Final Report: Evaluating the Impact of Rewards on Donations

W241 Section 01 | Summer 2025

Group 1: David Carlson, Jackie Loyola, Kristen Lin, Yiwen Hou, Yoko Morishita

Abstract

Charitable giving is often thought to stem from altruistic intent, yet the growing use of incentives to encourage donations has raised questions about how rewards influence donor behavior. Previous studies have highlighted a positive association between small, tangible rewards (e.g. gifts or public recognition) and increased donation rates, but much of the existing evidence stems from correlational research, making it difficult to disentangle motivation from self-selection. To investigate this relationship more rigorously, we implemented a theoretical survey in which participants, grouped by income and age, were randomly assigned either to control or one of two treatment groups. Our aim was to measure how the presence of a potential reward and the magnitude of the reward influenced both the likelihood and size of a donation.

Background

Offering a reward in exchange for a donation is predicted to increase giving behavior because it activates principles of reciprocity and perceived value, both well-documented in behavioral science. Cialdini (2009) outlines how reciprocity norms compel individuals to give back when they receive something, even if unsolicited. Additionally, Kessler and Milkman (2018) demonstrated that including small incentives (like gifts or thank-yous) in charitable campaigns can significantly increase donation rates. Studies in nonprofit marketing have also shown that tangible acknowledgments enhance feelings of appreciation and social identity, which in turn encourage charitable behavior (Winterich et al., 2013). Therefore, offering a small digital reward is grounded in empirical findings and is expected to increase both the likelihood and frequency of donations. The studies also suggest that participants in the treatment group were significantly more likely to donate, with higher donation frequencies and slightly larger average amounts given. Notably, younger individuals and those in middle-income brackets responded most strongly to the incentive. These findings imply that while intrinsic motivation remains a key factor, the presence of a reward may act as a practical nudge, particularly for certain demographic groups.

Research Question

For this study, we attempted to answer the question: *Are people more likely to donate money if given a reward in return?* The null hypothesis poses that (the possibility of) receiving a reward does not increase the likelihood for an individual to donate. However, based on prior research, we lean towards the alternative hypothesis and predict that (1) a potential monetary incentive and (2) a higher likelihood of receiving the incentive will increase the amount donated.

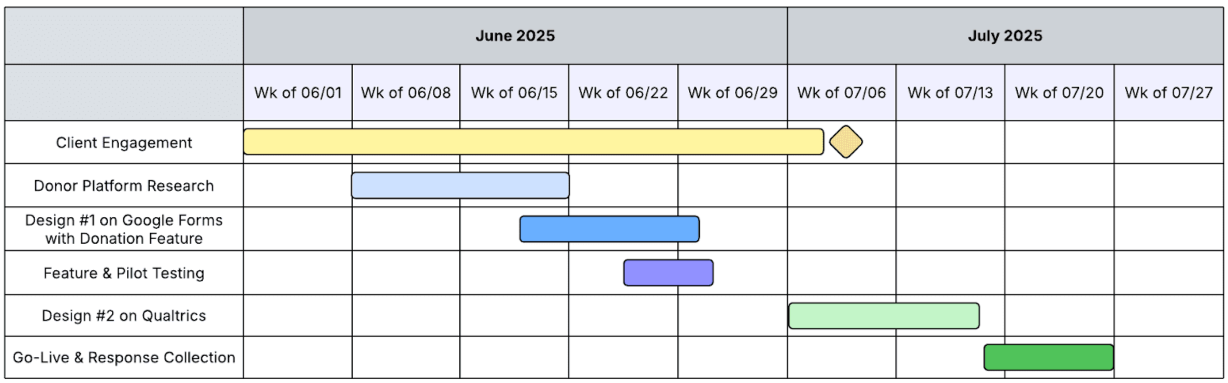
Experiment Design

Overview

We first attempted to design the experiment in collaboration with a client who had a large following of over 300,000 individuals over social media. We planned to initiate a donation campaign designed to support animal welfare initiatives for the client's organization. As part of the campaign strategy, we leverage a randomized link that would distribute two versions of a donation form for a difference-in-difference analysis. Although the client was initially enthusiastic, as we neared the execution stage, the client abruptly disengaged and no longer continued to communicate with us.

Due to unforeseen circumstances and having a deadline that left us with a little over two weeks, we had to salvage our study by approaching it from a different direction, in which we designed a theoretical survey on Qualtrics where baseline data would be collected in the first 10 days, and the remaining time would be spent for the core analysis. While we initially planned to continue with the animal welfare initiative, we believed that it would have been difficult to be able to encourage individuals to donate to our cause without a reputation to back us.

