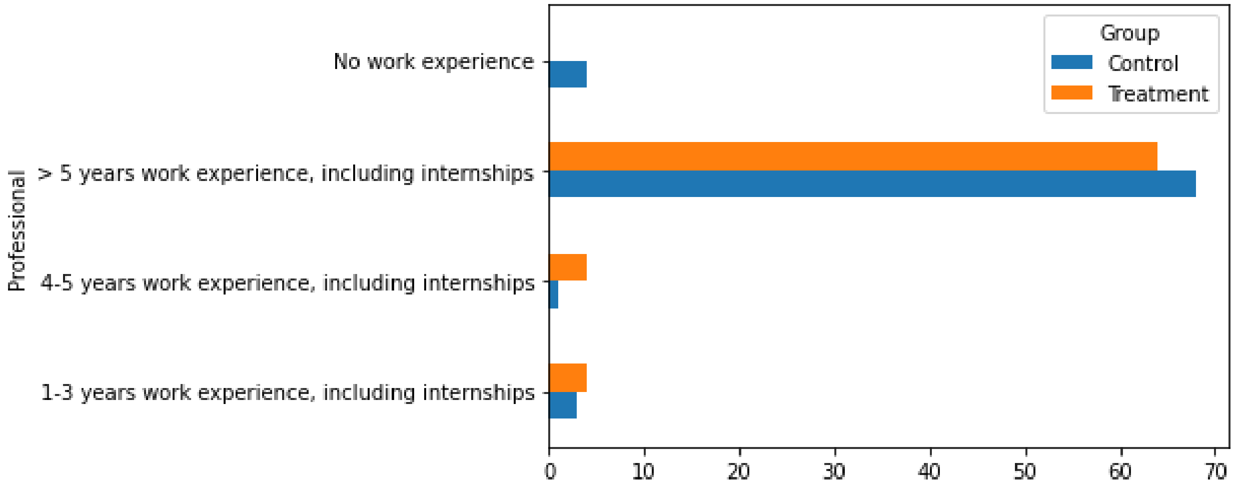


Figure 4C. Sample Distribution by Area(s) of Study and Assignment Group

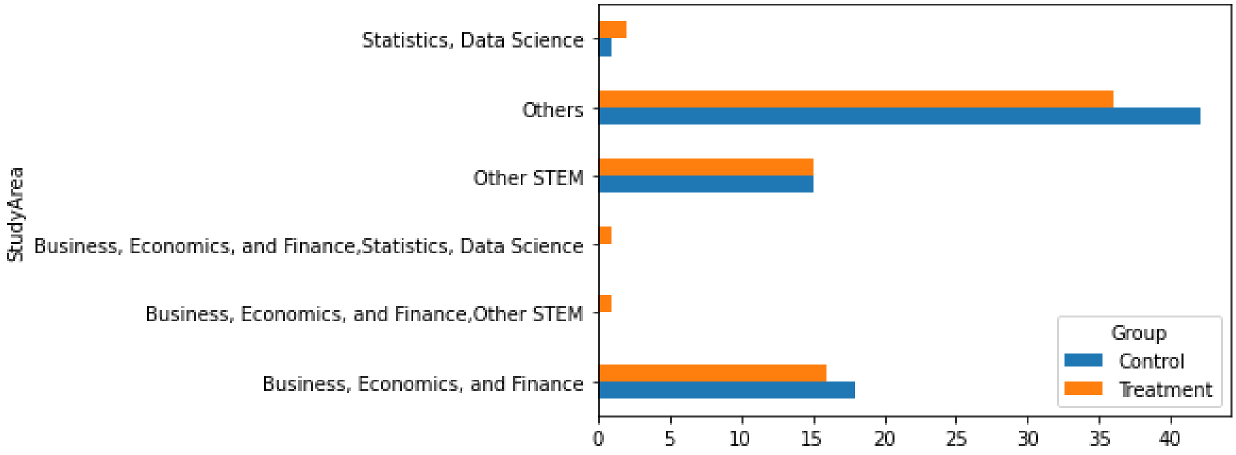


Figure 4D. Sample Distribution by Enrollment in 401(k) and Assignment Group

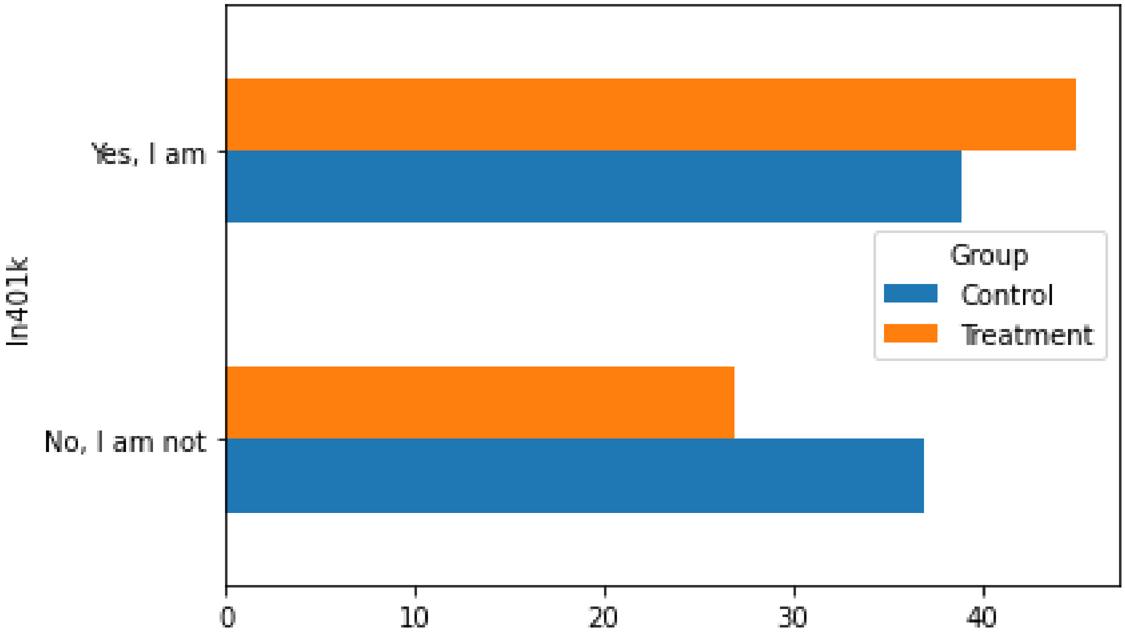


