UNIVERSITY OF CALIFORNIA SAN DIEGO

Essays on Inequality: Insights using Labor and Behavioral Economics

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Economics

by

Amanda Bonheur

Committee in charge:

Professor Gordon Dahl, Co-Chair
Professor Isabel Trevino, Co-Chair
Professor James Andreoni
Professor Prashant Bharadwaj
Professor Marta Serra-Garcia

2024

The Dissertation of Amanda Bonheur is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

The journey towards a PhD is a long one, and there are a number of people I would like to thank for their role in supporting me and my work throughout the process.

I would like to first acknowledge my advising team. I benefitted from a very intelligent and encouraging team of advisors who provided feedback, advice, and all-around support. It would not have been the same without them. I want to thank Isabel for her insights, encouragement, conference advice, holistic approach to advising, and for always saying things in a constructive way. I thank Gordon for pushing me to be exact about my research and holding me to a high standard. Prashant, thank you for developing my research capacity from early on in the PhD, insisting that I include models in my projects, fostering my teaching interests with the economics of discrimination course, and offering your home as a backup location for my dissertation defense. I am also grateful to Isabel, Gordon, and Prashant for extra support during the job market. I thank Marta for her logic, wisdom, and unwavering commitment to real-world applications. And to Jim, thank you for sharing your wealth of knowledge about the charitable giving literature and the IRB process, along with your excitement about research. I am grateful that each and every one of you learned new things and worked together to provide me with a range of expertise and fantastic support. I also want to thank you for trusting me and allowing me to prioritize family and conduct some of my PhD remotely. For everything, thank you.

I would also like to thank my CSWEP mentors – Prachi Jain, Sarah Jacobson, and Wenjia Zhu – for being so generous with your time. You volunteered to mentor a small group during a workshop, but continued to meet with us every 6 weeks to support the emotional side of the PhD process. These meetings, where we talked about the hidden curriculum, strategies to manage stress and various situations, how to cope with a rejection, and the self-care and personal management that it takes to complete a PhD, were invaluable. I always left our meetings feeling buoyed and ready to tackle the next stage. Thank you also for job market help, from mock interviews, to telling me I was ready, to feedback on my research/teaching statements. Perhaps most of all, thank you for reiterating the importance of both relaxing and celebrating.

A massive shoutout to my wife and biggest supporter. Thank you for always keeping me sane, encouraging me to take breaks, brainstorming and developing research with me, spending countless hours discussing what slacktivism is, giving me pep talks at various stages, supporting me, and always believing in me. I can only imagine how much harder the last six years would have been without you.

Thank you to CSQIEP for making the field more inclusive by organizing the LGBTQ+ seminar series, hosting LGBTQ+ events at conferences, and starting a mentoring program. Thank you, Tami Hayford, for being my CSQIEP mentor and giving your time, ear, and unique perspective.

Thank you to Tanner for being an amazing co-author and friend. Working with you makes the research more enjoyable. I have been lucky to have you as a friend and comrade every year of the PhD process, from problem sets to group projects to setting up a company together, with all the highs and lows in between.

Thank you to the UCSD Economics department more broadly. Everyone in the department contributed to my experience; my cohort, peers, staff, professors, seminar participants, TA professors, EconLab managers, ECONnected co-founders and students, et cetera. I am proud to have been a part of this department with you all.

I am grateful for my San Diego community of Jess Pickering & Pinar Yilmaz and Tanner & Allie Eastmond. I can't thank you enough for filling my time in San Diego with joy. Jess and

Pinar, I will never forget the hundreds of meals you cooked for me throughout the first year and beyond. Your excitement and cheerleading are things that will always be special to me.

Similarly, I am thankful for my DC/NYC community. Zoë Brouns, Leon Gray, Pierre Guigon, Alex Finnegan, Chris Allen, and Megan Tingley, you all contributed to my well-being by listening, giving encouragement and confidence-boosts, and reminding me of life outside of grad school.

Thank you to the San Diego Women's Rugby club, the Medstar Capitals Women's Hockey League, and the Washington Wolves Women's Ice Hockey club for being physical outlets, communities, and welcoming spaces for me to grow non-academically.

I would be remiss not to mention everyone who played a part in my decision to pursue a PhD in Economics. This includes my undergraduate advisors Tony Becker and Becky Judge and my colleagues at the Federal Reserve: Shifrah Aron-Dine, Laura Feiveson, Andrew Chang, Eugénio Pinto, Claudia Sahm, and Stacey Tevlin, to name a few.

Last but not least, thank you to the various funders and grant organizations who have not only made my research possible, but also encouraged me along the way. Thank you to the Russell Sage Foundation, the W.E. Upjohn Institute for Employment Research, the Yankelovich Center for Social Science Research, the Frieda Daum Urey Endowment, and the UCSD Economics Department. Thank you for your financial support and for reminding me that my work is of interest to others.

Some coffee shops and bakeries that gave me caffeine sustenance, gluten free baked goods, and inclusive and inspiring spaces to work in include: Café Bassam, Art of Espresso, Starry Lane Bakery, Tatte, Rise Bakery, Better Buzz, As You Are, Emissary, Busboys and Poets, Last Mile

Cafe, The Coffee Bar, and Caribou Coffee. Thank you also to UsFolk and their Belfast coffee map for unknowingly giving me daily motivation.

There are undoubtedly more people (and coffee shops) that could be included on this list, so to everyone, thank you.

Chapters 1 and 2 are currently being prepared for submission for publication of the material. Bonheur, Amanda. The dissertation author was the primary investigator and author of these materials.

Chapter 3, in part, is currently being prepared for submission for publication of the material and is coauthored with Eastmond, Tanner S. The dissertation author was the primary researcher and author of this material.

VITA

EDUCATION

2024 Doctor of Philosophy in Economics, University of California San Diego

2020 Master of Arts in Economics, University of California San Diego

2016 Bachelor of Arts in Economics and Mathematics, St. Olaf College

FIELDS

Labor Economics, Behavioral Economics, Inequality

PROFESSIONAL EXPERIENCE

2018 – 2024 Teaching Assistant, Economics Department, University of California San Diego

2020 – 2021 Research Assistant, Economics Department, University of California San Diego

2020 Consultant, Poverty and Equity Global Practice, World Bank Group

2016 – 2018 Sr Research Assistant, Research & Statistics, Federal Reserve Board of Governors

ABSTRACT OF THE DISSERTATION

Essays on Inequality: Insights using Labor and Behavioral Economics

by

Amanda Bonheur

Doctor of Philosophy in Economics

University of California San Diego, 2024

Professor Gordon Dahl, Co-Chair
Professor Isabel Trevino, Co-Chair

This dissertation investigates inequality-related issues across three chapters, focusing on social media activism and disparate impacts of policies on historically marginalized groups.