Overall Project Timeline



Survey Tooling

Google Forms was initially considered a viable option for collecting participant and survey data. While we were able to simulate the randomized distribution of theoretical donation surveys, our mock-up presented limitations in tracking sessions and native functionality for monitoring partial survey completions. We tried creating a separate audit log to capture link access events with timestamps, but this only added more complications in stitching the data as the timestamps proved to be unreliable if more than one user was accessing the survey at the same time.

We used our learnings from the Google Forms pilot study to inform our decision on using Qualtrics as our form distributor and randomizer. Through this platform, we were able to design our survey with ease and track how far along a user was able to complete the survey. The survey was natively designed to be compatible with both personal computers and devices, making it more accessible to our social media target audience.

Because our survey was designed to be publicly accessible from any platform, we also had to consider possible exposure to bad actors that may try to disrupt or manipulate our survey results. Within the first few hours, we noticed a highly improbable number of participants (>2000) that were answering our survey and integrated preliminary and post bot detection and traffic analysis shortly after to distinguish between botting behaviors and legitimate human participants. These features helped to flag suspicious behavior and contributed to the overall integrity of our feedback collection process.

Enrollment and Recruitment Process

To maximize participant engagement and survey results within a limited time frame, our team adopted an enrollment model that integrated both recruitment and participation at the same time. Because these platforms would have unpredictable user behaviors, we knew that it would have been more difficult and time-intensive to set up a recruitment period followed by an enrollment process. We also did not want to have to fight against the platforms' algorithms, and therefore extended participation to any individual who encountered our postings and clicked the survey link. Our messaging for these platforms was relatively consistent across the platforms:

> **Help Us Support Animals in Need 🐶🐱 1-Min Survey + Chance to Win $50 Amazon Gift Card**
>
> SURVEY
>
> We're UC Berkeley students conducting a research survey to understand how to better support animal welfare, but **we're still short at least 100 participants** and need your help **today**.
>
> The survey is **anonymous**, takes just **1–2 minutes**, and you can enter a raffle to win **one of ten $50 Amazon gift cards**.
>
> Whether or not you have pets, **your honest answers are critical** to help us get accurate, meaningful results.
>
> Please take a minute to support this important cause and help us cross the finish line 🐾
>
> Link: https://berkeley.qualtrics.com/jfe/form/SV_0N9SCg48xIsvDme

The landing page of our study provided details about our study and data management, time commitment, potential reward, and opportunity to opt out at any time. We considered those who proceeded beyond the initial page of the survey as a voluntary, consenting participant:

> **Give a Paw: Share Your Support for Animals!** 🐾
>
> We're UC Berkeley students conducting a survey for an educational project to better understand charitable giving. Your responses are completely **anonymous and confidential** and will help us learn how to encourage support for animals and pets in need. This survey is open to everyone, whether or not you own a pet, and **takes only 1–2 minutes to complete**. Participation is entirely voluntary, and you may opt out at any time.
>
> As a thank you, participants have a chance to win **one of ten $50 Amazon gift cards (US only) in a raffle**. We appreciate your time and perspective!
>
> *Your privacy matters. We will not be collecting any personally identifiable information, and your data will never be sold or shared for commercial purposes.*

Additionally, we attached a $50 lottery incentive for those who chose to participate and complete the theoretical survey. We distributed the survey across friends/families and multiple social media platforms, including several Facebook pages, X (Twitter), various sub-Reddits, and the Berkeley MIDS Slack. As a note, some sub-Reddits considered our survey as marketing, which violated their rules and led to the survey being removed from their pages.

Treatment Details and Randomization

To assess how the treatment of reward-based incentives would impact donation behavior, we set up the survey to randomize the final question of the survey between three scenarios at the moment the participant navigated to the question.

Control Group: Participants would receive a neutral question with no mention of incentives.

“If you had $100, how much would you be willing to donate to help animals/pets in need?”

Treatment (Version 1): Participants were presented with an offer to enter a lottery for a fully-expensed trip in exchange for their donation. This version would help measure whether the presence of a reward would motivate donor behavior.