Figure 4E. Sample Distribution by Investment Background and Assignment Group

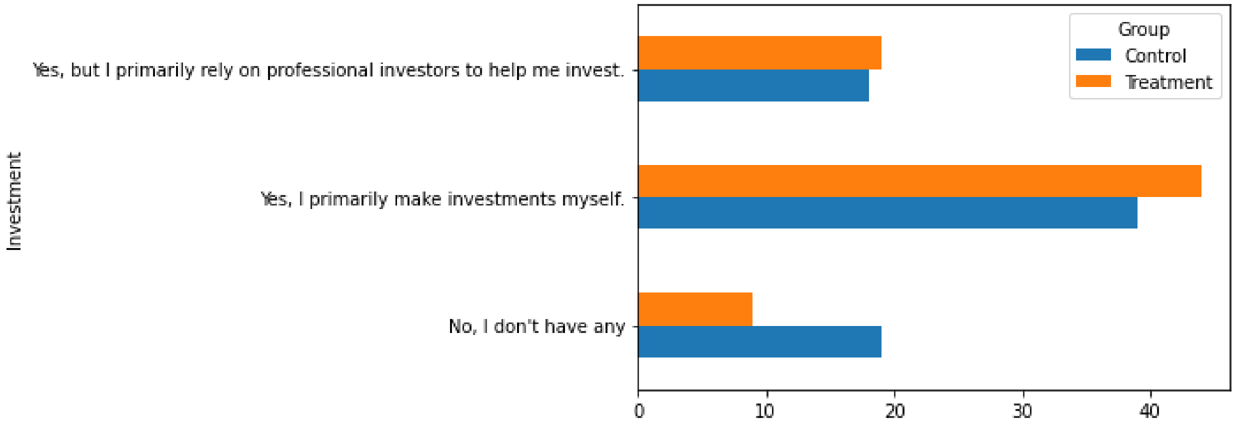


Figure 4F. Sample Distribution by Causal Analytics Training and Assignment Group

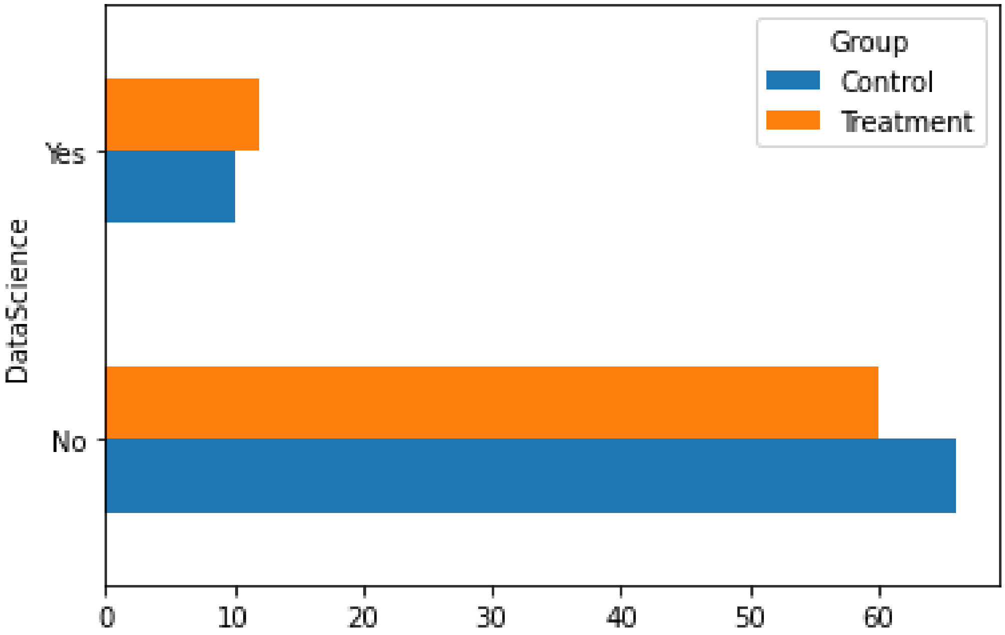


