Motivating Altruistic Behavior

Sumedh Shah, Devashish Kulkarni, Ryan Wilson
W241 Final Project, Fall 2022

Abstract

We call an act altruistic when it is a sacrifice that benefits others. We aim to experimentally investigate if altruistic behavior can be motivated by outlining specific benefits for others of an act. In our experiment, we recruit participants through Prolific to solve 5 puzzles in a limited time, with a financial reward for correctly completing all the puzzles. The treatment group is provided information that outlines the benefits of the task for elementary and middle school students, while the control group is provided no such information. Outcomes are measured as the number of puzzles attempted and the accuracy (measured as the fraction of correctly completed puzzles). We found no significant effect of the treatment on the outcome variables on average, however, we observed an increased rate of completion of all 5 puzzles in the control group and an increased puzzle accuracy in the treatment group. Presumably, the perceived altruistic benefit of the task caused participants in the treatment group to take greater care in answering the puzzles, while participants in the control group were focused on completing all the puzzles for the financial reward.

Background

The human behavior of altruism towards strangers represents a huge anomaly in the animal world[1]. In most animal species, altruistic behavior in the form of cooperation is often restricted to small groups, largely within a kin. Several explanations of the evolution of altruism suggest that psychological mechanisms, such as emotions, promote altruistic behaviors, rather than conscious calculation of personal gain[2]. Several studies[3] have been performed on altruistic behavior in the form of prisoner's dilemma games, public goods games and ultimatum games. Our experiment hopes to measure if a statement on the benefit of performing a puzzle solving task is effective in nudging the participants to, either complete more puzzles in a limited time, or show better performance at solving the puzzles (measured as accuracy).

Research Question and Hypothesis

The research question we hope to answer through this study is - Can we motivate altruistic behavior by outlining clear benefits for others? We hypothesize that participants who are provided with altruistic

[1] Fehr, Ernst, and Urs Fischbacher. "The nature of human altruism." *Nature* 425.6960 (2003): 785-791.
[2] Roberts, S. Craig, ed. *Applied evolutionary psychology*. Oxford University Press, 2012.
[3] Andreoni, James, William T. Harbaugh, and Lise Vesterlund. "Altruism in experiments." *Behavioral and experimental economics*. Palgrave Macmillan, London, 2010. 6-13.

survey instructions will attempt more puzzles than participants who receive general survey instructions. Age, sex and ethnicity of the participants are used as covariates. While we are controlling for three different demographic features, we primarily hypothesize that those of Eastern ethnicities (specifically Asian) will complete more puzzles than those of Western ethnicities because of the collectivist and familial nature of Eastern cultures. In addition, we hope to investigate if the performance of the participants in the puzzle solving task improves as a result of the altruistic instructions, measured in terms of accuracy.

Experimental Design

The design for this experiment is a posttest-only control group design. In this design, the subjects are randomly selected to be in the control or treatment group (R). The control and treatment groups varied in the set of instructions provided to them (X). The effect of the treatment is measured by the average number of puzzles completed and accuracy of puzzles completed between the two groups (O). The experiment design using RO notation is shown in Table 1.

Table 1: Experimental Design

	Randomization (R)	**Treatment (X)**	**Observation (O)**
Subject Pool	R_1	X	O_1
	R_2		O_2
Notes:	**Independent Variable: Variation of Instructions** **Dependent Variable: Number of Questions Completed, Accuracy**		

Survey/Randomization

The survey used for this experiment was hosted on Qualtrics, an online survey platform. Randomization of the subject pool into treatment and control groups was handled by the survey platform.

Puzzles

We ask participants to complete five "Factor 24" puzzles in 4 minutes. Four 1-digit numbers are provided to the participants, who are required to use addition, subtraction, multiplication and/or division operations on the provided numbers to get a total of 24. The puzzles were designed to have an increasing level of difficulty. The time limit for the task was set at 4 minutes, based on pilot survey data collected from the authors' family and friends. An example puzzle, with the expected response was shown to the participants before the start of a countdown timer. All the puzzles were shown on the same webpage, allowing the participants to answer the questions in any order they wished. The puzzles used in the survey and the answers to the puzzles are shown in Table 2.

