Gratitude

W241: Final Project

School of Information, UC Berkeley

Contents

Statement of research

There have been many studies conducted to examine expressions of gratitude and their impact. A handful of those have shown that gratitude has a measurable impact on an individual's level of happiness and other indicators of wellbeing. However, most of these studies have applied their treatment (mostly nightly journaling) and measured their outcomes over a prolonged period (one week to nine months). We sought to also investigate gratitude, but in a different format and with different outcomes: does a single act of expressing gratitude have a short-term impact on how one perceives and engages with the world around them?

In the event that we found statistically and practically significant effects of quickly triggering a change in behavior through the expression of gratitude, there could be broad applications. Practitioners could specifically use the experiment informed tactics for behavioral intervention in children and students, mediation in community spaces as well as heated conversations such as race relations. Moreover, gratitude treatments could be used in tactics for increasing donations to nonprofits.

Previous work

There have been previous research into the topic of gratitude that have shown to be promising. In one study, at the beginning of an academic quarter, participants were given a packet of 10 weekly reports. The packets were organized into three clusters representing three experimental conditions: a gratitude condition, a hassle condition and an events (control) condition. Over the 10 week period, physical symptoms, reactions to aid, and appraisals (asked to rate how they felt about their life as a whole during the week) were assessed. It was found that subjects in the gratitude group rated their life more favorably on health and "global appraisals" than subjects in the hassle or events group. People in the gratitude group also spent significantly more time exercising (Emmons et al 2003).

Another study found that the emotion of gratitude is associated with the formation of relationships in a sorority setting. The authors examined naturally occurring gratitude during a week when big sisters give presents to little sisters. The little sisters recorded reactions to the benefits they received during the week. At the end of the week and one month later, the sister who gave a gift and the sister who received a gift rated their relationships. Gratitude during the week predicted future relationship outcomes. This study suggests that gratitude may function to promote relationship formation and maintenance (Algoe 2008).

There are also studies that examined the types of gratitude that people experienced. One study sought to discover if people were more grateful for their material or experiential purchases. In the study 95 subjects were asked to list both experiences and materials they purchased and report which one made them feel more grateful. The study found no difference in the price of purchase types but participants expressed more gratitude for experiences than for possessions. The results of this study are only tangentially related to our research question, and it is actually more the method of carrying out the study that is related to our study. The researchers used Mechanical Turk in exchange for modest compensation. The subjects were given information about the difference between an experience and material related purchase and were asked to write about each purchase in detail and then asked to think about their emotions with the purchase. Since this study was found to be statistically significant, it encouraged to use a similar method (Walker et al 2016).

Finally, a meta-analysis of gratitude studies sought to examine different strategies for promoting gratitude. The researchers found that these strategies are simple and relatively easy to incorporate into a variety of treatment strategies. They wish to ultimately evaluate that if things like spending just a few minutes a day of pausing to express gratitude can help people avoid anxiety, depression and other mental health problems. The researchers found that gratitude interventions were in general marginally better than controls but when we adjust for publication bias it is not significant. Nonetheless, the

authors found positive but limited benefits. They believe that the potential of gratitude interventions has not been fully realized but researchers should temper their enthusiasm until it can be shown that experiments do not show diminishing effects over time (Davis et al 2016).

Research Design

Our experiment followed a classic design in which we had two different experimental groups and one control group. We created 3 conditions in Qualtrics, which we used to administer the survey. We instructed it to randomize our subjects into three groups which we refer to as “gratitude”, “control” and “anti-gratitude.” Below are descriptions of the texts that each group received:

- Gratitude: *Think of a person who has had a significant positive impact in your life. It could be a relative, mentor, parent, friend, etc. In the space below, write them a thank you note (45 words or more) in which you express appreciation for the specific role they played in your life and explain how it has impacted you*
- Anti-gratitude: *Describe something in your life that you find most frustrating at the moment. Please use at least 45 words.*
- Control: *In 45 words or more, describe this picture.*

Randomization engineering

The chosen format of Qualtrics and Mechanical Turk did not allow for us to block our randomization based on any demographic (or other) information, since Turkers are

recruited on the fly. However we did administer pre-treatment questions that allow us to check for balance in our analysis. These questions include gender, race, year of birth, religious attendance, political ideology, whether somebody would donate after a natural disaster, and if they donate to a charitable cause once a year. For all of the aforementioned questions with the exception of race, we performed a permutation test to test the sharp null hypothesis of no difference between treatment groups. We were interested in testing that the treatment groups did not have different in characteristics (heterogeneity) prior to having the treatment applied. Specifically, we used a two-sided Asymptotic General Independence Test. A failure to reject the sharp null hypothesis of no difference between groups implies balanced randomization between treatment groups.