Figure 5. Sample Distribution of Reading Accuracy by Assignment Group

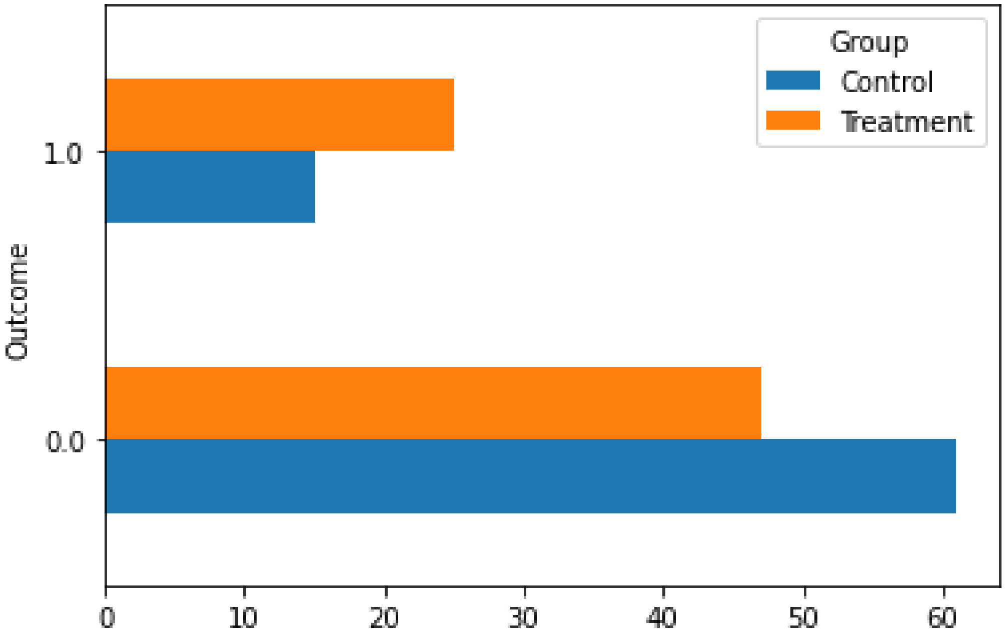
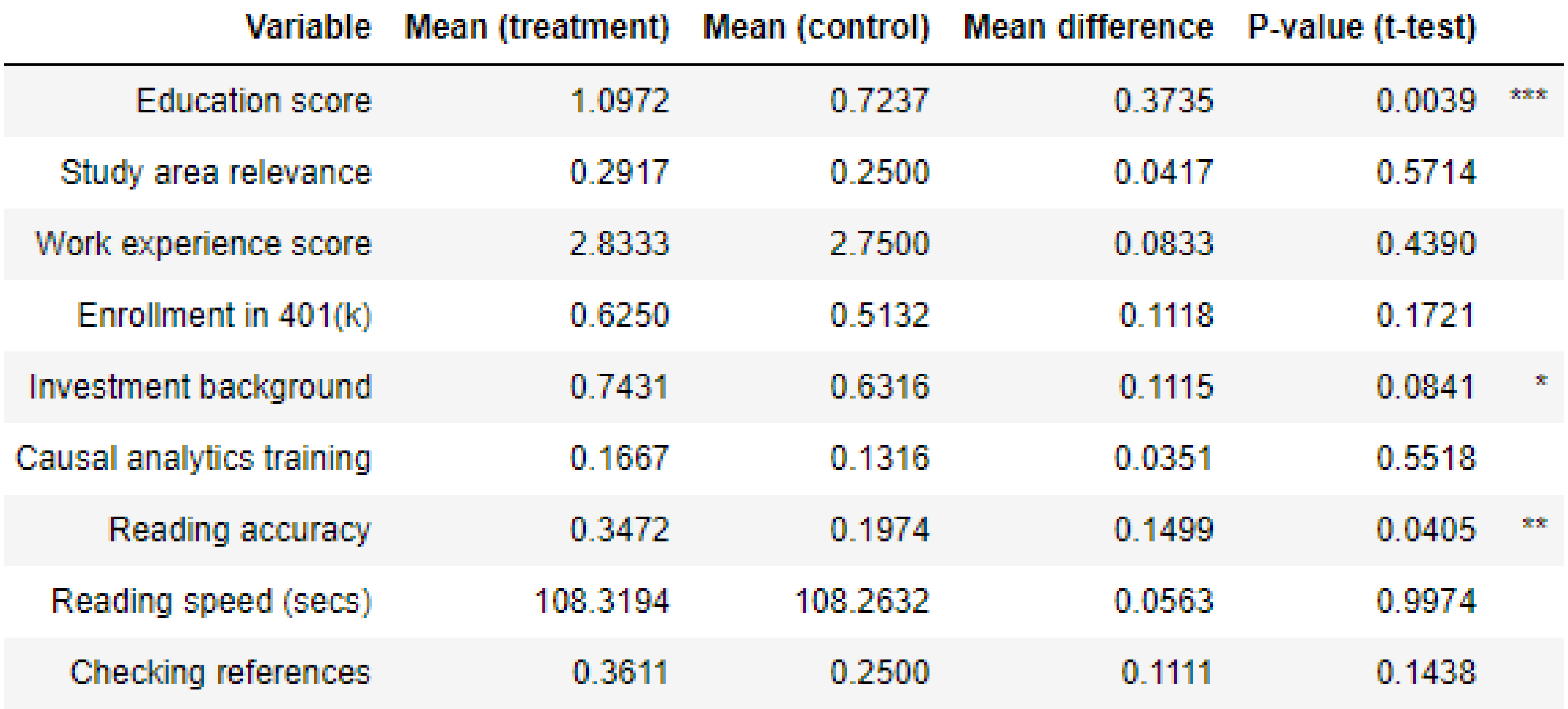


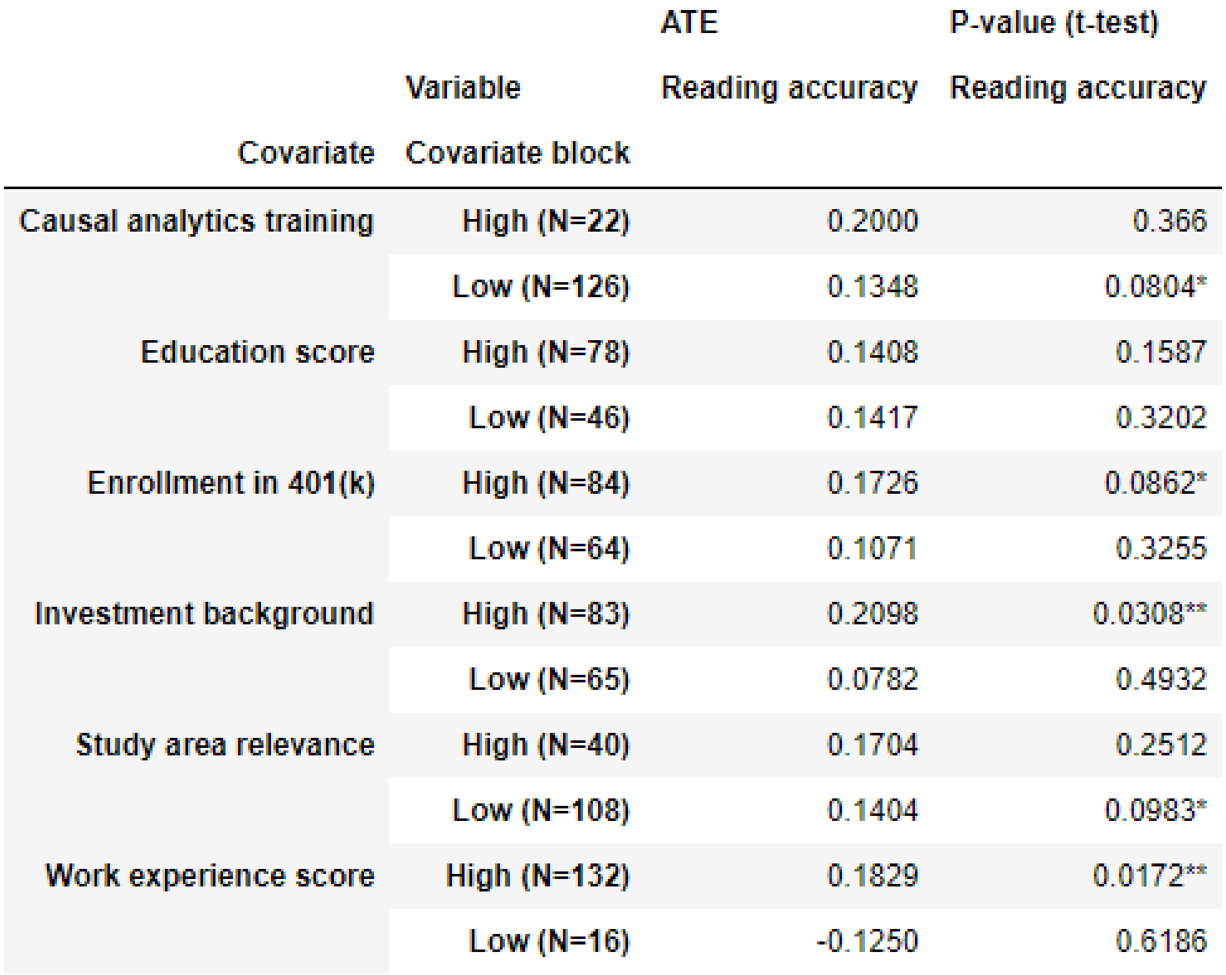
Table 1. Mean Difference in Covariates and Potential Outcomes