*“If you had $100, how much would you be willing to donate to help animals/pets in need, **knowing that any donation would enter you into a raffle to win a fully paid trip of your choice** (one entry per donor)?”*

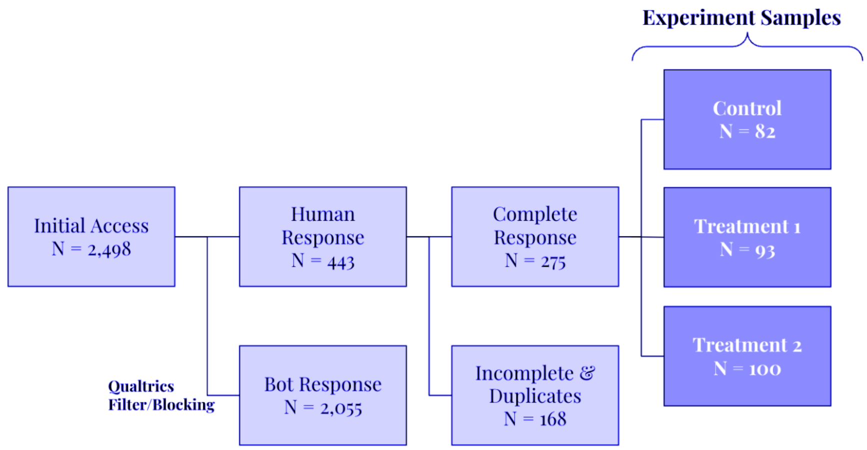
Treatment (Version 2): Building on treatment 1, this version introduced a scaled incentive structure where participants could increase their odds of winning the lottery, depending on the $ amount donated. Treatment 2 would help measure whether increasing the probability of receiving an incentive would further motivate donor behavior and donation amounts.

*“If you had $100, how much would you be willing to donate to help animals/pets in need, **knowing that every $10 you donate would enter you into a raffle to win a fully paid trip of your choice** (for example, $50 = 5 entries)?”*

Given the potential diversity in participant background and the theoretical nature of the survey, we capped donation amounts to help prevent any opportunity to exaggerate or inaccurately report donation amounts. The maximum allowable donation amount was set to $100, and this limit helped to reduce outliers while ensuring a more realistic range of donated amounts. We deliberately set this question to a dynamic slider format instead of a free-text numerical field to guide our participants to answer within this range. Overall, our design elements were selected carefully to fetch a more accurate analysis of donor behavior under varying incentive conditions.

During the data collection process, a Reddit participant attempted to exploit the survey to maximize their chances of receiving the $50 gift card incentive from the participation lottery. We leveraged Qualtrics’ built-in bot detection tools to automatically flag the 2,055 bot-generated responses. After removing the bot responses, we had collected a total of 443 responses (Figure 1). To monitor participant engagement, we implemented a progress tracker in Qualtrics to record how far each respondent advanced through the survey. We also included a hidden question at the beginning of the survey, which would autofill after a 0.1-second timer elapsed. Combined with the progress tracker, this mechanism allowed us to track if a user clicked on the survey and whether they exited before starting the survey. In most cases, participants clicked the survey link but exited from the landing page without answering any questions. We then manually combed through the 443 records to remove 168 incomplete surveys or duplicate participant responses. Ultimately, we were left with 275 valid responses, but because of the previous filtering measures and precautions taken to preserve data quality, the final sample resulted in unbalanced control and treatment groups: 82 to the control group, 93 to treatment 1, and 100 to treatment 2. We therefore conducted a randomization check to examine whether pre-treatment conditions were balanced among the groups, which will be discussed in a later section.

Figure 1: Randomization Flow



Outcome Measure and Covariates

Our outcome was a continuous variable: the USD value of the donation, ranging from \$0 to \$100. We hypothesized that demographic factors such as residential state, gender, age, pet ownership, and donation history could influence donation behavior. Therefore, we collected these pre-treatment covariates to examine whether randomization achieved balance between groups and to include them as control variables in our regression analyses.

Statistical Analysis

The primary objective was to estimate the causal effect of two treatment types on the amount donated by using linear regression models. To assess whether adding certain pre-treatment covariates increased the precision of our estimates, we compared a short model (without covariates) and a full model (including the controls). We evaluated whether these covariates served as 'good controls' by examining their impact on the model precision. While the full model potentially includes all of the covariates, we excluded interaction terms to avoid overfitting and unnecessary model complexity.

<u>Short model</u>

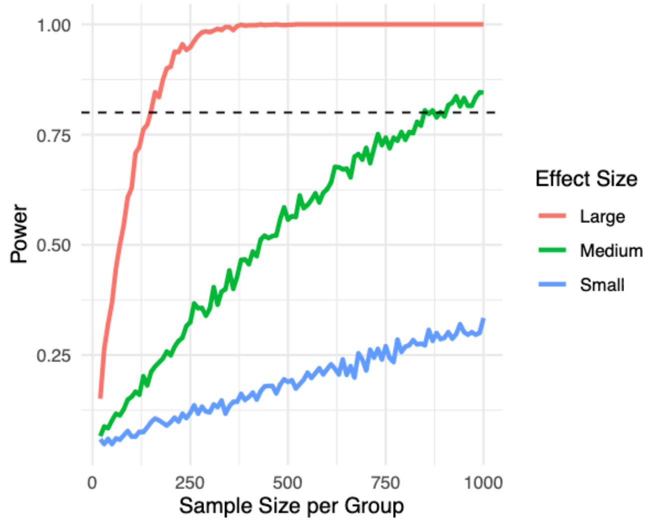
$$\text{Outcome} = \beta 0 + \beta 1 * \text{Treatment}$$