Question	**p-value**
Gender	0.8512
Political Affiliation	0.1186
Race	0.3739
Religious Attendance	0.7144
Donate in Natural Disaster	0.3004
Donate Yearly to charity	0.1384

With no p-values of any significant values, we concluded that all of these pre-treatment variables were balanced between groups. We also performed a linear regression predicted treatment groups on age and also found insignificant p-values of 0.602 and 0.69, which showed that age is balanced as well.

We studied literature to find an appropriate n-size for our experiment. An article outlined several rules that we followed. These include researchers needing to decide the rule for terminating data collection before research begins, researchers must collect at least 20 observations per cell, researchers must list all variables collected in a study, and researchers must list all experimental conditions.

Furthermore, in the case of eliminated observations, authors must report what the statistical results are if those observations are included, and finally, if an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate (Simmons 2011). We then did power calculations (alpha=.05 and beta=.8) and determined that a sample size of 50 per group was adequate. Due to randomization, we had 53 subjects in the anti-gratitude group, 49 subjects in control and 47 subjects in gratitude. This was one observation short of the 150 total we targeted, which had to do with logistical difficulties involving some respondents faking responses on mturk which we dropped and then re-recruited.

Survey Design

Following treatment, the subjects received two batteries of questions to measure treatment effect. Responses to questions were given on a 7-point Likert scale: strongly agree, agree, somewhat agree, neither agree nor disagree, somewhat disagree, disagree and strongly disagree. The first battery was intended to measure how people would act in certain situations. Because we did not measure how they actually *did* act, but rather how they stated they would *intend* to act, we referred to this battery as "intentions." Listed below are the reference codes we used for each outcome, along with the questions themselves.

- **Assistance**: *If someone in the parking lot needed assistance that I was able to provide, I would help them.*
- **Conversation**: *I would not be interested in engaging in a conversation with someone who has different opinions/beliefs than me.*
- **Mentor**: *If given the opportunity, I would like to be involved in mentoring youth in my community.*
- **Forgive**: *I would not be willing to forgive someone if they have shown no remorse.*
- **Respect**: *I would like to make it a goal to treat people with greater care and respect.*
- **Give**: *If I were given an additional $5 dollars right now and given the opportunity to donate it to a reliable charity or keep it for myself, I would keep it.*

The second battery was intended to measure beliefs about others (ref: beliefs):

- **Community:** *I believe people in my community have generally good intentions.*
- **Country:** *I believe people in my country have generally good intentions.*
- **Race_intelligence:** *People of other races are less likely to be intelligent than people of my race.*
- **Gender_intelligence:** *People of the opposite gender are less likely to be intelligent than people of my gender.*
- **Choices:** *In general, I believe people who are worse off than me are so because of personal choices they've made.*
- **Race_intentions:** *People of other races are just as likely to have good intentions as people of my race.*

To prepare the data for analysis, some variables had to be flipped on their scales because in the question "strongly agree" denoted a negative response while for others it denoted a positive response). Flipping these variables allows for an even analysis. The variables that we flipped were conversation, forgive, race_intelligence, gender_intelligence and choices.

Analysis

The central test we used in our analysis was the Mann-Whitney test (Wilcoxon rank-sum test). We used this test since our data was ordinal (non-linear), and because any attempts to sum, "stack", or average the data would lose integrity by assuming that the distance between answers was equal. (We do not know that the difference between "strongly agree" and "agree" is the same as the distance between "agree" and "slightly agree." Moreover, we do not know the difference between "strongly agree" and "agree" is the same in the Forgive measure as it is in the Gender_intelligence measure.)