Variable	Mean (treatment)	Mean (control)	Mean difference	P-value (t-test)	
Education score	1.0972	0.7237	0.3735	0.0039	***
Study area relevance	0.2917	0.2500	0.0417	0.5714	
Work experience score	2.8333	2.7500	0.0833	0.4390	
Enrollment in 401(k)	0.6250	0.5132	0.1118	0.1721	
Investment background	0.7431	0.6316	0.1115	0.0841	*
Causal analytics training	0.1667	0.1316	0.0351	0.5518	
Reading accuracy	0.3472	0.1974	0.1499	0.0405	**
Reading speed (secs)	108.3194	108.2632	0.0563	0.9974	
Checking references	0.3611	0.2500	0.1111	0.1438	

*p<0.1, **p<0.05, ***p<0.01.

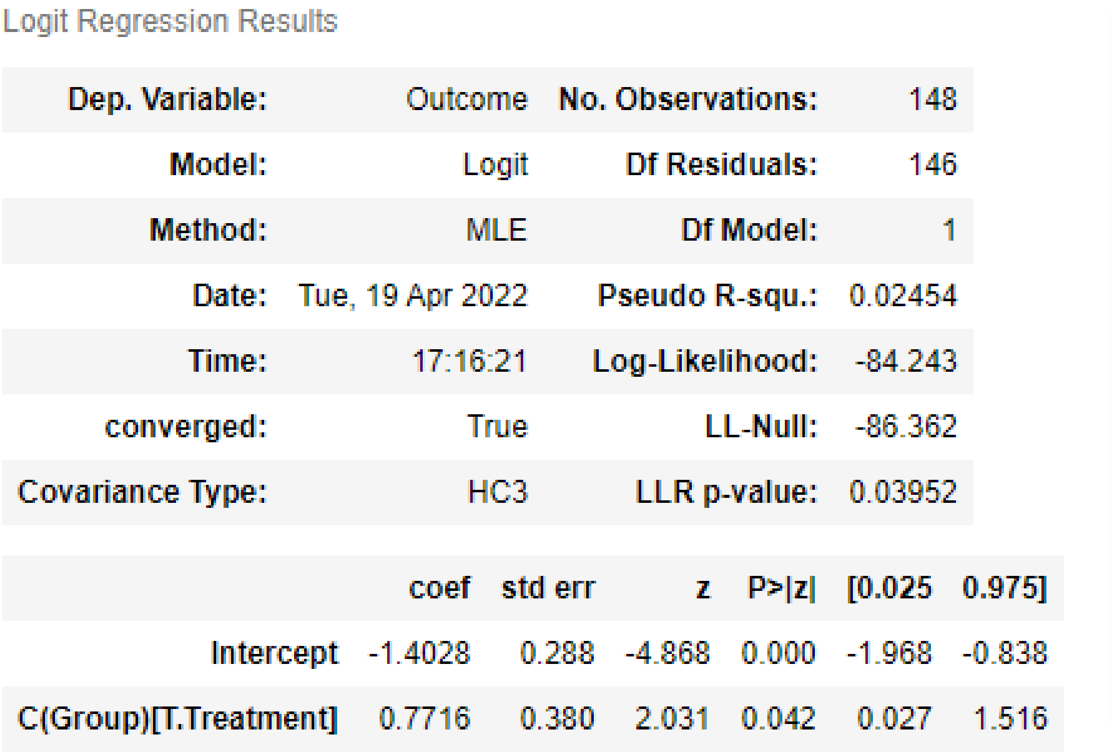
Table 2. ATE by Covariate Block



		ATE	P-value (t-test)
	Variable	Reading accuracy	Reading accuracy
Covariate	Covariate block		
Causal analytics training	High (N=22)	0.2000	0.366
	Low (N=126)	0.1348	0.0804*
Education score	High (N=78)	0.1408	0.1587
	Low (N=46)	0.1417	0.3202
Enrollment in 401(k)	High (N=84)	0.1726	0.0862*
	Low (N=64)	0.1071	0.3255
Investment background	High (N=83)	0.2098	0.0308**
	Low (N=65)	0.0782	0.4932
Study area relevance	High (N=40)	0.1704	0.2512
	Low (N=108)	0.1404	0.0983*
Work experience score	High (N=132)	0.1829	0.0172**
	Low (N=16)	-0.1250	0.6186

Notes: *p<0.1, **p<0.05, ***p<0.01. If N < 61, the sample lacks statistical power.