Chapter 1 provides the first causal evidence of slacktivism, the phenomenon whereby visible, low-effort forms of support on social media deter more vital, higher-cost actions. Using a laboratory experiment, I show that subjects who send –or “post”– a digital message to peers stating

‘I support racial justice’ are less likely to donate to related charities than those who could not publicize their support. While people believe their post is helpful, posts do not encourage others to give, and perceived effectiveness of posting is not driving results. Rather, self-interested biases explain behavior. Multiple mitigation strategies are unsuccessful, demonstrating the persistence of slacktivism.

Chapter 2 shows that retaliation disproportionately disadvantages women on peer-to-peer review platforms. I leverage an exogenous policy change implemented by Airbnb in 2014 which made reviews simultaneous reveal, preventing users from seeing each other’s review before posting their own. The policy enhanced honesty by removing the ability to retaliate to negative reviews and reciprocate extremely positive ones. Analysis of review data reveals that while both male and female guests wrote less positively worded reviews, the negative shift was more pronounced towards male hosts. This indicates that reviews for male hosts were artificially elevated due to fear of retaliation, confirming that review systems with retaliatory capacity exhibit gender biases.

Chapter 3 explores whether we can narrow application gaps and promote diversity by modifying language around qualification requirements in job ads. We established a non-profit company that acts as an intermediary in the job search process to conduct a large-scale, reverse audit study field experiment where we randomize the content of real job ads. We vary whether job seekers are encouraged to apply even if they don’t meet all the listed qualifications and whether they are informed that companies routinely hire individuals who do not have all qualifications. Preliminary results show this intervention encourages applications, and we hypothesize that it will have larger impacts on women and individuals from underrepresented racial groups by changing perceptions of the hiring process.

Chapter 1

Is Slacktivism Harmless? Unintended Consequences of Social Media Activism

Amanda Bonheur[1]

Abstract

People show support for causes on social media, which has raised the 'slacktivism' concern that visible, low-effort displays of support may crowd out vital, higher-cost actions. Using a between-subject laboratory experiment, I show that subjects who send - or "post" - a digital message to peers stating 'I support racial justice' are *less* likely to donate to racial justice charities than those who could not publicize their support. While people believe their post is helpful, posts do not encourage others to give, and perceived impactfulness of posting is not driving results. Instead, self-interested biases explain behavior. Multiple mitigation strategies are unsuccessful, demonstrating the persistence of slacktivism.

[1] Department of Economics, University of California San Diego. Email: abonheur@ucsd.edu. This research was pre-registered on AsPredicted as #90396 and conducted under IRB #800210. Funding was provided by the Frieda Daum Urey Endowment/Matching Fellowship and the UCSD Economics Department through the Diversity Research Fellowship and the Travel and Research Grant. Thank you to Isabel Trevino, Gordon Dahl, Prashant Bharadwaj, Marta Serra-Garcia, Jim Andreoni, the UCSD applied and behavioral seminar participants, the ACLEC and ESA conference attendees, and CSWEP mentors for valuable comments. I thank Oishee Misra and Jay Cizeski for excellent research assistance. I wish to acknowledge the support I have received from Ifeyinwa Bonheur. All opinions, conclusions, and errors are my own.

1.1 Introduction

The prevalence of low-effort, visible ways of showing support for a cause, especially on social media, has raised concerns that these might crowd out more impactful forms of support. This concept is known as slacktivism and has received significant media attention.[2] Taking cues from the political science and journalism literatures, this paper defines slacktivism as "the phenomenon where someone posts to social media in support of a cause, but in doing so reduces their total contribution to that cause relative to if they had not posted." (Anderson et al 2018, Harlow and Guo 2014, Morozov 2009a,b, Robertson 2020). Millions of individuals use social media platforms to express support, and this may influence other forms of engagement. Donations and protests are economically and socially meaningful, so understanding this phenomenon is important.

This paper performs the first causal examination of the existence and extent of slacktivism. Being able to show support for a cause online may influence other forms of support in a variety of ways. Some may post and feel that is sufficient, perhaps exacerbated by overestimating how their post impacts others' actions. Conversely, posting may foster a stronger connection to the cause and motivate individuals to become more involved. Similarly, those who see posts of support from others may assume others are donating more and free ride, or be encouraged to contribute more.

There are three main difficulties in detecting slacktivism that I overcome. The first is that individuals who post support for causes on social media are self-selected, and we can't see counterfactual decisions in a world without social media. The second is a lack of data on people's actions of support, including an inability to observe both people's social media use and their private actions of support (e.g. private donations, semi-private signing of petitions, participating in protests, etc). Thirdly, the impact of posting on the cause is critical to measuring slacktivism (Karpf 2010, 2012; Halupka 2014). Someone's total net contribution to a cause is the sum of all of their actions of support. Posts of support are visible and may influence others, so this needs to be included in the calculation of one's total contribution. I am able to circumvent all three of these issues empirically

[2] Major news outlets, including NBC, the New York Times, BBC Future, have recently covered slacktivism in relation to the Black Lives Matter movement (Ho 2020, Haberman 2016, Fisher 2020, Robertson 2020). Other names are performative activism (Abdi 2020), armchair activism (Elias 2020), and clicktivism (Haberman 2016, Fisher 2020).

using a laboratory experiment.

I study contributions towards racial justice, a pertinent example in the slacktivism debate due to its significant digital presence, contentious nature, and reliance on grassroots fundraising. Showing support for racial justice online has surged, exemplified by the shift from under 86,000 daily tweets with the Black Lives Matter hashtag in 2019 to an average of 3.7 million tweets per day following George Floyd's murder (Anderson et al. 2020). Notably, Blackout Tuesday on June 2, 2020 saw 28 million people post a black square in support of Black Lives Matter (Ho 2020). Social justice charities are often especially reliant on individual donations (Boris et al 2010; NPT 2022; WPI 2022), so they are more vulnerable to slacktivist behavior. While I focus on social media and racial justice, the implications of my findings are broad because slacktivism can manifest in numerous contexts and is not limited to this case.[3]

I first formalize slacktivism using a theoretical model of behavior. Donations have economic consequences and are one of the most prevalent ways people directly support causes,[4] and so serves as the higher cost way people can support the cause in my study. Traditional donation models involve a trade-off where donating provides warm glow, joy of contributing to the public good or cause, and a reputation of being prosocial if donations are revealed, while not donating provides consumption utility (Bénabou & Tirole 2006). My model expands on this framework, allowing people to contribute through two actions: private donations and public posts of support. Warm glow and the satisfaction of contributing can stem from either or both actions, while reputation is tied solely to visible posts of support. An individual's contribution is comprised of their direct donation and the indirect effect of their post on others, should they choose to post. In the context of slacktivism, if posting support positively influences others, some substitution of donations and posts is logical. This underscores the importance of measuring the impact of a post on others' donations to fully assess changes in one's total net contribution. Furthermore, beliefs about

[3] For example, slacktivism can include displaying a Ukrainian flag in one's window or adding a 'Save the Planet' frame to a profile picture if these actions reduce the likelihood of donating coats to displaced Ukrainians or opting for the higher-cost energy plan that supports renewable sources.