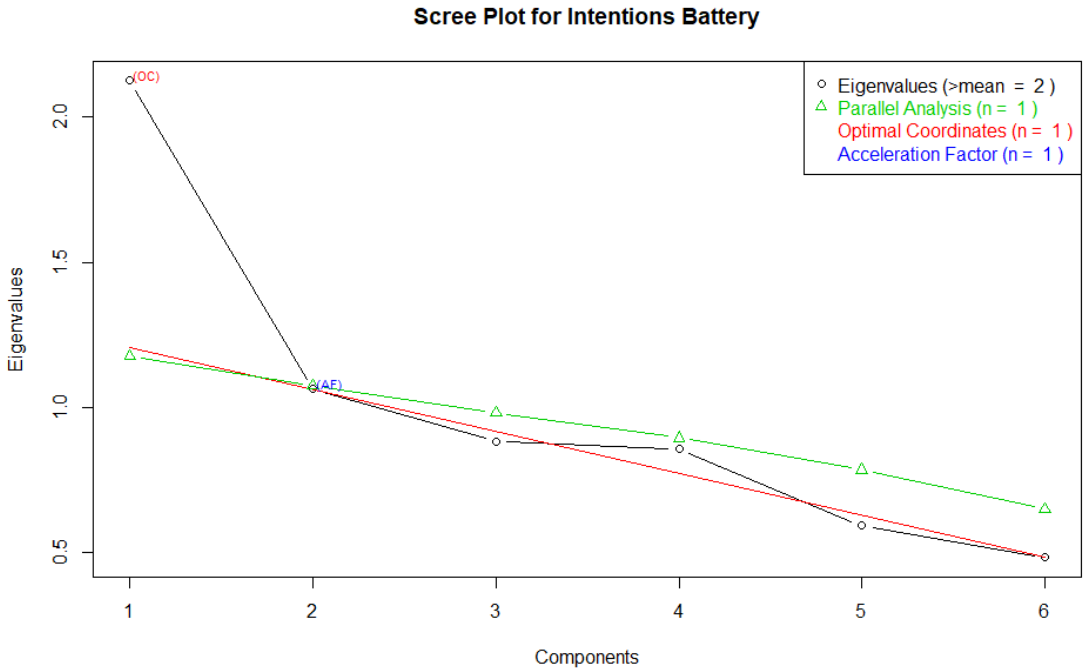
Our first analysis involved looking at answers to each individual question. A limitation of the Mann-Whitney test is that it can only analyze two groups at a time. This meant we ran three tests for each outcome measure - one to compare the Gratitude treatment responses to Anti-gratitude, one to compare Gratitude to Control, and one to compare Control to Anti-gratitude. We had a strong hypothesis

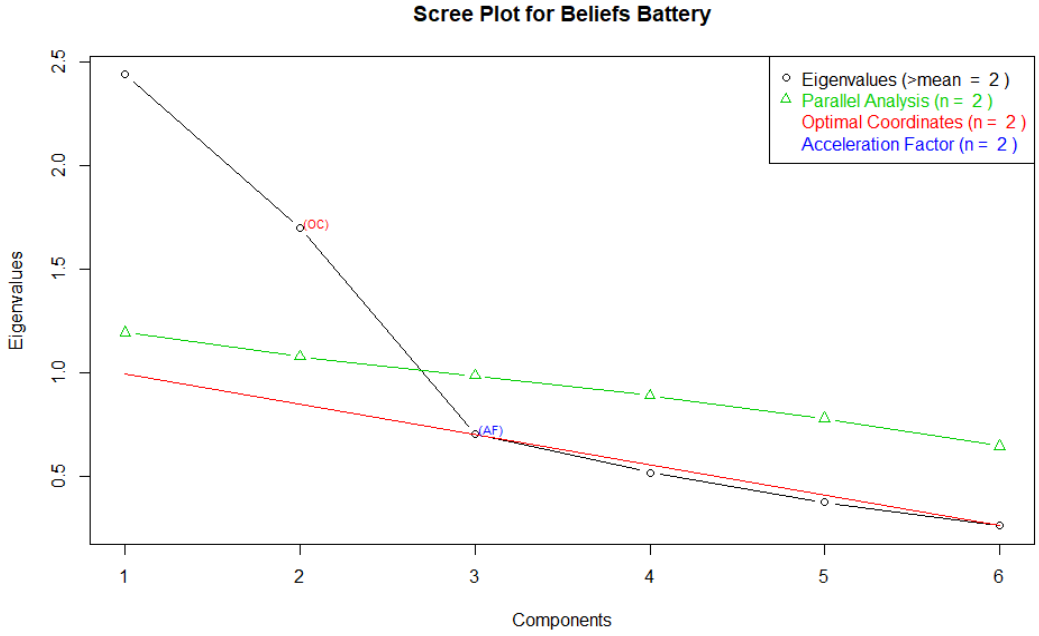
that the Gratitude group would yield more “generous” responses than Control, and Control than Anti-gratitude, so we used a one-tailed test in each instance.