Table 3A. Regression Model 1: Intercept + Treatment Only

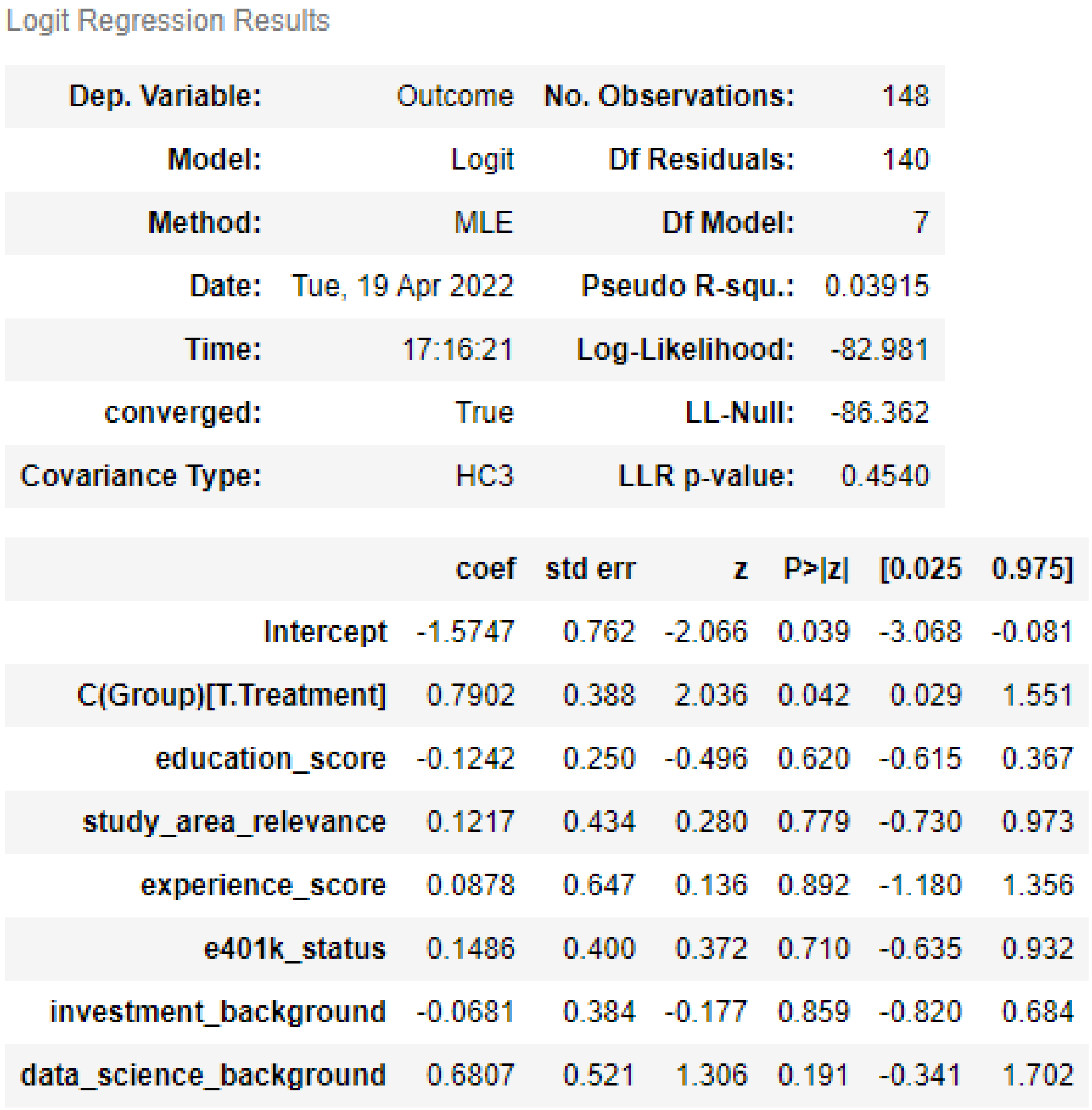


Logit Regression Results

Dep. Variable:	Outcome	No. Observations:	148
Model:	Logit	Df Residuals:	146
Method:	MLE	Df Model:	1
Date:	Tue, 19 Apr 2022	Pseudo R-squ.:	0.02454
Time:	17:16:21	Log-Likelihood:	-84.243
converged:	True	LL-Null:	-86.362
Covariance Type:	HC3	LLR p-value:	0.03952

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.4028	0.288	-4.868	0.000	-1.968	-0.838
C(Group)[T.Treatment]	0.7716	0.380	2.031	0.042	0.027	1.516

Table 3B. Regression Model 1: Intercept + Treatment + Covariates



Logit Regression Results

Dep. Variable:	Outcome	No. Observations:	148
Model:	Logit	Df Residuals:	140
Method:	MLE	Df Model:	7
Date:	Tue, 19 Apr 2022	Pseudo R-squ.:	0.03915
Time:	17:16:21	Log-Likelihood:	-82.981
converged:	True	LL-Null:	-86.362
Covariance Type:	HC3	LLR p-value:	0.4540

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.5747	0.762	-2.066	0.039	-3.068	-0.081
C(Group)[T.Treatment]	0.7902	0.388	2.036	0.042	0.029	1.551
education_score	-0.1242	0.250	-0.496	0.620	-0.615	0.367
study_area_relevance	0.1217	0.434	0.280	0.779	-0.730	0.973
experience_score	0.0878	0.647	0.136	0.892	-1.180	1.356
e401k_status	0.1486	0.400	0.372	0.710	-0.635	0.932
investment_background	-0.0681	0.384	-0.177	0.859	-0.820	0.684
data_science_background	0.6807	0.521	1.306	0.191	-0.341	1.702

REFERENCES

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. In: Oxford University Press Oxford, UK.

Choudhry, A., Sharma, M., Chundury, P., Kapler, T., Gray, D. W., Ramakrishnan, N., & Elmqvist, N. (2020). Once upon a time in visualization: Understanding the use of textual narratives for causality. *IEEE transactions on visualization and computer graphics, 27*(2), 1332-1342.

Garcia-Retamero, R., & Hoffrage, U. (2006). How causal knowledge simplifies decision-making. *Minds and Machines*, *16*(3), 365-380.

GoPeaks. (2021). Text2CausalGraph. *Accessible at* *https://drive.google.com/file/d/1_sCfEqC3fUt7cVicnbbAO4CP964pStrv/view?resourcekey*.

GoPeaks. (2022a). Knowledge graph-based research synthesis. *Accessible at* *https://www.gopeaks.org/kg-based-research-synthesis*.