[4] Two-thirds of all individuals in the US donate to charities, which is larger than the one-third of individuals who volunteer (NPT 2022, Philanthropy Roundtable n.d.).

a post's impact will substantially influence decisions.

I run a real donation experiment to simultaneously test for donation crowd out, measure the impact of posting, and evaluate total contributions as a result of social media-like posts. I randomly assign individuals to one of three groups: a control group with no social media environment, a treatment group where they can post their support digitally (senders), or a treatment group where they can only observe posts of support from others (receivers). All participants have the same option to donate privately to racial justice charities. Those with the option to post additionally choose whether they want to send a digital pop-up message of support saying 'Person at desk # supports racial justice.' I measure the causal impact of seeing one additional post on donation behavior using receivers who observe up to 5 posts of support from 5 randomly matched senders. This measure is the full impact of the posts in the controlled laboratory environment, since there is limited awareness raising capacity.[5] Donations are private, meaning they are not revealed to other participants, and are costly since any amount donated is subtracted from their earnings and sent to the charities instead (Eckel & Grossman 1996). Posts are public and monetarily free to send. I compare donations and net individual contributions across subjects with no social media-like environment (control), those with the option to post support (sender), and those who see posts from others (receiver).

I find that people who are randomly given the option to post are significantly less likely to donate than individuals in the control group. The majority of senders (75%) send a digital post of support for racial justice. Both senders who choose to post support and those who don't crowd out their donations relative to control individuals who do not have the option to send a costless message of support. The extensive margin crowd out by senders is large; logistic regressions show that senders are half as likely to donate relative to individuals without a social media environment. The shift is coming from those who would have donated a small amount ($1 or $2), who now

[5] Outside of the lab environment, posts could raise awareness about the cause and this would need to be included in one's total contribution. My sessions are run during 2022 with 95% of my sample agreeing that racial justice is an important cause, meaning awareness levels are already high. Further, posts are short-lived and only exist on receivers' screens for the duration of the decision-making portion of the experiment. Subjects do not think posts will affect their decisions in the future, further emphasizing the limited capacity for posts in the experiment to influence attitudes and behavior beyond donations of receivers.

post and don't donate. High-donators are relatively unaffected, meaning the ability to post has a negative, albeit insignificant, effect on the average amount donated per person.

Importantly, posts of support do not encourage others to give. Receivers are also only half as likely to donate relative to the control group. Receivers are randomly matched with five senders and see up to five posts of support from those senders, depending on those sender's decisions. Those who are matched with more senders who post observe more posts, but do not exhibit different donation behavior compared to those who saw fewer posts. Using this random variation in the number of posts that a receiver observes, the marginal impact of a post in my setting is statistically indistinguishable from zero.[6]

Combined, I find that slacktivism does occur and is harmful. People who can post their support digitally often do and crowd out their own donations. Their post does not motivate others to give. Taken together, those who post in support of racial justice end up reducing their total net contribution relative to if they had not posted, exemplifying slacktivism. Not only does the posting environment reduce net contributions, it lowers the number of people who donate. Both senders and receivers are less likely to donate, translating to the posting environment causing a 15% reduction in the overall donor base.

The results are not driven only by people who think they are affecting change. A common critique of slacktivism is that people may not be intentionally crowding out, but rather may genuinely believe they are making a difference (Anderson et al 2018; Harlow & Guo 2014). The vast majority of senders do overestimate the impact of their post on receivers' donations. However, even senders who believe the posts will have a neutral or negative impact on the cause are less likely to donate than control individuals, and many of them continue to post.[7]

The mechanisms driving slacktivism are self-interested behavioral biases such as motivated beliefs, justification effects, and moral licensing. There is widespread motivated reasoning, where

[6] The regression to measure the impact of a post was included in the pre-registration. I can rule out effect sizes larger than positive 27 cents and more negative than -$1.07.

[7] The same is true for senders who are told that posts have a negative impact. I have information treatment arms where subjects are told a data-driven value of how posts impact others' donation behavior in previous sessions. Behavior of senders who are told posts have a negative impact doesn't suffer from endogeneity since information treatment assignment is randomized. Additionally, belief elicitation is incentivized.

half of senders recognize that posts of support won't influence their own donation choice, but continue to believe posts will increase others' donations. There is some correlation between holding these motivated beliefs and donating less often, suggesting that motivated beliefs allow some people to justify their own non-contribution. There is also evidence of a justification effect. People who had the option to show support publicly give significantly shorter explanations of their donation decisions and are more willing to use financial need as an excuse. These patterns are consistent with moral licensing, which is "the cognitive process by which individuals justify morally questionable behavior by having previously engaged in moral behavior" (Blanken et al 2015). Moral licensing is common in many contexts (Dütschke et al 2018; Monin & Miller 2001; Simbrunner & Schlegelmilch 2017), including charitable giving (Grieder et al 2023; Zhang & Peng 2022). The phenomenon is especially pronounced when the good behavior was observed (Rotella et al 2023). People even experience vicarious moral licensing, where they act more immorally after seeing others engage in moral behavior (Kouchaki 2011; Meijers et al 2019). My results indicate that self-benefitting biases are strong, since even short-lived, generic posts significantly change behavior. Further, the ubiquity of these self-interested behavioral biases leads me to believe that the observed mechanisms are not limited to this setting, but rather are representative of behavior outside of the laboratory.

Given that slacktivism occurs, a natural next question is how to mitigate this behavior. The first mitigation strategy I test is an information treatment where I provide a signal about the impact of posting. I anticipated that beliefs about the impact of a post would be important determinants of decisions due to previous articles (Jones 2015; Hogben & Cownie 2017) and insights from my model. I implemented an information treatment arm in which senders were informed about the impact of posting using data from receivers in all previous sessions.[8] Throughout data collection I re-calculated the pre-registered regression of the marginal impact of seeing an additional post on receivers' donation behavior, to be able to provide varying information about posts without

[8] The value senders are told is the coefficient of the pre-registered regression used to measure how receivers respond to observing an additional post. Timing of the sessions was determined by lab participation rates, and I did not look at what the value would be before deciding whether or not to run an information or no information treatment session.

deception. The measure of the impact of a post bounced around the true null effect as the sample grew, so some senders were told posts have a small positive impact while others were told a small negative impact. Senders do update their beliefs of a posts impactfulness down towards the value they were told, and some no longer post, but interestingly, those informed of a negative impact exhibited similar donation behavior to those informed of a positive impact and those who received no information regarding a post's impact. The lack of donation response to these signals implies that simply providing information about the impact of posting is unlikely to meaningfully reduce the incidence of slacktivism.