We primarily looked for statistical significance at the p = 0.05 level. Since there were 12 outcome variables, in any one comparison, we would expect to find between 0 and 1 measures with a p-value below 0.05 by random chance. In the comparison between Gratitude and Anti-gratitude, we found 2 measures that were statistically significant (p-values below 0.05): the Conversation metric had a p-value of 0.038 and the Forgive metric 0.00233. The latter is below the p-value that the Bonferroni adjustment would suggest (0.05/12 = 0.004), suggesting true statistical significance of the Forgive metric. Not surprisingly, forgiveness also show statistical significance when comparing the Gratitude group to Control. There was no statistical significance found (below p = 0.05) between Control and Anti-gratitude.

When we expanded to look at significances within the p = 0.1 threshold, we saw similar trends. Whereas we would expect to see 1-2 p-values below that level by random chance, 3 between Gratitude/Anti-gratitude, 1 between Gratitude/Control, and 2 between Control/Anti-gratitude. None of the statistically significant measures (using either threshold), however, showed large effect sizes, meaning practical significance was minimal for each. For a full schedule of Mann-Whitney p-values and effect sizes for statistically significant outcome measures, see Appendix A.

Next, we performed factor analysis as a way to combine questions to analyze them together. We examined the intentions battery separately from the beliefs battery and found that three factors is appropriate for both of them.





We first examined the Intentions battery using an oblique rotation method since the factors are correlated above 0.33. The first factor was composed of conversation and forgive (1.05 and 0.27,

respectively). The second factor was composed of assistance and respect (0.99 and 0.23, respectively). The third factor was composed of mentor, forgive, respect and give (0.76, 0.33, 0.36 and 0.22, respectively). The first factor demonstrates an openness to new people and perspectives, so we called it "openness". The second factor demonstrates treating people well, so we called it "treat_well" and the third factor is somewhat of a miscellaneous factor, but an element of giving back to the community is what differentiates it from the rest of the factors, so we called it "give_back."

Again using the Mann-Whitney test, when examining the openness factor, we found that the subjects in the treatment group were more open than those in the anti-gratitude group with a p-value of .04275. Nonetheless, we found that there is no statistically significant difference between the gratitude and anti-gratitude groups for treat_well or give_back with p-values of 0.84 and 0.11, respectively. The results of analyzing the treatment versus control groups were largely consistent with the analysis of treatment versus anti-treatment. We found a marginally significant p-value of 0.05118 for the openness factor, while the treat well and give back factors are not statistically significant with p-values of 0.75 and 0.26,respectively.

Factor analysis of the beliefs battery also warranted an oblique rotation method since the factors are correlated above 0.33. The factors load in a more clean manner than the Intentions battery. We found that factor 1 was composed of race_intelligence and gender_intelligence with loadings of 0.85 and 0.83, respectively. We called this factor to be "In Group Superiority" since it shows subjects' belief that their own race and gender are more intelligent than other races and gender. Factor 2 was composed of community and country, with loadings of 0.73 and 0.84, respectively. These questions asked if people in the subjects' community and country have good intentions, so this factor was labelled as "Neighbors Good Intentions." The third factor was only composed of race_intentions, with a loading of 0.81. We simply called this factor "Race Intentions."

A Mann-Whitney test of the factors in the beliefs battery when comparing treatment versus anti-treatment did not appear as significant as the intentions battery. The factor of “In Group Superiority” had a p-value of 0.98 and was therefore not statistically significant. The second factor, “neighbors good intentions” had a p value of 0.56 and therefore was also not statistically significant. Finally, the third factor of “Race Intentions” had a p-value of 0.4904 when performing the Mann-Whitney test and is therefore also not statistically significant. Comparing the treatment versus control group was consistent with the comparison of treatment versus anti-treatment. The factor of “In Group Superiority” had a p-value 0.87 and was therefore not statistically significant. The second factor, “neighbors good intentions” had a p-value of 0.66 and was also not statistically significant. Finally, the third factor of “Race Intentions” had a p-value of 0.8297 and was also not statistically significant.