GoPeaks. (2022b). meta2causalgraph. *Accessible at* *https://github.com/GoPeaks-AI/meta2causalgraph*.

Jaimini, U., & Sheth, A. (2022). CausalKG: Causal Knowledge Graph Explainability using interventional and counterfactual reasoning. *arXiv preprint arXiv:2201.03647*.

Kirschenbaum, S. S., & Arruda, J. E. (1994). Effects of graphic and verbal probability information on command decision making. *Human Factors, 36*(3), 406-418.

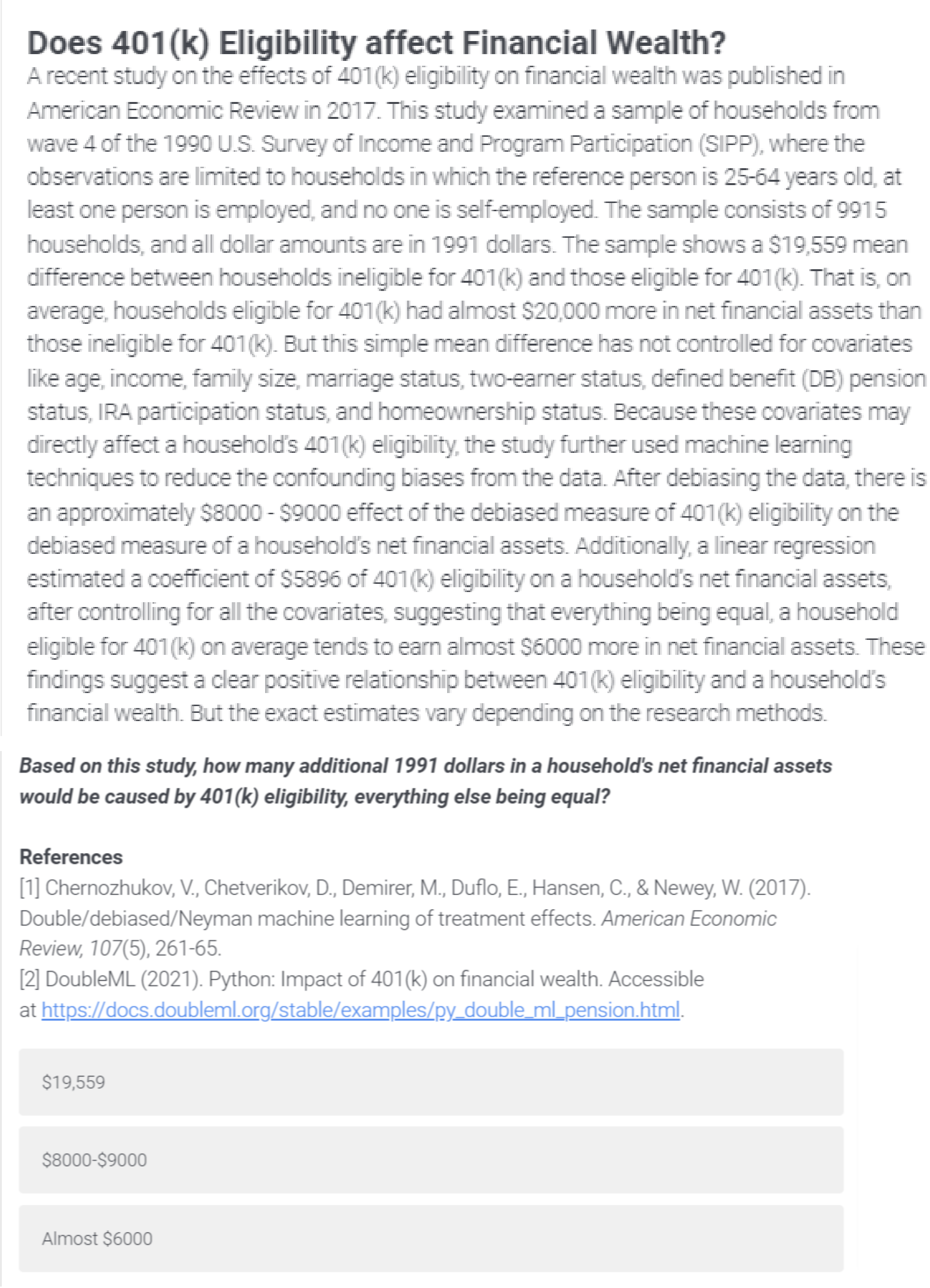
Lurie, N. H., & Mason, C. H. (2007). Visual representation: Implications for decision making. *Journal of Marketing, 71*(1), 160-177.

Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*: Basic books.

Serdar, C. C., Cihan, M., Yücel, D., & Serdar, M. A. (2021). Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia medica, 31*(1), 27-53.

Viegas, F. B., & Wattenberg, M. (2006). Communication-minded visualization: A call to action. *IBM Systems Journal, 45*(4), 801.

Appendix 1A. Reading Task for Readers in the Control Group



Does 401(k) Eligibility affect Financial Wealth?

A recent study on the effects of 401(k) eligibility on financial wealth was published in American Economic Review in 2017. This study examined a sample of households from wave 4 of the 1990 U.S. Survey of Income and Program Participation (SIPP), where the observations are limited to households in which the reference person is 25-64 years old, at least one person is employed, and no one is self-employed. The sample consists of 9915 households, and all dollar amounts are in 1991 dollars. The sample shows a $19,559 mean difference between households ineligible for 401(k) and those eligible for 401(k). That is, on average, households eligible for 401(k) had almost $20,000 more in net financial assets than those ineligible for 401(k). But this simple mean difference has not controlled for covariates like age, income, family size, marriage status, two-earner status, defined benefit (DB) pension status, IRA participation status, and homeownership status. Because these covariates may directly affect a household's 401(k) eligibility, the study further used machine learning techniques to reduce the confounding biases from the data. After debiasing the data, there is an approximately $8000 - $9000 effect of the debiased measure of 401(k) eligibility on the debiased measure of a household's net financial assets. Additionally, a linear regression estimated a coefficient of $5896 of 401(k) eligibility on a household's net financial assets, after controlling for all the covariates, suggesting that everything being equal, a household eligible for 401(k) on average tends to earn almost $6000 more in net financial assets. These findings suggest a clear positive relationship between 401(k) eligibility and a household's financial wealth. But the exact estimates vary depending on the research methods.