Table 2: Puzzles and Answers

Puzzle	Answer
8, 8, 7, 1	8 + 8 + 7 + 1
2, 6, 8, 2	2 + 6 + (2 * 8)
3, 3, 6, 5	3 + 6 + (5 * 3)
11, 5, 2, 4	11 + 5 + (2 * 4)
7, 2, 6, 8	(7 - 6 + 2) * 8

Treatment

The treatment is provided through a variation in instructions provided at the start of the survey. A statement on the benefit of the survey to elementary and middle school students is included in the treatment instructions, and no such statement is included in the control instructions. The instructions used in this experiment are shown in Figure 1.

Figure 1: Treatment

Control Prompt

Thank you for participating in our survey. We are asking that you complete up to 5 puzzles in 4 minutes. If you correctly answer all 5 puzzles, you will be entered into a raffle to win a $50 Amazon gift card.

Treatment Prompt

Thank you for participating in our survey. We work in ed-tech and are developing a natural language processing math app for elementary and middle school mathematics. Your responses will help us train our algorithm to help provide real time feedback to students as they work on our platform. We are asking to complete up to 5 puzzles in 4 minutes. Answering all 5 puzzles will provide us a diverse set of responses to provide meaningful feedback to the students using our platform. If you answer all 5 puzzles correctly, you will also be entered into a raffle to win a $50 Amazon gift card.

Recruitment/Incentives

Participants were recruited through the online survey platform Prolific. Each participant was paid the recommended rate by Prolific ($12.00/hr) to participate in and complete the survey. The survey required about 5 minutes of the participants' time, resulting in a payment of ~$1.00 per participant. In order to incentivize the participants to complete all the puzzles correctly, a raffle for $50 Amazon gift cards was included for participants that answered all the puzzles correctly. Based on the budget available for this study, 300 participants were recruited into the study and two $50 Amazon gift cards were raffled off.

Power Calculation

Prior to conducting the experiment, a power analysis was performed to determine the sample size on desired effect size. Three different scenarios were developed to estimate the power of the observations of

the experiment. Three different scenarios were considered with varied distributions of the outcome variables (number of puzzles completed) for the treatment and control groups, as follows:

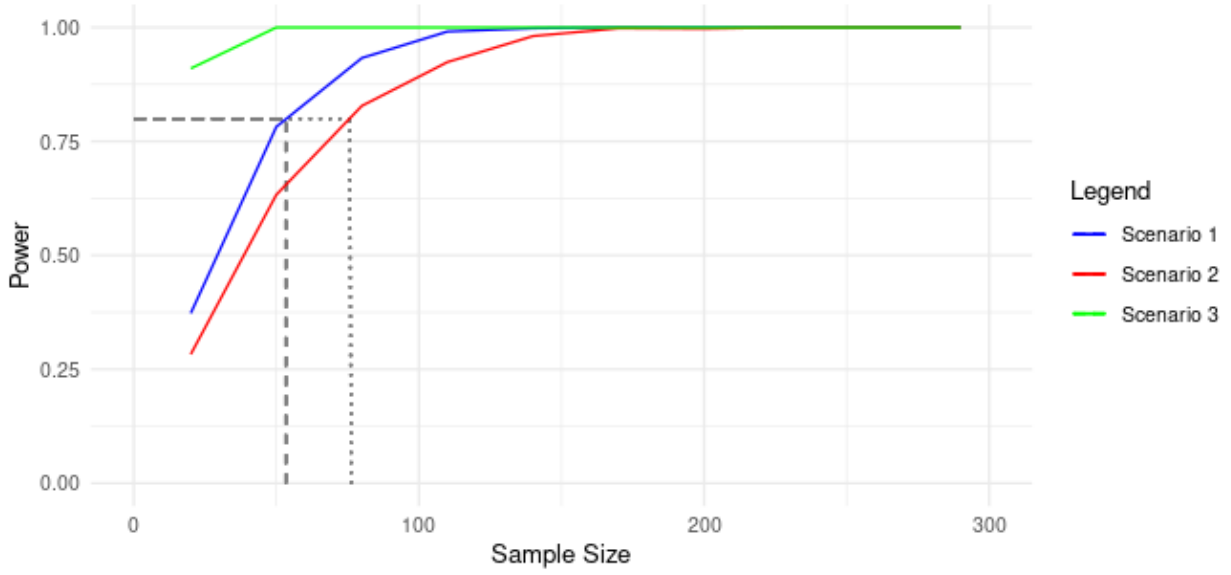
- **Scenario 1**: Low effect size (0.5), low dispersion
    - Control group: 10% fails to solve the first puzzle, 80% solve only 1 puzzle, 10% solve 2 puzzles. (mean = 1)
    - Treatment group: 10% fails to solve the first puzzle, 40% solve only 1 puzzle, 40% solve 2 puzzles, 10% solve 3 puzzles. (mean = 1.5)