In sum below is a table showing the results of all the significance tests from the variables derived from factor analysis.

	Gratitude Vs. Anti-Gratitude	Gratitude Vs. Control
Openness	Yes**	Yes*
Treat Well	No	No
Give Back	No	No
In Group Superiority	No	No
Good Intentions	No	No
Race Intentions	No	No
	*=0.1, **=0.05, ***=0.001	

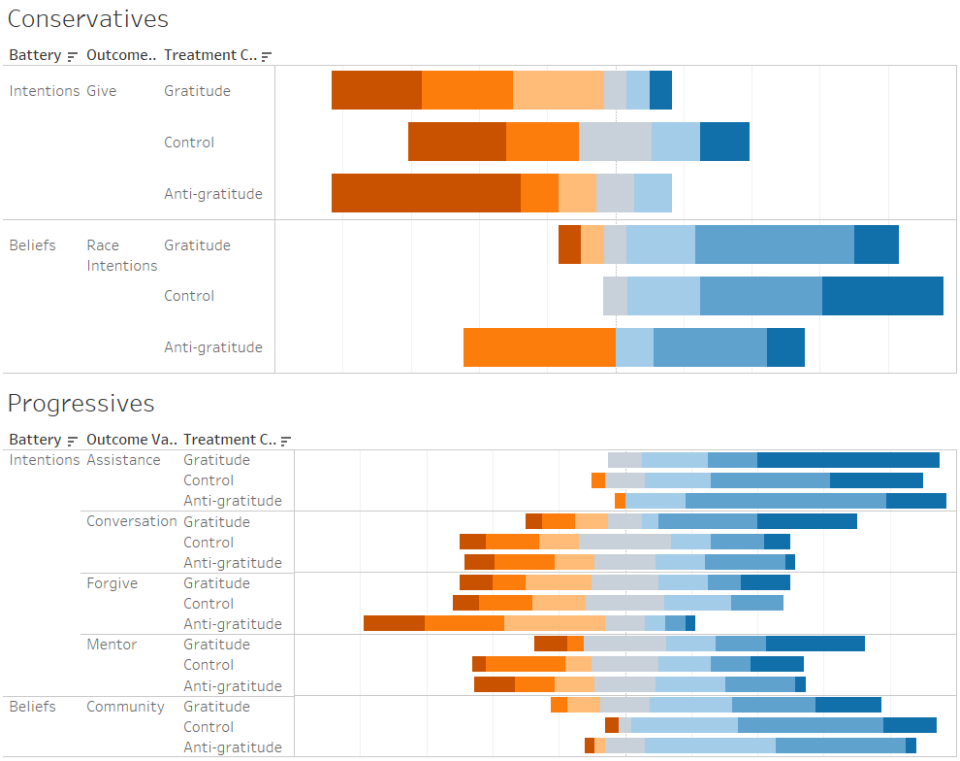
Subpopulation analysis

Because there was balance across demographic variables, we were able to do analysis on subpopulations. The next several images from our visual analysis highlight some of the most interesting differences found in these analyses. There were notable differences both in how the subpopulations answered in general (e.g. women more likely to answer generously on the “Gender Intelligence”