Based on this study, how many additional 1991 dollars in a household's net financial assets would be caused by 401(k) eligibility, everything else being equal?

References

[1] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/Neyman machine learning of treatment effects. *American Economic Review, 107*(5), 261-65.

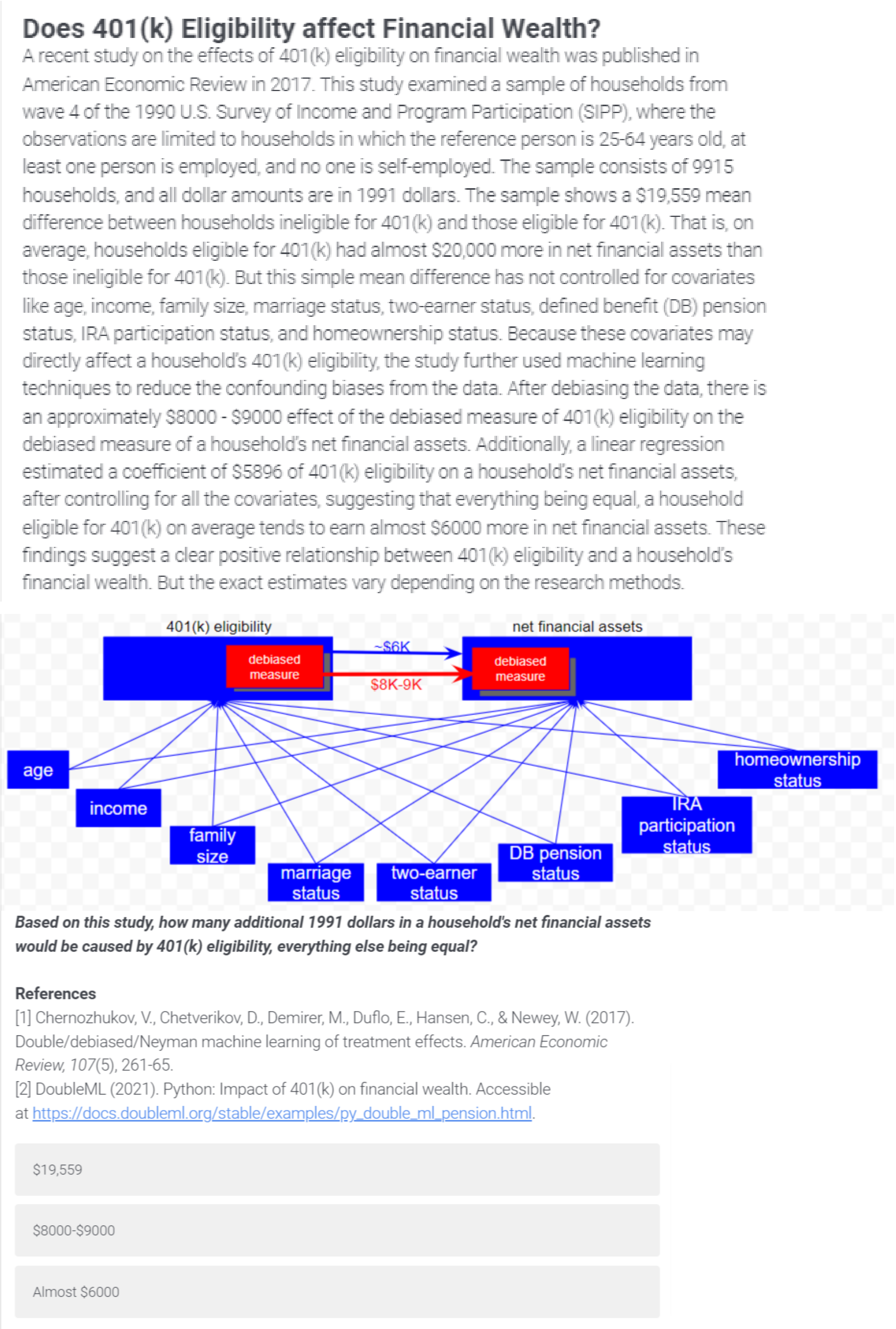
[2] DoubleML (2021). Python: Impact of 401(k) on financial wealth. Accessible at https://docs.doubleml.org/stable/examples/py_double_ml_pension.html.

$19,559

$8000-$9000

Almost $6000

Appendix 1B. Reading Task for Readers in the Treatment Group



Does 401(k) Eligibility affect Financial Wealth?

A recent study on the effects of 401(k) eligibility on financial wealth was published in American Economic Review in 2017. This study examined a sample of households from wave 4 of the 1990 U.S. Survey of Income and Program Participation (SIPP), where the observations are limited to households in which the reference person is 25-64 years old, at least one person is employed, and no one is self-employed. The sample consists of 9915 households, and all dollar amounts are in 1991 dollars. The sample shows a $19,559 mean difference between households ineligible for 401(k) and those eligible for 401(k). That is, on average, households eligible for 401(k) had almost $20,000 more in net financial assets than those ineligible for 401(k). But this simple mean difference has not controlled for covariates like age, income, family size, marriage status, two-earner status, defined benefit (DB) pension status, IRA participation status, and homeownership status. Because these covariates may directly affect a household's 401(k) eligibility, the study further used machine learning techniques to reduce the confounding biases from the data. After debiasing the data, there is an approximately $8000 - $9000 effect of the debiased measure of 401(k) eligibility on the debiased measure of a household's net financial assets. Additionally, a linear regression estimated a coefficient of $5896 of 401(k) eligibility on a household's net financial assets, after controlling for all the covariates, suggesting that everything being equal, a household eligible for 401(k) on average tends to earn almost $6000 more in net financial assets. These findings suggest a clear positive relationship between 401(k) eligibility and a household's financial wealth. But the exact estimates vary depending on the research methods.

Based on this study, how many additional 1991 dollars in a household's net financial assets would be caused by 401(k) eligibility, everything else being equal?

References

[1] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/Neyman machine learning of treatment effects. *American Economic Review, 107*(5), 261-65.
[2] DoubleML (2021). Python: Impact of 401(k) on financial wealth. Accessible at https://docs.doubleml.org/stable/examples/py_double_ml_pension.html.