The second mitigation approach involves a policy-oriented nudge treatment aimed at encouraging donations within the posting environment and curbing the behavioral mechanisms. To design this follow-on nudge intervention, I draw upon insights from this study and use straightforward language that non-profit organizations can apply in real-world scenarios.[9] People want to use posts to justify not donating whilst retaining a reputation of being a supporter. To make this more difficult, senders are informed that only posting is insufficient to make a meaningful impact on the cause, while receivers are told that relying solely on others to post is inadequate. All subjects are made aware that people who post are less likely to donate to create a disconnect between posting and reputation. Everyone is reminded that donations are crucial for charities and encouraged to donate, regardless of whether they choose to post (or irrespective of the posts they encounter from others, in the case of receivers). Unexpectedly, donation behavior and mechanisms are unaffected by the nudge. Overall, slacktivism is a persistent phenomenon that is difficult to mitigate.

As far as I am aware, this is the first paper to causally test whether slacktivism exists. Previous social science work has found that there is generally a positive correlation between online and offline activism since some people engage more than others (Chon & Park 2020, Greijdanus et al 2020, Ogilvy & CSIC 2011). A small but growing literature explores public statements of support including social media on subsequent activist actions. Most find a demobilizing effect (Schumann & Klein 2015, Kristofferson et al 2014, Wilkens et al 2019; Hogben & Cownie 2017)

[9] The 'main' sample for this study was run in 2022 and includes control sessions and treated sessions (receivers and senders) with and without information. The nudge intervention sessions were run roughly a year later in 2023.

while Lee & Hsieh (2013) find modest consistency and moral balancing effects. I build on the foundation of these papers and identify the causal nature of social media on both donations and total contributions. I am the first to quantify the possible impact of posting, an important piece of net individual contributions.

This paper provides a framework to think about digital age extensions to charitable giving. Research has shown that when donations are visible and indicate generosity or effort, people donate more (Andreoni & Petrie 2004; Ariely et al 2009; Bénabou & Tirole 2006; Bracha & Vesterlund 2017) because they care about how they are perceived (Bursztyn & Jensen 2017; Meier 2006; Ottoni-Wilhelm et al 2017). Social media changes the observability of actions in this space; posting to social media is an inherently observable action that is independent from one's donation. People use the public forum to check-the-box as having done something, allowing themselves to justify not donating. Seven in ten of Americans use social media (Pew Research Center 2021), so this interaction between visible, low-cost actions and private, high-cost actions is crucial for organizations and social justice movements.

Outside of social media, this paper also contributes to discussions of substitution in prosocial giving. There is mixed evidence about whether awareness or targeted fundraising efforts towards one charity crowd out donations to another (Filiz-Ozbay & Uler 2019; Gee & Meer 2019; Harwell et al 2015; Scharf et al 2022). Most notably, Thunström (2020) finds that prayers for hurricane victims crowd out donations to the hurricane relief effort. This is direct evidence of slacktivism in a different, offline context.

This paper confirms that slacktivism is a real and insistent phenomenon. I find concrete evidence of slacktivism whereby people who can show digital support are less likely to donate to that cause. Simple information and nudge interventions are ineffective in raising donations to the level seen in the no social media environment, confirming that the slacktivist trade-off between low-cost, visible actions and high-cost, private actions cannot be ignored. This paper also provides a framework for both hypothesizing about and evaluating slacktivism in other contexts.

1.2 Model

To formalize the concept of slacktivism, I build on the charitable giving and public economics literatures by expanding standard donation choice models to include public posts of support. Canonical models of charitable giving describe how people choose if and how much to donate (Andreoni 1990; Andreoni 2006; Andreoni & Payne 2013; Kahneman et al 1986; Marwell & Ames 1981; Samuelson 1954). However, donations are not the only way to contribute to a cause. People can post their support, and this may contribute to the cause in its own way or affect other decisions. My model characterizes this joint decision environment, allowing us to analyze not just donations, but also the broader concept of total own contributions to the cause from all actions of support.

People can perform two different actions of support: donate and post support. Define d_i as the private *donation* and p_i as the visible *post* for individual i. People can either post or not post, so p_i is binary and equals 0 or 1. In my setting, people can donate between $\$0$ and $\$12$, and anything they donate is deducted from their earnings they keep, x_i.

An individual's total own contribution towards the cause is the combination of all of their actions to support the cause. In order to measure person i's individual contribution, denoted as y_i, we need to be able to sum these different types of actions. Donations add to contributions in a straightforward way; $1 donated is $1 contributed. How a post of support contributes needs to be translated into dollar terms.

A post could have impact by encouraging others to donate more or by raising awareness about the cause. In my controlled lab setting, I can measure how posts influence others' donations, and at the same time, the awareness-raising capacity of posts is limited. Posts are unlikely to raise awareness in this context because awareness levels are already high and posts are not elaborate. Subjects are informed about the racial justice charities during the instructions before making decisions, and sessions are conducted after the surge of public support for Black Lives Matter in Summer 2020. Nearly all (95%) of subjects agree that racial justice is an important cause.[10] Posts are generic in that they say 'I support racial justice' but are not charity specific and do not contain

[10] This question is asked during the exit survey because I didn't want to prime individuals.

personalized messages. Posts exist for a short sub-duration of the experimental session, a couple minutes at most, and are not posted to actual social media accounts.

If posts were to raise awareness, we would expect that to affect subjects' actions in the future. I find no evidence of people expecting themselves or others to contribute more in the future as a result of posts. Most people don't think themselves nor others will donate more next week, and some think it will increase next week's actions a little. There is no difference in this pattern for senders or receivers relative to the control.[11] There could be some benefit from knowing how many of your peers support racial justice, and this could encourage recievers to donate more, but this is unlikely to have a large impact outside the lab. This means that posting support in my experiment is unlikely to raise awareness, and can realistically only contribute in my setting through encouraging others to donate more during the session.

I measure how posts affect others' donations using receiver behavior. Receivers see posts of support from senders, and due to random variation, I am able to measure the impact of an additional post on receivers' donations. This measurement translates the impact of a post into a dollar value of its contribution towards the cause. Mathematically, I write this as $\kappa^* p_i$. When someone posts, then $p_i = 1$ and the impact of that post is \$$\kappa$, which is the amount that receivers changed their donation by due to seeing the post.

Now that we have the dollar-value contribution from donating and the dollar-value contribution from posting, we can sum them to an individual's total contribution. Person i's total contribution is equal to their donation plus the impact of their post, so $y_i = d_i + \kappa^* p_i$.

In line with previous papers, people care about their individual contribution (warm glow and altruism), the total amount contributed, money they keep for consumption, and any personal net benefits of visible actions (e.g. reputation). The total amount contributed is the total amount donated by everyone in my lab setting. If posts had an effect besides influencing others' donations, then you could have posts contributing directly to the cause as well.[12]

[11] For details, see Section 1.4.