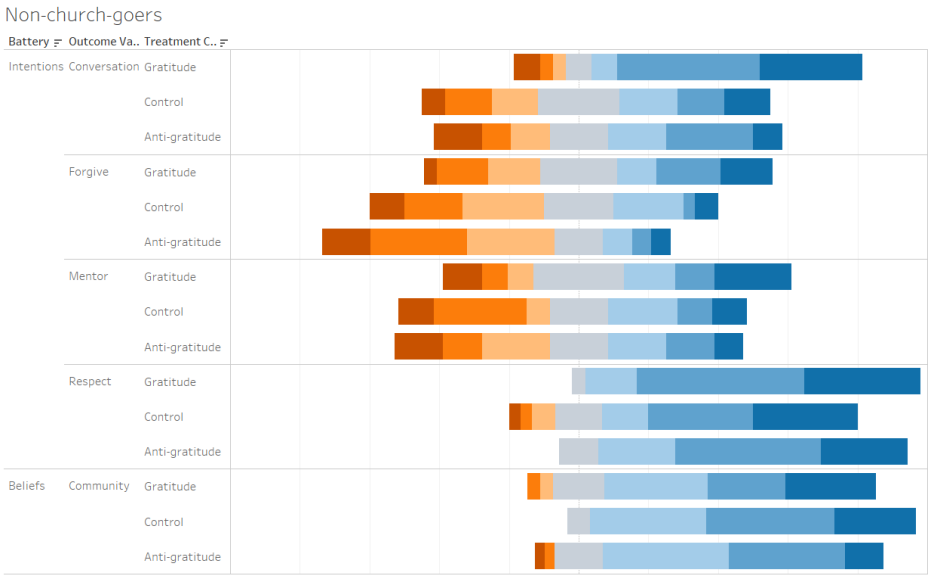
measure) and in subpopulations' sensitivity to treatment and control. Because the latter is more aligned with the purposes of our experiment, we will focus there.

The visuals shown for each subpopulation below are only those found to be statistically significant via the Mann-Whitney test (at the p = 0.1 level).

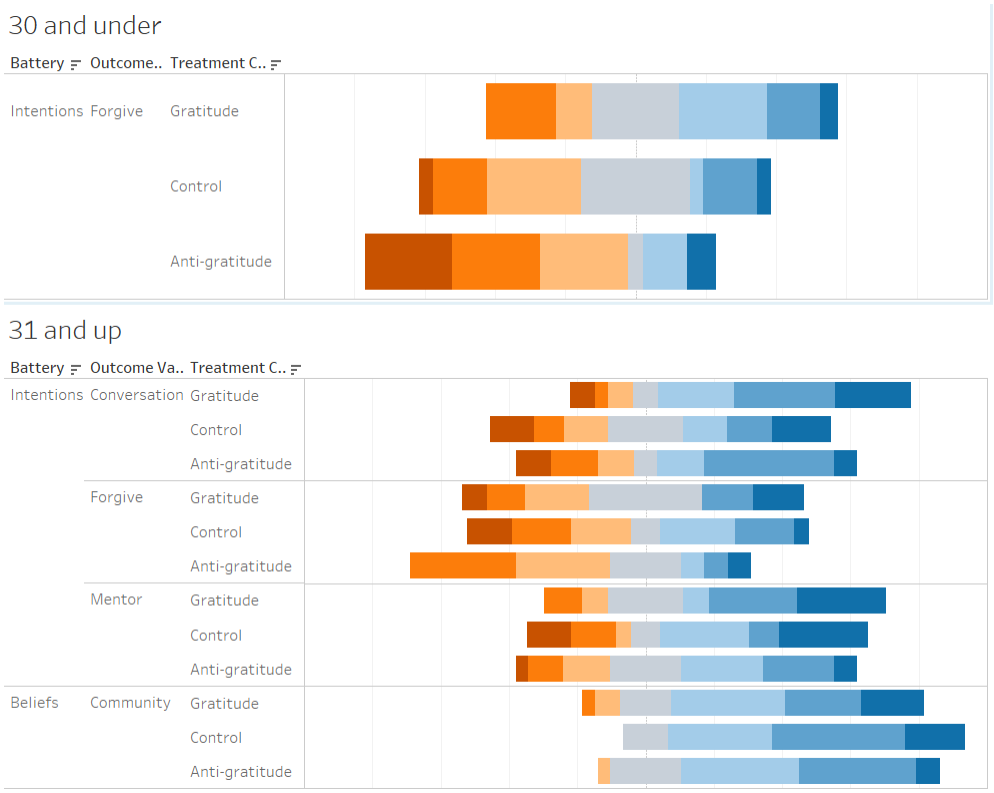
As can be seen from the above visual, we found statistically significant responses to 7 outcome measures for women, but only 3 for men. Notably, the two most statistically significant for women (i.e. conversation and assistance) don't show up for men. Also notable is that the difference in the conversation measure for women had one of the largest effect sizes found in this study.