- **Scenario 2**: Low effect size (0.5), high dispersion
    - Control group: 10% fails to solve the first puzzle, 40% solve only 1 puzzle, 40% solve 2 puzzles, 10% solve 3 puzzles. (mean = 1.5)
    - Treatment group: 10% fails to solve the first puzzle, 20% solve only 1 puzzle, 40% solve 2 puzzles, 20% solve 3 puzzles, 10% solve 4 puzzles. (mean = 2)

- **Scenario 3**: High effect size (1.5), high dispersion
    - Control group: 10% fails to solve the first puzzle, 40% solve only 1 puzzle, 40% solve 2 puzzles, 10% solve 3 puzzles. (mean = 1.5)
    - Treatment group: 0% fails to solve the first puzzle, 10% solve only 1 puzzle, 20% solve 2 puzzles, 40% solve 3 puzzles, 20% solve 4 puzzles, 10% solve 5 puzzles. (mean = 3)

A plot of statistical power with the sample size for each of the three scenarios is shown in Figure 2. A sample size of at least 75 participants is required to generate statistical power of 0.8 for the worst case scenario.

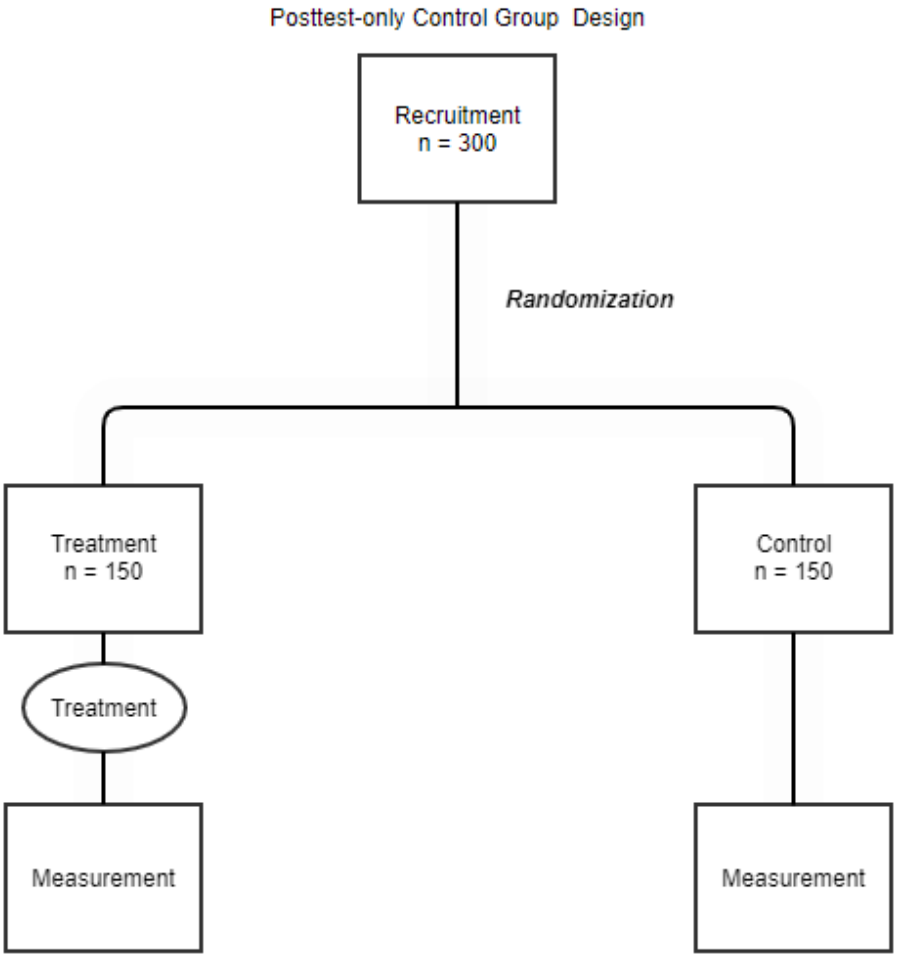
Figure 2: Statistical Power



To summarize, the goal of the experiment is to investigate if providing the benefits of research will elicit a response in the participants in answering more questions. Participants are recruited through Prolific and

randomly assigned to treatment and control groups by Qualtrics. Participants in each group are asked to complete five puzzles, with only the introductory text being varied between the two groups. At the conclusion of the survey, the average difference in the number of questions completed and accuracy of questions completed between the two groups were calculated. Figure 3 represents the flow of the experiment in terms of recruitment, randomization, treatment, and measurement.