$19,559

$8000-$9000

Almost $6000

Appendix 2. Qualtrics Survey Questions

Section I. Background of the Respondent

1a. Your highest education level is (single choice):

[ ] < Undergraduate

[ ] Undergraduate

[ ] Master’s

[ ] > Master’s

1b. What is (are) your study area(s)? (multiple choice)

[ ] Business, Economics, and Finance

[ ] Statistics, Data Science

[ ] Other STEM

[ ] Others

3. What is your professional experience? (single choice)

[ ] No work experience (<1 year)

[ ] 1-3 years, including internships

[ ] 4-5 years, including internships

[ ] >5 years, including internships

4a. Are you currently enrolled in 401(k)? (single choice)

[ ] Yes, I am.

[ ] No, I am not.

4b. Do you have any investment experience? (single choice)

[ ] No, I don’t have any

[ ] Yes, but I primarily rely on professional investors to help me invest.

[ ] Yes, I primarily make investments myself.

5. Do you have any formal training in data science focusing on experiments or causal inference? (single choice)

[ ] Yes

[ ] No

6a. Have you studied any of the following subjects in school? (multiple choice)

[ ] Business and Management

[ ] Economics and Strategy

[ ] Finance and Accounting

[ ] Statistics and Math

[ ] Data Science and Information

[ ] Other Science, Technology, Engineering, and Math Subjects

[ ] Others

6b. How many years of postsecondary education have you completed? (single choice)

[ ] > 6 years

[ ] 4 – 6 years

[ ] < 4 years

7. Randomizer (see Appendix 1)

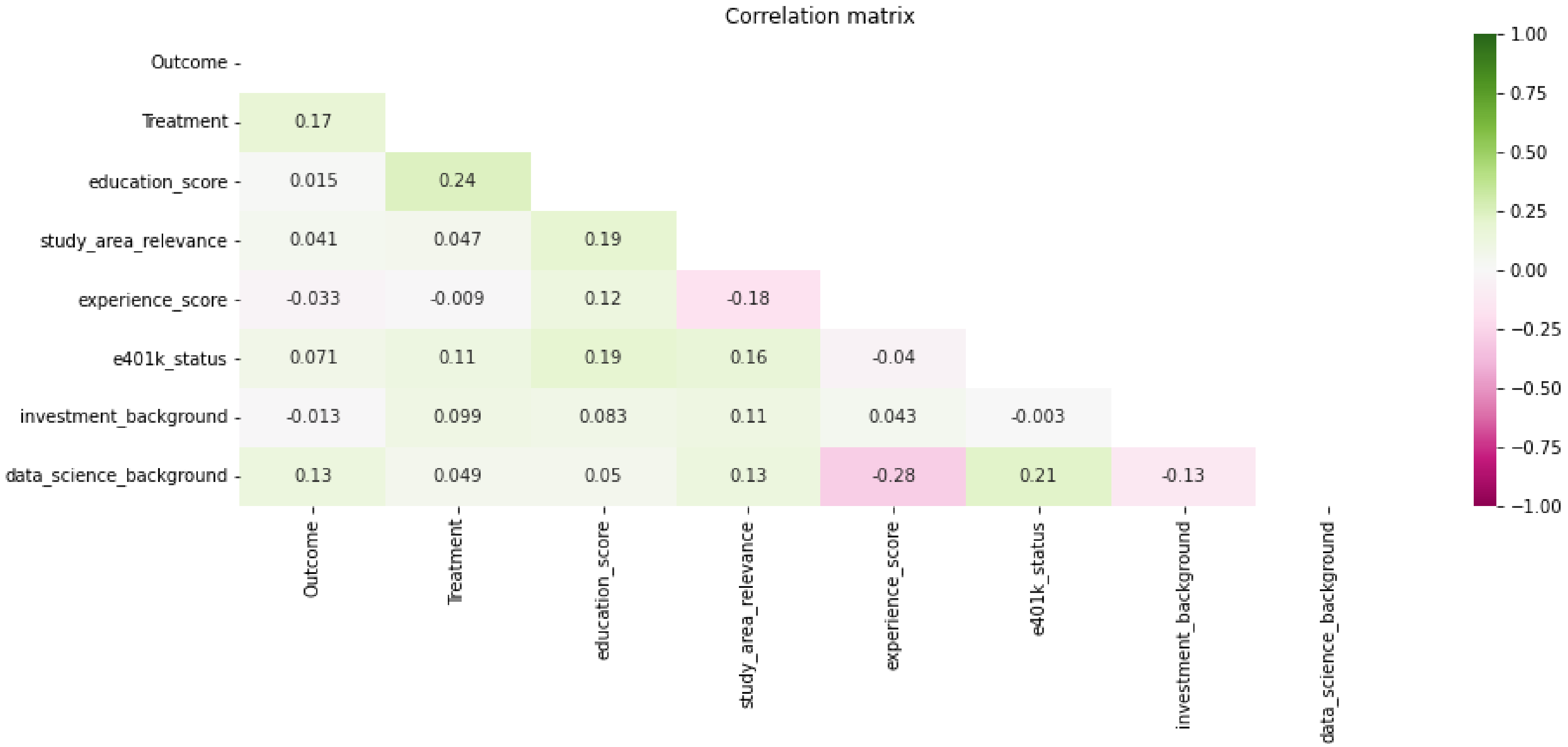
8. Did you visit the listed references to help you comprehend the reading task?

[ ] Yes

[ ] No

Note: 6a and 6b were repeated questions of 1b and 1a respectively under a different framing.

Appendix 3. Correlation Matrix of All Variables in Regression Model 2



Correlation matrix

	Outcome	Treatment	education_score	study_area_relevance	experience_score	e401k_status	investment_background	data_science_background
Outcome								
Treatment	0.17							
education_score	0.015	0.24						
study_area_relevance	0.041	0.047	0.19					
experience_score	-0.033	-0.009	0.12	-0.18				
e401k_status	0.071	0.11	0.19	0.16	-0.04			
investment_background	-0.013	0.099	0.083	0.11	0.043	-0.003		
data_science_background	0.13	0.049	0.05	0.13	-0.28	0.21	-0.13	