Another demographic variable we tracked was self-reported political leanings. While there were more Progressives that responded to our survey (and we did not block on this variable), we also had a decent sample size (38) of Conservatives (and fewer Moderates). We therefore decided to only compare Progressives to Conservatives. We found that there were only 2 statistically significant outcome measures for Conservatives, but 5 for Progressives in this study. While this would suggest that liberal respondents are more sensitive to our treatment, a counter to that is that the 2 outcomes that can be seen above for Conservatives were in our top 5 highest effect sizes.



The subpopulations of non-church-goers (n = 93; anyone who marked themselves as "Rarely/Never" going to church) as visually represented above highlights the 5 statistically significant differences for non-church-goers, but none for church-goers sub population (n = 56; anyone who went to church at least a several times a year). This is an important place to note that our findings would necessarily indicate that non-church-goers are *more generous* than church-goers, but rather that they are *more sensitive to our treatments*. A hypothetical story that elucidates this point could be that church-goers are more likely to be accustomed to exercises in gratitude, and therefore our treatments are not novel to that population and have a minimal effect.



Finally, we looked at the population divided by age. It was found that there were 5 statistically significant differences for people over age 30, only 1 for those 30 and under. Notably, “Forgive,” one of the outcome measures that appears as statistically significant over and over throughout our analysis, is the one show to be statistically significant for the 30 and under population. Despite it being the only measure with statistical significance, it has the largest effect size of any measure we found (and therefore has the most practical significance).

Findings: Summary

All-in-all, the study had findings of statistical significance that slightly exceed what we could have expected from random chance. We found that of the well-powered comparisons:

- 14% were significant at p = 0.1
- 6.5% were significant at p = 0.05

In other words, we found 30%-40% more significant findings than we would expect by random chance. While findings varied for different subpopulations, they seemed to be consistently strong for the outcome measures of Forgive, Conversation, and Mentor. In general, more statistically significant findings were found for the outcome measures that measured respondents' intended actions rather than their beliefs.

Findings of practical significance were however smaller, only 10% of the statistically significant findings had a meaningful effect size. We recommend a few potential next steps in the next section to aid future studies that might build on the study ' Gratitude' as we structured.

Potential Next steps

If continuing to improve the study and build from here we recommend the following ways to tweak and build-on the current study:

1. We recommend to use a lab-type environment, to promote authentic and focused responses to treatment. While we attempted to encourage this via word minimums (and by rejecting submissions where text was clearly copied and and pasted from another source), we recognize that Turkers are incentivized to complete work as quickly as possible to earn additional money, thus disincentivizing thoughtful (more time-consuming) responses.

2. We recommend to expand as well as refine the survey questions. Expansion on scopes of questions on topics that were most statistically significant such as below might help enhance the outcome of the study:
    - Forgive,
    - Assistance,
    - Mentor, and
    - Conversation

    One specific suggestion for adjustment is in testing participants' willingness to give to a charity. Two factors limited our ability to assess this outcome. The first was that people on Mechanical Turk are specifically there to earn money. This is likely not a group that could generalize well to the population at large on this outcome measure. Secondly, this is an area where people's self-reported *intention* to act may be different from their actual action. We would be interested to measure results of actually giving participants money and seeing what they do with it when presented with actual charities of various sorts.
3. We would also recommend to Block on pretreatment variables that were found to be impactful (gender, political leanings, age, church attendance).
4. Seeking a larger sample size would enable the future studies to have an increased power.

As stated above, we do believe that the number of statistically significant findings indicate that there is some impact of our treatments, but the revisions listed here would allow us to strengthen our findings to determine with more clarity the true significance of our findings, along with a greater understanding of applications of the findings and generalizability.

References

- Algoe, Sara B., et al. “Beyond reciprocity: Gratitude and relationships in everyday life.” *Emotion*, vol. 8, no. 3, 2008, pp. 425–429., doi:10.1037/1528-3542.8.3.425.

- Davis, Don E., et al. “Thankful for the little things: A meta-Analysis of gratitude interventions.” *Journal of Counseling Psychology*, vol. 63, no. 1, 2016, pp. 20–31., doi:10.1037/cou0000107.