[12] See Appendix 1.A for further discussion of how the direct impact of posts could be added to the model.

I use a Cobb-Douglas utility function to represent the decision problem as follows.[13]

$$
\begin{aligned}
max_{d_i,p_i} \quad & \alpha \ln(y_i) + \omega \ln(D) + \gamma \ln(x_i) + \delta p_i \\
& \text{s.t. budget constraint } d_i + x_i \leq I \\
& \qquad \text{production fn } y_i = d_i + \kappa^* p_i \\
& \text{where } p_i = \{0, 1\};\ D = \sum_{i=1}^{n} d_i
\end{aligned}
$$

The parameter α gives the warm glow and altruism benefits of one's own contribution to the cause, while ω measures the relative utility weight on total contributions from the group. The parameter γ is the weight placed on money kept, while δ represents the net private benefits of posting. For instance, δ captures personal benefits of reputation, being seen as a supporter of the cause, being part of a group, etc., and personal costs like social anxiety (I don't post often or the risk of a debate with friends who have different political views).

This model scales up the typical donate-keep tradeoff to a donate-keep-post tradeoff. This joint donation-post decision problem is particularly interesting since one decision is private and has a known impact in dollar terms, while the other is public and beliefs about its effectiveness may vary. The budget constraint shows that monetarily, income (I) is only divided between the self and the charity because the post is free to send.[14] At the same time, posting can affect an individual's donation decision, their utility, and others' donations. Total donations are the sum of one's own donation plus others' donations, which may depend on your own posting choice, $D = \sum_{i=1}^{n} d_i = d_i + D_{-i}(p_i)$.

In the absence of social media, an agent is only maximizing over their private donation choice. Donations are private, so person i does not know how much others are donating but makes their own decision conditional on their expectation of what others will donate. I will denote all others excluding individual i with $\{-i\}$ and expectations using the hat notation, so $\hat{D}_{-i} = E_i(D_{-i})$. Without posts, the utility function reduces to the standard decision problem and the optimal dona-

[13] The posting decision, p_i, is binary so taking the natural log is unnecessary and only changes the interpretation of δ. Note that a Cobb-Douglas utility function is simply a representation that makes the various components tangible, but other utility functions could be used.

[14] The cost of donating is 1 and the post is free to send, so these costs are left out of the budget constraint for simplicity. The costs could be included in a general way to translate them to other settings.

tion amount depends on income, how much they expect others to donate, how much they care about the cause, the warm glow they derive from contributing, and the weight they place on consumption utility (Eqn 1.1).[15]

$$\begin{aligned} d_i^*(\hat{D}_{-i}) = \frac{1}{2}\left(\left(\frac{\alpha+\omega}{\alpha+\omega+\gamma}\right) I - \left(\frac{\alpha+\gamma}{\alpha+\omega+\gamma}\right)\hat{D}_{-i}\right) \\ + \frac{1}{2}\left(\frac{1}{\alpha+\omega+\gamma}\right)\left\{\left(-\hat{D}_{-i}(\alpha+\gamma) + I(\alpha+\omega)\right)^2 + 4\hat{D}_{-i}I\alpha\left(\alpha+\omega+\gamma\right)\right\}^{\frac{1}{2}} \end{aligned} \tag{1.1}$$

With posts, agents maximize over both their private donation and their public post choice. Solving the model yields the following optimal donation and posting decision rules.

The optimal donation decision with digital posts of support has an extra negative term from the impact of posting reducing the need to donate, along with additional interaction terms, shown in blue (Eqn 1.2). If people post and the impact of that post is positive, then their optimal donation decision is lower than in the world without posts. This is represented by the negative κp_i term. This effect is logical. For instance, since person i was able to contribute κ to the cause costlessly, they can donate κ less and keep \$$\kappa$ more, while contributing the same amount to the cause as before.[16]

$$\begin{aligned} d_i^*(p_i, \hat{D}_{-i}) = \frac{1}{2}\left(\left(\frac{\alpha+\omega}{\alpha+\omega+\gamma}\right) I - \left(\frac{\alpha+\gamma}{\alpha+\omega+\gamma}\right)\hat{D}_{-i} - \left(\frac{\omega+\gamma}{\alpha+\omega+\gamma}\right)\kappa p_i\right) \\ + \frac{1}{2}\left(\frac{1}{\alpha+\omega+\gamma}\right) * \left\{\left[-(\alpha+\gamma)\hat{D}_{-i} + I(\alpha+\omega) - (\omega+\gamma)\kappa p_i\right]^2\right. \\ \left. + 4(\alpha+\omega+\gamma)\left(\hat{D}_{-i}I\alpha + I\kappa p_i\omega - \hat{D}_{-i}\kappa p_i\gamma\right)\right\}^{\frac{1}{2}} \end{aligned} \tag{1.2}$$

The optimal posting rule depends on a number of factors. Conceptually, people post when posting makes them better off. This equates to posting when the resulting net change in the components multiplied by their utility weights is positive (Eqn 1.3).

[15] The model being solved is $\alpha\ln(d_i) + \omega\ln(d_i + \hat{D}_{-i}) + \gamma\ln(x_i)$. See Proof 2 in Appendix 1.A.

[16] See Appendix 1.A, Proof 4 and 3 for the derivation of Eqn 1.2 and that $d_i^*(p_i, \hat{D}_{-i})$ is decreasing in p_i. Notice that someone who is given the option to post but chooses not to has the same optimal donation amount as they would without the option to post, holding parameters fixed. In reality, people's posting decision is endogenous, meaning the causal comparison we can make using the data is between those who have the option to post and those who do not.

$$p_i^*(\hat{D}_{-i}) = \begin{cases} 1 & U\left(d_i^*(p_i=1), p_i=1, \hat{D}_{-i}(p_i=1)\right) > U\left(d_i^*(p_i=0), p_i=0, \hat{D}_{-i}(p_i=0)\right) \\ & i.e.,\ \alpha \ln\left(\frac{d_i^*(p_i=1)+\kappa}{d_i^*(p_i=0)}\right) + \omega \ln\left(\frac{\hat{D}_{-i}(p_i=1)}{\hat{D}_{-i}(p_i=0)}\right) + \gamma \ln\left(\frac{I-d_i^*(p_i=1)}{I-d_i^*(p_i=0)}\right) + \delta \geq 0 \\ 0 & otherwise \end{cases} \tag{1.3}$$

Returning to our definition of slacktivism, we have the phenomenon where someone posts because their utility is higher when posting, $(U|p_i = 1) > (U|p_i = 0)$, but in doing so they reduce their own contribution to the cause $(y_i|p_i = 1) < (y_i|p_i = 0)$. Two intuitive conditions must be met for slacktivism to occur. First, people have to want to post, making the first condition the same as the optimal posting decision rule. Second, an individual's total contribution needs to be lower when they post than when they don't post. In other words, person i reduces their donation amount because they posted, and that reduction is larger than the impact of their post. Note that donations being crowded out by posts is not sufficient for slacktivism, we need to analyze total own contributions to the cause.