<u>Full model</u>

$$\text{Outcome} = \beta 0 + \beta 1 * \text{Treatment} + \beta 2 * \text{State} + \beta 3 * \text{Income level} + \beta 4 * \text{Age} + \beta 5 * \text{Gender} + \beta 6 * \text{Pet ownership} + \beta 7 * \text{Donation history}$$

Our initial hypothetical treatment effect was estimated to be a 5 - 25% increase in the donation amount. Pre-experiment power analysis indicates that a sample size of 100 per group would yield approximately 80% statistical power to detect a large effect (25% increase), 20% power for a medium effect (10% increase), and 10% power for a small effect (5% increase).

Figure 2: Pre-treatment Power Analysis



* Effect Size: Large (25%), Medium (10%), Small (5%)

Results

Raw Data and Data Clean-up

Data pre-check - The survey responses were monitored on the backend on a daily basis. During the experience period, there were two instances where the response rate increased significantly. After carefully reviewing the backend data—such as Meta Version, Resolution, Operating System, Browser, etc.—we were able to identify and exclude bot-generated responses directly through the Qualtrics backend. We then removed incomplete responses and duplicate entries based on email account identifiers, resulting in a final sample of 275 responses.

Data modification - We preprocessed the dataset by converting relevant variables into categorical (factor) types and creating derived variables for analysis. Demographic variables such as state, gender, age, income, and pet ownership were re-coded into categorical formats. Donation history variables were converted into logical values. Binary indicators were created to represent assignment to control or treatment groups, and a unified outcome variable was generated. Finally, a treatment group factor was constructed to indicate each respondent's experimental condition.

Figure 3: Distribution of Donation Amounts

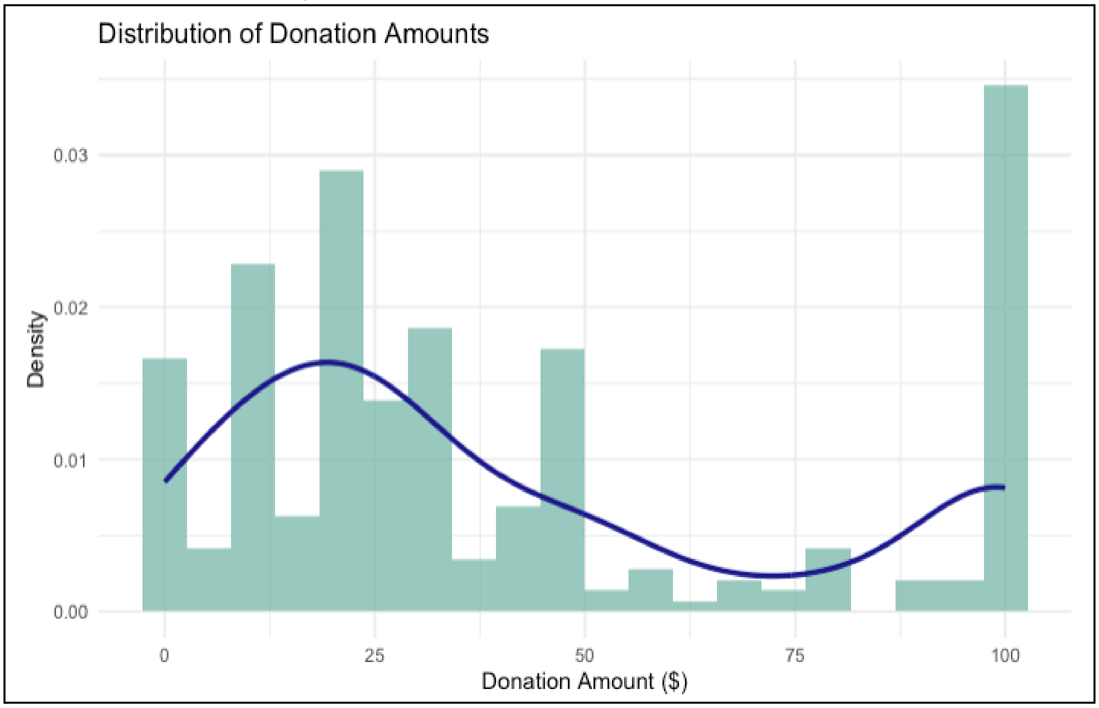
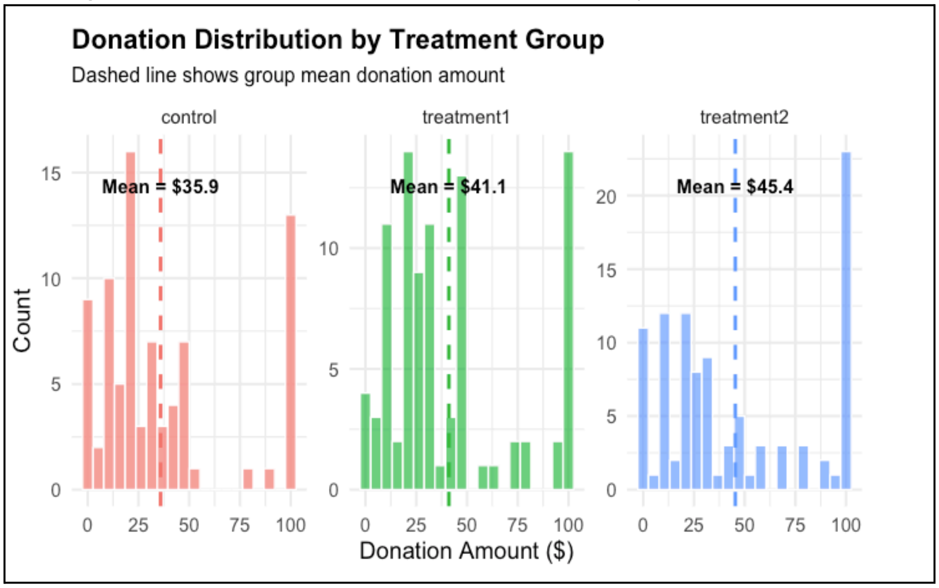