Figure 3: Experiment Flow Diagram



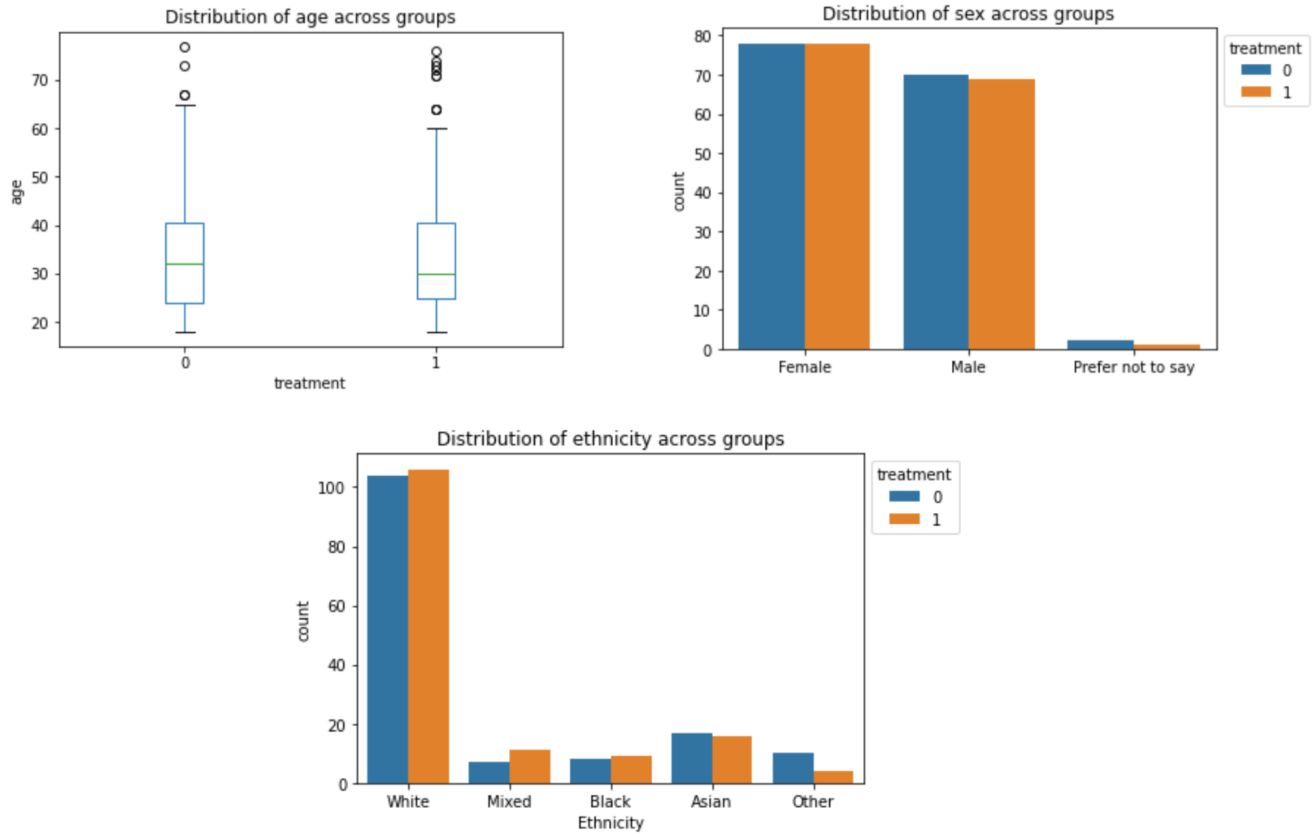
Results

The survey was published on Prolific at ~6 PM PST on Thursday, Nov 17th 2022. The required number of responses were received within 90 minutes of publishing the survey. The survey responses were manually graded by the authors and a value for number of puzzles completed and accuracy (number correct/number completed) was assigned to each response. In addition, demographic data for each participant, including age, sex and ethnicity was provided by Prolific.

Covariate balance check

Figure 4 shows the distribution of the covariates between the treatment and control groups. The survey participants had a mean age of ~30 years old, with participants as young as 18 years old and some outliers over 70 years old. A slightly larger number of the participants were females as compared to males, with some participants identifying as neither. A large majority of the participants were of white ethnicity, since the subject pool is recruited in the US. Asian was the second largest ethnicity of the participants. The distribution of the covariates is very similar in both treatment and control groups, indicative of a reliable randomization process.

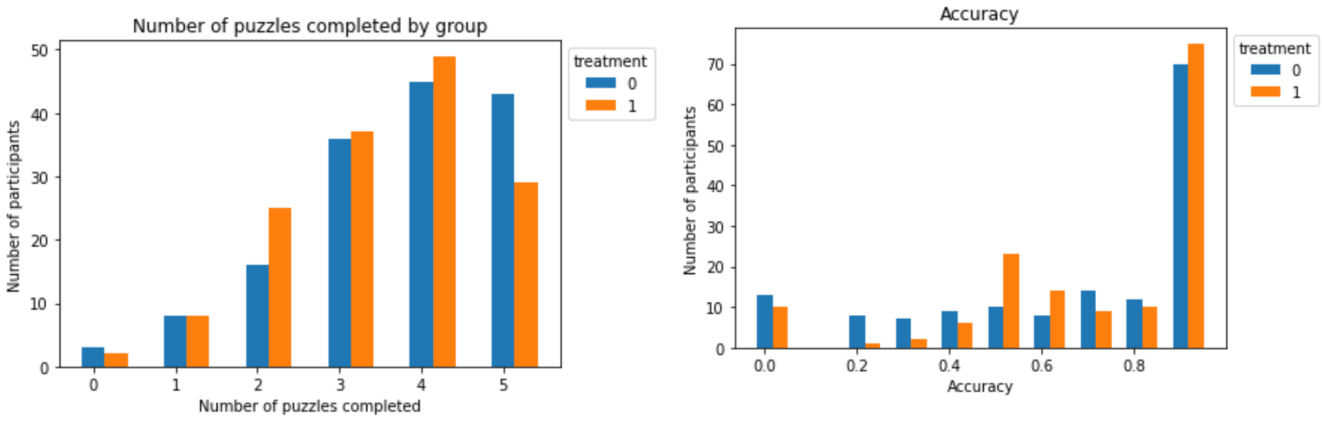
Figure 4: Distribution of covariates



Outcome Distributions

Figure 5 shows the distribution of the outcome variables between the two groups. We see a good spread of counts of participants for each value of number of puzzles completed, including 0 and 5 puzzles completed. Interestingly, more participants in the treatment group completed 2, 3 and 4 puzzles than participants in the control group. However, the reverse is seen for the number of participants completing 5 puzzles, with a higher count for the control group (43) compared to the treatment group (29). A chi-squared test for goodness of fit on these counts yields a p-value of 0.09896, which indicates the probability of observing these counts by chance, assuming an expected split of 50:50 between the groups for participants completing 5 puzzles. If believed to be real, this effect suggests that participants in the control group are more likely to complete all 5 puzzles to be included in the raffle for the financial reward, while participants in the treatment group are more likely to spend more time and care on the initial puzzles, due to their altruistic intention in completing the tasks, and less likely to complete all 5 puzzles. The distribution of accuracy reveals that most participants correctly answered all the puzzles they attempted. A small fraction of participants did not get any puzzle correct. No obvious difference is seen between the groups in terms of accuracy.