Slacktivism Condition 1: Want to post

$$U\left(d_i^*(p_i=1), p_i=1, \hat{D}_{-i}(p_i=1)\right) > U\left(d_i^*(p_i=0), p_i=0, \hat{D}_{-i}(p_i=0)\right), i.e.$$

$$\alpha \ln\left(\frac{d_i^*(p_i=1)+\kappa}{d_i^*(p_i=0)}\right) + \omega \ln\left(\frac{\hat{D}_{-i}(p_i=1)}{\hat{D}_{-i}(p_i=0)}\right) + \gamma \ln\left(\frac{I-d_i^*(p_i=1)}{I-d_i^*(p_i=0)}\right) + \delta \geq 0$$

$$\text{where } \hat{D}_{-i}(p_i=1) = \hat{D}_{-i}(p_i=0) + \kappa$$

Slacktivism Condition 2: Reduce individual total contribution

$$y_i^*(d_i^*(p_i), p_i\, |p_i=1) < y_i^*(d_i^*(p_i), p_i\, |p_i=0), i.e.$$

$$d_i^*(p_i=1) + \kappa < d_i^*(p_i=0)$$

There are a few discussion points around the slacktivism conditions.

First, if people post and reduce their own contribution such that condition 2 is met, then the first term of condition 1 is negative but condition 1 can still be satisfied. People could crowd out their contribution but still want to post due to a combination of the other terms being sufficiently positive. This could be from thinking posts sufficiently influence others' donations, being able to keep more money for themselves, and reputation gains from posting. The higher the net private benefits of posting (δ), the more likely someone is to post.[17]

Second, the slacktivism conditions depend on beliefs, which may or may not be accurate, so people may engage in unintentional slacktivism. How an individual thinks their post impacts others affects their understanding of their own contribution and their belief of others' donations. If senders are not told the impact of posts, then they are acting on their best estimate of how posts influence receivers' donations. People are likely to believe posts make a difference (Harlow & Guo 2014; Anderson et al 2018). They could overestimate the impact, so that they don't realize they have crowded out their own contribution. Mathematically, this scenario would mean i's own contribution is crowded out according to Slacktivism Condition 2 which is $d_i^*(p_i = 1) + \kappa < d_i^*(p_i = 0)$, but they don't think they are because their expectations of κ far exceed the truth, $\hat{\kappa} \gg \kappa$, so they don't think they are crowding out their own contributions such that $d_i^*(p_i = 0) \leq d_i^*(p_i = 1) + \hat{\kappa}$. On the other hand, intentional slacktivism would be someone who knowingly reduces their contribution to social impact relative to if they didn't post. While it is possible to have this behavior by a rational agent in the model, the conditions are easier to satisfy for agents who reduce their donations by larger amounts due to inaccurate beliefs.

The model formalizes the idea that the impact of posting on others is critical to the valuation of one's own total contribution, and therefore the existence of slacktivism. The model also highlights that beliefs of the impact of post are important for determining behavior. As a result, I elicit beliefs about the impact of posts from participants in an incentivized manner.

Studying slacktivism provides insights into additional theories of behavior. First, consider warm glow, which is typically thought of as an emotional satisfaction from taking prosocial action (Andreoni 1990). If people show costless support for a cause and uncertainty exists about its

[17] If there are no private benefits to posting, it is still possible to have slacktivism but will be harder to satisfy Slacktivism Condition 1. If it was costly to post support, it would be easier to satisfy condition 2 since donations would be further reduced through the budget constraint, but it is not necessary for the slacktivism conditions to be met.

effectiveness, it is unclear whether they will experience warm glow even if they don't donate. My results suggest that costless signaling can provide warm glow.[18] Second, beliefs are extremely important motivations of behavior (Battigalli & Dufwenberg 2022; Bicchieri et al 2023). Beliefs about a post's impact are crucial to donation decisions and slacktivism, and this setting offers a window into the extent of self-interested heuristics present in people's beliefs. People who show support for a cause often care about its success, but I find that motivated reasoning leads people to overestimate their post's impact to justify under-donating even among supporters.

This model can also act as a framework that can be adapted to think about other contexts with visible, low-effort actions and private, high-cost actions.

Given the extent of media attention on slacktivism, previous correlational studies, these model predictions, and the role of beliefs, I expect to find a non-zero fraction of people engaging in slacktivism. Not only are beliefs key, the true impact of a digital post of support outside the lab is unknown, so belief construction is essentially unbounded. I hypothesize that at least some people will crowd out their individual contribution to racial justice when they have the ability to send digital posts of support.

1.3 Experimental Design

I use a between-subject laboratory experiment to identify how the signaling environment affects private donation decisions and total individual contributions to the cause.[19] The randomized, experimental setting allows me to observe and control for factors that we could not with observational data. A private donation is the costly action of support a participant can take, while a public, digital post of support mimics a social media post.

Sessions for the main sample were conducted in the EconLab at the University of California

[18] This is in line with the expanded model by Evren & Minardi (2017), which suggests that people might feel warm glow from costless action if they believe it enhances their reputation for being prosocial.

[19] I use a between-subject rather than a within-subject design since posts of support provide information about others. Multiple rounds would be influenced by posts observed in previous rounds. Additionally, people could use multiple rounds of donation decisions to hedge their payoffs across rounds.

San Diego (UCSD) from March to October 2022 and coded using oTree (Chen et al 2016). The follow-on nudge experiment was conducted between July and October 2023 in the same EconLab. Participants are undergraduate students at UCSD. Each session lasted 35 to 45 minutes and had an average of 16 participants. Sessions were conducted in-person to encourage people to consider their social image even though it is a short-lived interaction, which may matter for slacktivism.

Subjects receive a show-up payment of $6 and an additional $12 that they can choose to keep or donate to three racial justice-oriented charitable organizations. Individuals can donate any amount between the minimum and maximum to allow testing of both the extensive margin (did they donate or not) and the intensive margin (amount donated). These amounts were chosen to fairly compensate subjects for their time, be large enough to be taken seriously, and minimize clustering at heuristic norms such as $5 and $10.[20] Donation decisions are private; they are never revealed to the other subjects or the experimenter.[21] Any amount that they donate is sent directly to the charities. This real donation method is commonly used in behavioral economics studies and was first used in Eckel & Grossman (1996). Subjects know the amount they donate is subtracted from their earnings, and they are invited to stay after the session finishes and watch the experimenter send the money directly to the charities.

If an individual chooses to donate a positive amount, they choose which of the following 3 organizations they would like to donate to: Equal Justice Initiative, NAACP Legal Defense Fund, and Race Forward. They can choose to donate to 1, more than one, or all three charities. This flexibility is used to encourage donations, with their donation split equally to their chosen charities. These organizations complement each other to reflect the Black Lives Matter (BLM) movement.[22] The instructions provide the mission statement and action items of each organization.