Figure 3 shows that the distribution of donation amounts is heavily right-skewed, with a prominent spike at $100, indicating a common maximum donation choice. A substantial proportion of donations fall between $0 and $30, suggesting smaller contributions are far more frequent. This skewness may impact model assumptions, making it important to consider transformations or robust methods in the analysis.

Breaking the data down by treatment group (Figure 4), we observe a steady increase in mean donation amounts from the control group ($35.9) to treatment 1 ($41.1) and treatment 2 ($45.4). The right skew and $100 spike remain consistent across all groups, suggesting a common behavioral pattern regardless of treatment. These trends indicate that incentives may positively influence donations, but further statistical testing is needed to confirm significance.

Figure 4: Distribution of Donation Amounts by Treatment Group



Randomization Check

Randomization was conducted through the Qualtrics platform. Figure 5 is the distribution of pre-treatment covariates by treatment group, revealing slight imbalances but no major concerns. To formally assess the success of the randomization, a Chi-square test was conducted for each covariate across treatment groups. Table 1 summarizes the resulting p-values, indicating that there are no significant differences in covariates among the groups.

Figure 5: Distribution of Covariates by Treatment Group

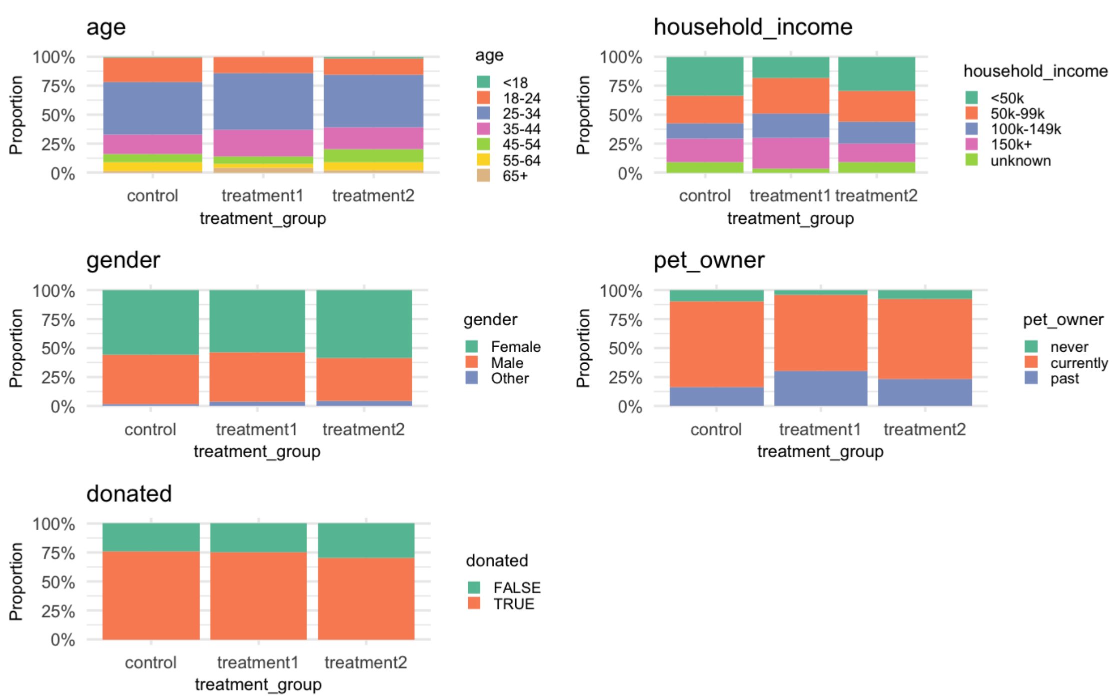


Table 1: Summary of the Chi-squared test for each covariate across treatment groups

variable <chr>	p_value <dbl>
state	0.1740343
age	0.6885807
household_income	0.1361910
gender	0.7322818
pet_owner	0.1849070
donated	0.6174360

Regressions

Before we tested treatment effects, we conducted multivariate regression analysis to explore the dataset and identify correlations between outcome and covariates. Table 2 shows that the coefficients of donation history and pet ownership are 10.9 (p-value < 0.05) and 23.8 (p-value < 0.01), respectively. This result indicates that both covariates were significantly correlated with donation amount before the treatment intervention. Robust standard errors (HC1) were used to account for potential heteroskedasticity issues.