Figure 5: Distribution of outcomes



Regressions

Tables 3 and 4 show regression results for each of the outcome variables. For each outcome, three different regression models are presented: one only with a treatment indicator, one with treatment and the covariates, and one with treatment, covariates, and the interaction of treatment with an indicator for Asian ethnicity. Finally, robust standard errors are calculated for each of the regressions. The first model for each outcome indicates whether the treatment, or lack thereof, had any impact on the number of puzzles completed by the participants and the accuracy with which they completed the puzzles. For our covariates, we were primarily interested in people from Eastern cultures, or those of Asian ethnicity, but also include the participant's age and sex. Outside of the Asian ethnicity, all of the other participants' ethnicities were either mixed or from a Western culture and were therefore grouped together. To reiterate, we hypothesize that since Eastern cultures are more collectivistic and giving, participants coming from these cultures (or having Asian ethnicity) would answer more puzzles if they were given the treatment of

an altruistic reason for performing the puzzle solving task. We therefore wanted to investigate the heterogeneous treatment effect of being an Asian that received the treatment.

Table 3 shows the regression results on the number of puzzles completed. The first regression shows an average treatment effect (ATE) of -0.196, or an average of 0.196 less puzzles completed if the participants were in the treatment group, and an average of 3.596 puzzles completed as a baseline. While the treatment coefficient was not statistically significant, the baseline constant is statistically significant ($p < 0.01$) with similar values and p-values across the other two regressions with the covariates and interaction term. The addition of the covariates and interaction term in the next two regressions barely impact the treatment coefficient and don't reduce the standard error. The covariates and interaction term don't explain any other variance in the data, either. In addition, the standard errors are extremely high for all of the coefficients. Because the constant terms are so large and highly statistically significant, they have an overwhelming contribution to the outcome and all other coefficients are negligible in comparison. Finally, contrary to our hypothesis, we observe that participants of Asian ethnicity didn't necessarily answer more or less questions than those of other ethnicities. Furthermore, no heterogeneous treatment effect was observed for those of Asian ethnicity. Because the coefficients were already so small, with high standard errors, including a standard error higher than the interaction term coefficient, any variation of data could've affected the coefficients and the explained variance.

Finally, Table 4 shows the regression results on accuracy. The first regression shows an ATE of 0.043, or an average of 4.3% more accuracy if the participants were in the treatment group, and an average of 0.719 or 71.9% as a baseline. Similar to the previous outcome, the treatment coefficient was not statistically significant while the baseline constant was highly statistically significant ($p < 0.01$) with similar values and p values across the other two regressions with the covariates and interaction term. While impact is still fairly minimal, the addition of the covariates somewhat contributes to the outcome and explains a very small portion of the variance in our data. The covariates, however, didn't impact the treatment coefficient nor reduce its standard error in the second two regressions. Finally, the interaction term with the Asian ethnicity didn't impact our outcome and didn't reduce the standard errors. Similar to our previous outcome and regressions, the coefficients are relatively close to zero compared to the constant terms and the standard errors are extremely high for all of our coefficients. We observe a few more interesting results with these regressions, though. The Asian ethnicity coefficients are 0.088 (8.8% accuracy) and 0.118 (11.8% accuracy) for the second and third regressions and are statistically significant ($p < 0.1$). While we didn't see a noteworthy difference in how many puzzles they completed, we do notice they overall answered puzzles more accurately. However, it also appears from the interaction term that the Asians in treatment answered puzzles slightly less accurately, albeit with a standard error larger than the coefficient. This could be the result of a much smaller set of Asian participants in the study, compared to non-Asians. While it appears that Asians answered the puzzles more accurately, we still don't observe any heterogeneous treatment effect with Asians. Finally, we notice that participants that didn't wish to specify their sex also had statistically significant ($p < 0.1$) coefficients in the second and third regressions (18.1% and 18.2% higher accuracies). This, however, may be a result of the very few data points where people didn't wish to specify their sex, and is likely due to chance.