[20] The earliest sessions had a show-up fee of $4 and an additional $8 that they could donate or keep. The increase to $6 and $12 was made to encourage more people to sign up for sessions, while keeping the show-up fee to donate/keep earnings ratio constant. All donation results will be shown as scaled donations so the maximum amount for every subject is $12. Results do not change if I instead use raw donation amounts.

[21] The private nature of donations is made clear in the instructions and confirmed as a comprehension question that must be answered correctly to continue.

[22] I use a charity portfolio because Black Lives Matter is seen as a nebulous and there is lack of clarity on how donation dollars are used. The BLM movement encompasses a large range of efforts and certain charities may not be as well-known. This set balances familiarity and racial justice methods.

All participants can support racial justice causes by giving money privately, while some can also announce or observe others' support publicly. Individuals in the control group are only given the option to donate money privately. Subjects in the treatment sessions are randomly assigned to be a sender or a receiver of digital posts of support, in addition to the same donation choice as the control.

Senders have the option to publicly announce their support via a digital "pop-up" message shown on the receivers' screens, to emulate a social media post. The post is a yellow text box that says a participant (indicated by the desk number they are seated at) supports racial justice, along with an image of a black fist, a well-known depiction of solidarity and anti-racism. Posting in the lab experiment is free to imitate social media which does not cost money and can be done in seconds.

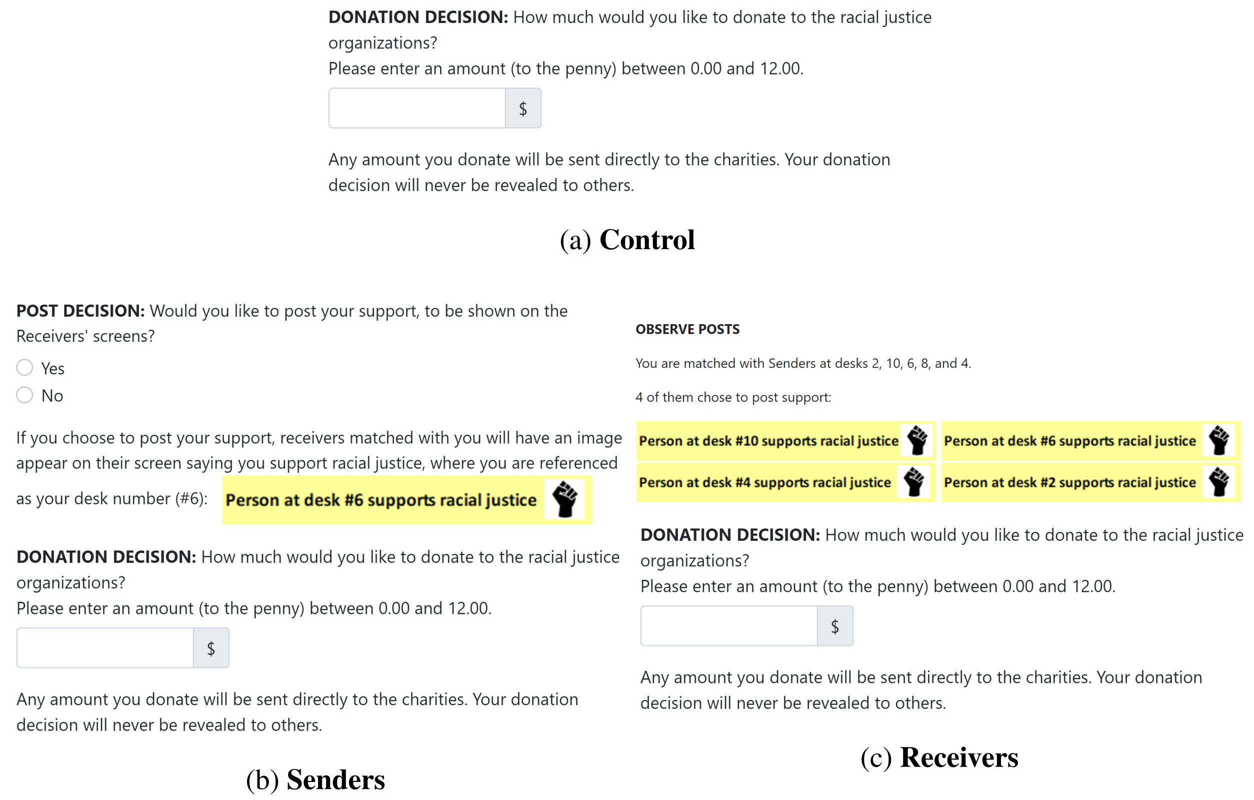
Receivers do not have the ability to send digital posts of support, but rather observe any posts from 5 randomly selected senders in the room. Each receiver is shown any posts from the 5 senders they are matched with. This creates random variation in the number of posts that receivers observe, and allows me to identify the impact of a post in this context. Specifically, the impact of a post is the amount that the average receivers' donation decision responds to each additional post they observe.

The experimental setup is as follows.[23] All participants read instructions, answer comprehension questions, are informed that they are a sender or a receiver (if treated), make/observe posting choices (if treated), make their donation choice, report their beliefs, and complete an exit survey. Instructions are read aloud so that senders and receivers know that everyone in the room understands the posting environment. The survey asks about demographic information, thoughts during the experiment, and related actions outside of the lab.

The decision-making portion of the experiment for each role, including images of the posts, is shown in Figure 1.1. Senders make their donation and posting decisions at the same time on the same screen. The post option is independent of the donation decision; subjects can post and not

[23] See Appendix 1.B for more details about the experimental design.

donate, or any other combination of the 2 (post) x 2 (donate) decision matrix. This independence is key to imitate social media where posts are unverified.[24] All senders finish making their decisions before receivers make their decisions. Receivers observe posts from the senders they are matched with on the same screen as their own donation choice.



DONATION DECISION: How much would you like to donate to the racial justice organizations?
Please enter an amount (to the penny) between 0.00 and 12.00.

$

Any amount you donate will be sent directly to the charities. Your donation decision will never be revealed to others.

(a) **Control**

POST DECISION: Would you like to post your support, to be shown on the Receivers' screens?

○ Yes
○ No

If you choose to post your support, receivers matched with you will have an image appear on their screen saying you support racial justice, where you are referenced as your desk number (#6): **Person at desk #6 supports racial justice** ✊

DONATION DECISION: How much would you like to donate to the racial justice organizations?
Please enter an amount (to the penny) between 0.00 and 12.00.

$

Any amount you donate will be sent directly to the charities. Your donation decision will never be revealed to others.

(b) **Senders**

OBSERVE POSTS

You are matched with Senders at desks 2, 10, 6, 8, and 4.