Table 2: Multivariate regression analysis

```
pre_test <-lm(outcome~donated+pet_owner+gender+household_income+age+state,data = df)
robust_se7<- vcovHC(pre_test, type = "HC1")
coeftest(pre_test, robust_se7)
```

```
t test of coefficients:

                          Estimate Std. Error t value  Pr(>|t|)
(Intercept)                10.5464    23.9321  0.4407 0.6598685
donatedTrue                10.9051     5.5288  1.9724 0.0497897 *
pet_ownercurrently         23.7872     8.6378  2.7539 0.0063734 **
pet_ownerpast              10.2485     9.3691  1.0939 0.2751872
genderMale                 -2.8477     5.5301 -0.5149 0.6070972
genderOther               -10.7893    12.4369 -0.8675 0.3865827
household_income50k-99k    -2.8126     6.2763 -0.4481 0.6544904
household_income100k-149k   4.0559     7.4848  0.5419 0.5884413
household_income150k+      -1.4129     7.6093 -0.1857 0.8528608
household_incomeunknown     4.1135     9.6810  0.4249 0.6713141
age18-24                   12.2158    18.2544  0.6692 0.5040583
age25-34                   17.4132    18.7807  0.9272 0.3548281
age35-44                   30.9108    19.5031  1.5849 0.1143959
age45-54                   16.3118    19.6982  0.8281 0.4085023
age55-64                   13.6351    21.0618  0.6474 0.5180458
age65+                     18.5132    20.6286  0.8975 0.3704400
state4                    -32.6109    12.5590 -2.5966 0.0100381 *
state5                     -9.2458    10.6541 -0.8678 0.3864263
state6                    -34.8537    11.7854 -2.9574 0.0034353 **
```

An ANOVA test was conducted to assess whether adding donation history and pet ownership as control variables improves model fit. The significant results (p-value < 0.001 and F statistic) shown in Table 3 indicate that these factors jointly explain more variance in the outcome. A limitation of the ANOVA model is that it assumes homoskedastic errors, so violations of this assumption can undermine its performance.