- Emmons, Robert A., and Michael E. Mccullough. “Counting blessings versus burdens: An experimental investigation of gratitude and subjective well-Being in daily life.” *Journal of Personality and Social Psychology*, vol. 84, no. 2, 2003, pp. 377–389., doi:10.1037/0022-3514.84.2.377.

- Simmons, Joseph, et al. “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.” *PsycEXTRA* , 2011, *Dataset*, doi:10.1037/e519702015-014.

- Walker, Jesse, et al. “Cultivating gratitude and giving through experiential consumption.” *Emotion*, vol. 16, no. 8, 2016, pp. 1126–1136., doi:10.1037/emo0000242.

Appendix A.

Schedule of Statistically significant findings in population and subpopulations.

Population	Outcome Measure	Data Comparisons	p-values	Effect Sizes
All	conversation	Gratitude/Anti-gratitude	0.0375	-0.1000
All	mentor	Gratitude/Anti-gratitude	0.0706	-0.1000
All	forgive	Gratitude/Anti-gratitude	0.0023	-0.1000
All	conversation	Gratitude/Control	0.0515	-0.1021
All	forgive	Control/Anti-gratitude	0.0344	-0.0990
All	community	Control/Anti-gratitude	0.0869	0.0000
Women	assistance	Gratitude/Anti-gratitude	0.0417	0.0000
Women	conversation	Gratitude/Anti-gratitude	0.0016	-0.2887
Women	mentor	Gratitude/Anti-gratitude	0.0801	-0.1443
Women	forgive	Gratitude/Anti-gratitude	0.0528	-0.1443
Women	conversation	Gratitude/Control	0.0315	-0.1414
Women	forgive	Control/Anti-gratitude	0.0614	-0.1361
Women	respect	Control/Anti-gratitude	0.0695	0.0000
Women	community	Control/Anti-gratitude	0.0510	0.0000
Women	country	Control/Anti-gratitude	0.0769	-0.1343
Men	forgive	Gratitude/Anti-gratitude	0.0092	-0.1387
Men	mentor	Gratitude/Control	0.0532	-0.1474
Men	forgive	Gratitude/Control	0.0342	-0.1474
Men	respect	Gratitude/Control	0.0926	-0.1474
Conservatives	give	Control/Anti-gratitude	0.0901	-0.2085
Conservatives	race_intentions	Control/Anti-gratitude	0.0292	-0.2085
Progressives	assistance	Gratitude/Anti-gratitude	0.0900	0.0000
Progressives	conversation	Gratitude/Anti-gratitude	0.0118	-0.1374
Progressives	mentor	Gratitude/Anti-gratitude	0.0321	-0.1374
Progressives	forgive	Gratitude/Anti-gratitude	0.0114	-0.1374
Progressives	assistance	Gratitude/Control	0.0841	0.0000
Progressives	conversation	Gratitude/Control	0.0231	-0.1491
Progressives	mentor	Gratitude/Control	0.0948	-0.1491
Progressives	forgive	Control/Anti-gratitude	0.0162	-0.1313
Progressives	community	Control/Anti-gratitude	0.0439	0.0000
Non-church-goers	conversation	Gratitude/Anti-gratitude	0.0026	-0.1260
Non-church-goers	mentor	Gratitude/Anti-gratitude	0.0992	-0.1260
Non-church-goers	forgive	Gratitude/Anti-gratitude	0.0038	-0.1260

Non-church-goers	conversation	Gratitude/Control	0.0041	-0.1325
Non-church-goers	forgive	Gratitude/Control	0.0571	-0.1325
Non-church-goers	respect	Gratitude/Control	0.0782	0.0000
Non-church-goers	forgive	Control/Anti-gratitude	0.0983	-0.1231
Non-church-goers	community	Control/Anti-gratitude	0.0426	0.0000
30-	forgive	Gratitude/Anti-gratitude	0.0044	-0.3015
30-	forgive	Control/Anti-gratitude	0.0141	-0.1414
31+	conversation	Gratitude/Anti-gratitude	0.0919	-0.1336
31+	mentor	Gratitude/Anti-gratitude	0.0615	-0.1336
31+	forgive	Gratitude/Anti-gratitude	0.0833	-0.1336
31+	conversation	Gratitude/Control	0.0505	-0.1414
31+	community	Control/Anti-gratitude	0.0822	0.0000