Table 3: Regression results (Outcome: Number of puzzles completed)

	Dependent variable:		
	num_completed		
	(1)	(2)	(3)
treatment	-0.196	-0.202	-0.169
	(0.143)	(0.145)	(0.156)
ethnicityAsian		-0.086	0.057
		(0.205)	(0.280)
sexMale		0.204	0.205
		(0.145)	(0.145)
sexNaN		-0.916	-0.915
		(2.084)	(2.108)
sexPrefer not to say		0.197	0.203
		(0.446)	(0.439)
age		-0.004	-0.004
		(0.006)	(0.006)
treatment:ethnicityAsian			-0.294
			(0.404)
Constant	3.596***	3.665***	3.647***
	(0.103)	(0.230)	(0.231)
Observations	301	298	298
R2	0.006	0.020	0.022
Adjusted R2	0.003	0.00001	-0.002
Residual Std. Error	1.233 (df = 299)	1.226 (df = 291)	1.227 (df = 290)
F Statistic	1.903 (df = 1; 299)	1.001 (df = 6; 291)	0.916 (df = 7; 290)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 4: Regression results (Outcome: Accuracy)

	Dependent variable:		
	accuracy		
	(1)	(2)	(3)
treatment	0.043	0.043	0.050
	(0.036)	(0.037)	(0.040)
ethnicityAsian		0.088*	0.118*
		(0.052)	(0.069)
sexMale		-0.007	-0.006
		(0.038)	(0.038)
sexNaN		-0.108	-0.108
		(0.470)	(0.466)
sexPrefer not to say		0.181*	0.182*
		(0.098)	(0.098)
age		-0.002	-0.002
		(0.002)	(0.002)
treatment:ethnicityAsian			-0.061
			(0.105)
Constant	0.719***	0.781***	0.778***
	(0.027)	(0.063)	(0.064)
Observations	301	298	298
R2	0.005	0.026	0.027
Adjusted R2	0.001	0.006	0.004
Residual Std. Error	0.315 (df = 299)	0.316 (df = 291)	0.316 (df = 290)
F Statistic	1.423 (df = 1; 299)	1.306 (df = 6; 291)	1.156 (df = 7; 290)
Note:			*p<0.1; **p<0.05; ***p<0.01

Conclusions and Future Enhancements

We did not find a statistically significant effect of including the treatment, or instructions of helping a math application for elementary and middle school students, with the number of puzzles completed nor the accuracy of puzzles completed. The inclusion of demographic covariates did not contribute to either of the outcomes or add any explained variance to the regressions. We also did not observe any heterogeneous treatment effect or different impact from participants of Eastern ethnicities in this study. Even though the results were fairly insignificant, this study shows a proof-of-concept that the inclusion of instructions with an altruistic purpose may have some effect on the accuracy of puzzles completed.

However, we did observe an interesting effect of the treatment on the number of participants in each group that completed all 5 puzzles. This may suggest that the treatment did induce a change in behavior of the participants, but the outcome variables, as defined, may not be adequate to capture its effect. For example, since the benefits of altruism are positive emotional feelings, a question asking the participants to rate their enjoyment of the puzzle solving task may reveal a difference between the groups. A different scoring system, where the difficult puzzles are worth more points, could reveal differences between the behaviors of the two groups' that the current outcome variables did not, for example, if the treatment group answers the difficult puzzles with more accuracy than the control group. A different incentive structure, with a guaranteed reward for correctly completing all puzzles rather than a raffle, may strengthen the 'greedy' motivation of the control participants and amplify the difference in behaviors of the two groups.

There are a few key tweaks we could make in the future to better explore our hypothesis and findings from the study. Because we utilized Prolific for our participant recruitment, we are collecting responses from people that are doing survey after survey and not necessarily reading the instructions, or treatment prompt, very carefully. These people could also just be powering through surveys for the monetary rewards. For a more reliable and thorough experiment, we could recruit random friends, family members, and coworkers to complete the survey and have them complete it in person. The treatment statement might be more impactful if we delivered it in person and then we might see more significant results with the number of puzzles completed and the puzzle accuracy. Finally, further pilot studies need to be completed for the time limit given to complete all the puzzles. We gave a 4-minute time limit and found that this worked fairly well for people to complete puzzles but experimenting with slightly higher time limits would be beneficial for finding a sweet spot between completing puzzles and correctly figuring out the puzzles.