Table 3: ANOVA Testing

```
Analysis of Variance Table

Model 1: outcome ~ treatment_group
Model 2: outcome ~ treatment_group + donated + pet_owner
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    272 315075
2    269 287047  3     28028 8.7553 1.466e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Below is a detailed analysis of the models presented in Table 4, with a particular focus on the *Model with Covariates* - including donation history and pet ownership as covariates -, identified as the most appropriate model for this study. This analysis includes checks for normality and heteroskedasticity to assess model assumptions, as well as a series of Wald tests comparing all models in Table 4 to evaluate the significance of additional predictors. Together, these diagnostics provide a comprehensive understanding of model performance and the robustness of the results.

The key observation of the model is that treatment 2 and control group coefficients are statistically significant at a p-value of 0.05 level. According to the analysis, the ATE for treatment 2 is 11.06 (~60% of constant) at a significant level, which is a surprisingly higher value than our initial assumption, and the ATE for treatment 1 is 6.56 at a non-significant level. However, a direct test of the difference between treatment 1 and treatment 2 suggests no statistically significant difference between the two treatments (Table 5: $F(1, 269) = 0.88$, $p = 0.35$). The uncertainty in estimation means treatment 1 and treatment 2 could have similar or different effects, but any difference between them cannot be established as statistically significant. This is possible because there’s too much variability or not enough power to distinguish the two treatments from each other, given the sample size.

Table 4: Regression Models with Constant Term Added

	Dependent variable: outcome				
	Simple Model	Semi-Complex Model	Model With Covariates	Full Model	Stepwise Model
	(1)	(2)	(3)	(4)	(5)
Constant	39.21***	27.32***	18.51***	21.78**	23.45***
	(3.12)	(3.44)	(3.87)	(4.62)	(3.91)
treatment_grouptreatment1	5.254	5.305	6.561	5.293	4.815
	(5.776)	(4.778)	(4.639)	(5.379)	(4.690)
treatment_grouptreatment2	9.556	10.398**	11.057**	9.574*	10.069**
	(6.169)	(5.048)	(4.915)	(5.552)	(4.824)
donated		15.006***	12.154***	11.654**	9.073**
		(4.327)	(4.283)	(5.478)	(4.451)
pet_ownercurrently			18.869***	22.780**	22.272***
			(6.942)	(8.848)	(7.421)
pet_ownerpast			2.771	8.437	6.312
			(7.368)	(9.567)	(7.849)
genderMale				−2.113	
				(5.387)	
genderOther				−12.034	
				(12.785)	
household_income50k-99k				−3.932	
				(6.356)	
household_income100k-149k				3.129	
				(7.513)	
household_income150k+				−1.645	
				(7.716)	
household_incomeunknown				3.581	
				(9.756)	
age18-24				13.014	8.043
				(17.267)	(9.855)
age25-34				16.819	13.655
				(17.880)	(9.854)
age35-44				29.896	29.008***
				(18.638)	(10.685)
age45-54				16.024	17.796
				(18.831)	(11.668)
age55-64				13.019	11.212
				(20.175)	(12.776)
age65+				18.269	16.081
				(20.034)	(16.586)
state5				−8.871	
				(8.475)	
state10				−23.274**	
				(10.608)	
Observations	275	275	275	275	275
R^2	0.013	0.051	0.101	0.252	0.142
Adjusted R^2	0.006	0.040	0.084	0.077	0.106
F Statistic	1.776 (df = 2; 272)	4.813*** (df = 3; 271)	6.024*** (df = 5; 269)	1.437** (df = 52; 222)	3.948*** (df = 11; 263)

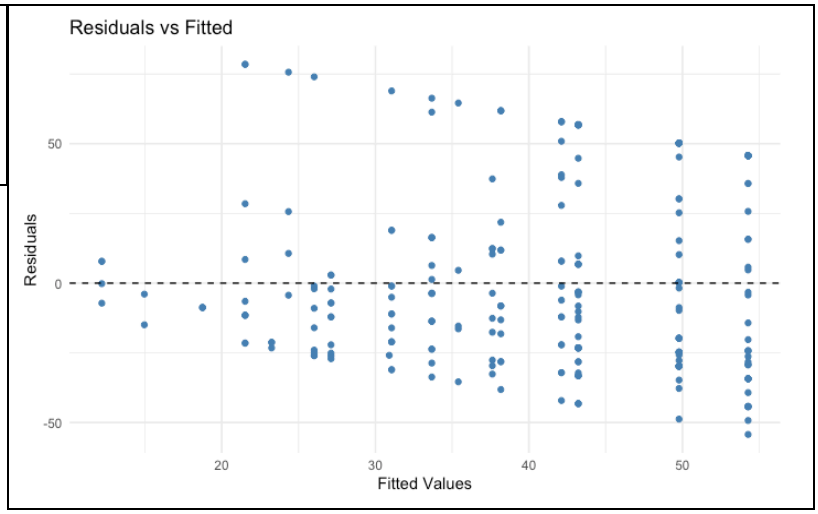
Note: *p<0.1; **p<0.05; ***p<0.01

Table 5: Linear Hypothesis Test

```
Linear hypothesis test:
treatment_grouptreatment1 - treatment_grouptreatment2 = 0

Model 1: restricted model
Model 2: outcome ~ treatment_group + donated + pet_owner

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F Pr(>F)
1    270
2    269  1 0.8814 0.3487
```

Figure 6: BP Test & Residual vs Fitted Plot

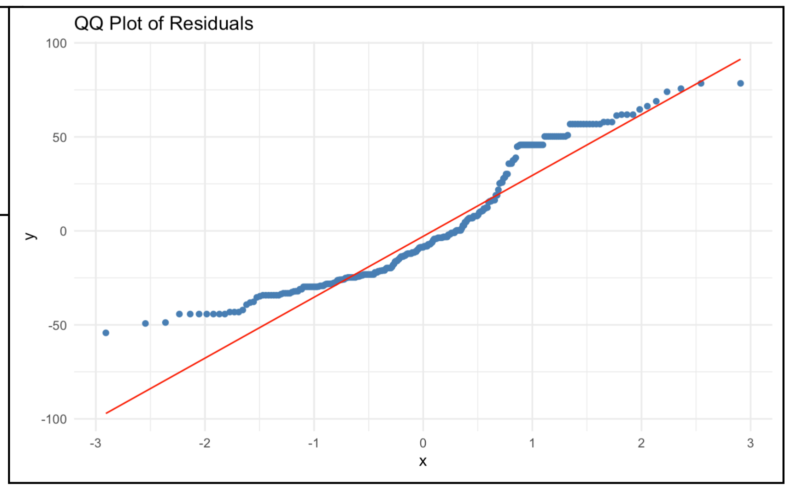
```
	studentized Breusch-Pagan test

data:  model_6
BP = 10.59, df = 5, p-value = 0.06015
```



We conducted the Breusch–Pagan (BP) test alongside a Residual vs. Fitted plot for the *Model with Covariates* presented in Table 4. The BP test returned a p-value of 0.06015, which is slightly above the 0.05 significance threshold, suggesting mild heteroskedasticity but not at a statistically significant level. The Residual vs. Fitted plot visually supports this, showing some variation in residual spread across fitted values, but not enough to raise major concerns about model reliability.

Figure 7: Shapiro-Wilk Normality Test and QQ Plot

```
	Shapiro-Wilk normality test

data:  residuals(model_6)
W = 0.91162, p-value = 1.184e-11
```



From both the Shapiro–Wilk normality test results and the QQ plot of residuals, we can see that the *Model with Covariates* does not meet the assumption of normality. The significant p-value ($p < 0.001$) and the pronounced deviations in the tails from the 45-degree reference line confirm this. Given the skewed distribution of our 275 observations, this result is unsurprising; to address potential issues in inference, we applied robust standard errors (HC1), which help mitigate the effects of non-normal residuals on our coefficient estimates.

Model Comparison	F-Statistic	P-Value
Simple Model vs Semi-Complex Model	10.76	0.001173
Semi-Complex Model vs Model with Covariates	7.49	0.00068
Model with Covariates vs Full Model	0.95	0.5628
Full Model vs Stepwise Model	2.10	0.05408

Table 6: Wald Test Results

Conducting the Wald Test shows that there is no significant difference between the *Model with Covariates* and the Full Model (p = 0.5628), confirming that the *Model with Covariates* is the most efficient choice. The significant results in earlier comparisons indicate that adding covariates improves model fit. This also confirms that differences between control, treatment 1, and treatment 2 are statistically significant, validating their impact when testing different incentives against no incentive.

Discussion

The main points for discussion focus on the ATE, unbalanced sample sizes across treatment groups, the use of "fake money" in the study, participant bias toward helping animals, the impact of bot activity on responses, and the short timeframe for data collection. Our ATE may have been underestimated due to the $100 donation cap, which restricted potential high-value contributions. This limit was intentional since no real money was exchanged. There was also a slight imbalance between treatment groups (82, 93, and 100 participants from control to treatment 2). We addressed this through Chi-squared testing and visual inspection. Although our survey was distributed to a broad audience, those most likely to click the link were people inclined to donate to animals, which may have introduced a positive bias. Additionally, our survey was targeted by bots, resulting in thousands of fake entries that we had to manually filter, which significantly reduced the usable sample size. Finally, the experiment lasted only 10 days, which limited both our recruitment scope and the diversity of participants.

Conclusion

Circling back to our motivating concept, we have evidence suggesting that reward structures influence donor decision-making by activating social norms like reciprocity. Across all the models shown in Table 4 and throughout our EDA, we consistently observed that both treatment groups donated more than the control group. This finding was reinforced through multiple regression models, heteroskedasticity checks, normality assessments, and Wald tests. From both our EDA and statistical testing, we identified the "Model with Covariates" as offering the strongest explanatory power, providing statistically significant evidence that supports our theory of social norms from the motivating concept.

Looking ahead, future studies could strengthen this research by partnering with actual organizations, removing donation caps, targeting a broader audience, and extending the experiment period on a high-traffic platform. This would allow us to test whether the statistically significant patterns observed in our EDA hold up in real-world donation scenarios.