4 of them chose to post support:

Person at desk #10 supports racial justice ✊ **Person at desk #6 supports racial justice** ✊
Person at desk #4 supports racial justice ✊ **Person at desk #2 supports racial justice** ✊

DONATION DECISION: How much would you like to donate to the racial justice organizations?
Please enter an amount (to the penny) between 0.00 and 12.00.

$

Any amount you donate will be sent directly to the charities. Your donation decision will never be revealed to others.

(c) **Receivers**

Figure 1.1: **Decision Screens: All subjects have the same donation decision, senders have an additional post decision, and receivers observe posts**

Subjects take an exit survey after making all experimental decisions. I use 7-point Likert scale questions to ask about underlying support for the cause, beliefs of how posts influenced themself and others, and their history of support. The survey includes standard demographic questions about age, gender, race, and background. These are important to ask because research has shown that donation decisions vary by gender, age, education, and sexual orientation (Aksoy et al 2023; Eckel & Grossman 1998; Piper & Sylke 2007; Skidmore & Sellen 2021; WPI 2022). Information

[24] This differs from many papers in the literature where a public display of support is shown only if the individual makes a minimum contribution (e.g. someone who donates at least $100 gets their name on the back of the orchestra program).

on the subject's race and background is important since the charities will be U.S. focused and racial justice oriented. Previous research on slacktivism shows that attitudes may vary by race (Anderson et al 2018). In addition, there are free response questions about why subjects made the decisions they did.

I elicit beliefs of others' actions using the binarized scoring rule (BSR) incentive scheme with a short, intuition-focused explanation. I use the BSR since it is incentive compatible even for people who are not expected utility maximizers or risk neutral (Erkal et al 2020; Hossain & Okui 2013; Schotter & Trevino 2014). I keep the explanation as short and intuitive as possible since previous research has found that there is more truth telling when less information about the specifics of the BSR is provided (Wilson & Vespa 2016; Danz et al 2022). Since I conduct sessions in a no-deception lab, a detailed explanation of the BSR payment scheme is provided via an optional pop-up box.[25] Participants could earn a bonus $2 per question if their squared error was less than the random number drawn between $[0, \bar{K}]$, and $0 otherwise.[26] Subjects were not aware that they would have the opportunity to earn bonus payments after their decisions. Participants were asked to report their best guess of the average amount donated by everyone else in the room, the number of people who posted in the room (if treated), and the average change in receivers' donations for each additional post of support observed (if treated and sender). For each question, $\bar{K}$ was chosen as the maximum possible squared error such that everyone had the possibility of receiving the bonus payment.[27]

Each role provides valuable insights, and the confluence allows me to identify slacktivism. The control experiment is simple and establishes a baseline of donation choices to the racial justice charities in the absence of low-effort, visible actions. The senders give us donation choices for those who have the option to post. We can also compare donation behavior for those who post

[25] Fewer than 25% clicked for more information. This is balanced across treatment (control, treated), role (sender, receiver) and information session (no info, info) status.

[26] I chose $2 per question to induce cognitive effort since it is large relative to previous payments (Burdea & Woon 2022).

[27] For example, the $\bar{K}$ for the average amount donated by everyone in the room was 144 ($= 12^2$) when people could donate between $0 and $12. Of those asked each question, 92, 97, and 100% of subjects received the bonus payment, respectively. This equates to an extra $2 - $6, depending on role.

and don't post, and measure senders' beliefs about the impact of posts when no information was provided. At the same time, the receivers' donation decisions and how they vary by the number of posts observed identifies the impact of posts in this setting.

To better investigate mechanisms and ways to improve donations, I have a treatment arm that provides senders with information about the impact of posts using data from previous sessions. One-third of treatment sessions are no information sessions (No Info) and two-thirds provide information (Info). The difference between the no information and information treatment sessions is minor, but has large benefits to understanding mechanisms and the role of beliefs. Senders in the information sessions are told that "Using data from previous sessions, we see that *Receivers donate $x more on average for each additional post of support they observe from Senders*." I wanted to study how different information affects behavior without using deception, so I continuously calculated the $x with updated data from all previous sessions. Given the random timing of when information treatment arms were conducted, the data-driven values bounced around the true value of zero. The values senders in the information sessions were told ended up being $x \in \{\$0.56, 0.16, 0.07, -0.11, -0.38\}$. When x is negative, the word 'more' is replaced with 'less' and the value is repeated in parentheses with a minus sign to ensure clarity, for example "*Receivers donate $0.11 less on average (-0.11).*"[28] Information session senders are told this value when they find out they are a sender, which is before they make their donation and posting choices. They then see the same decision-making screen as senders in the No Info treatment arm. The behavior of the senders with information let us observe how beliefs and donation behavior reacts to signals about the impact of posts.

Overall, these treatment variations provide the necessary components to test for slacktivist behavior resulting from publicly observable posts of support. Treatment variations disentangle the role of being able to send versus observe posts, beliefs about others, and beliefs of the effectiveness of posts. The lab simulates how social media affects an individual's donation behavior and net contribution towards non-profit, social justice organizations.

[28] This is done to ensure that the negative sign is salient since it goes against most people's expectations.

There are 322 subjects in my main sample, with 73 in the control condition, 126 senders, and 123 receivers. Subjects are balanced by observable characteristics, including demographics and underlying support for the cause (Appendix Table 1.C.1). The subject pool matches that of the UCSD undergraduate economics major population, with majority Asian or Asian American individuals, roughly half men and half women, and a substantial portion of international students.

1.4 Results

This section breaks down how the social media environment affects participants' donation behavior. The main sample consists of control and treated individuals made up of senders and receivers across information and no information sessions.[29] After establishing a baseline of donations from the control group, I explore donation behavior among those given the option to post. All results shown pool senders across information and no information sessions, unless otherwise noted. This is done for precision, and is reasonable given the similarity of behavior across treatment groups. Then I look at how often senders post, and how that correlates with donation behavior. Next I study the impact of posts on others' donations. I combine the two pieces of the slacktivism puzzle to get an individual's net contribution. I present mechanisms driving slacktivist behavior and heterogeneous effects. Finally, I discuss whether informing people about slacktivism can mitigate the effect and other policy implications.

I use ordinary least squares (OLS) for continuous measures and logit models for binary outcomes. I estimate causal effects whenever comparing categories that were randomly assigned. All regressions control for variables that are significant predictors of donation behavior and increase my precision. My preferred specification controls for age, gender (man, woman, other gender), and answers to 7-point Likert scale questions of how often they donated in the last two years, how often they participated in offline activism in the last two years, how much they agree/disagree with the statement 'I want to be one of the people who contributes', and 'I need the money'.[30] I refer to

[29] The follow-on nudge experiment is not included in the main sample, but analyzed in Section 1.4.4.

[30] The latter two questions are proxies for the α and γ model parameters, respectively.

these as demographics, past activism, and parameters in the tables going forward.

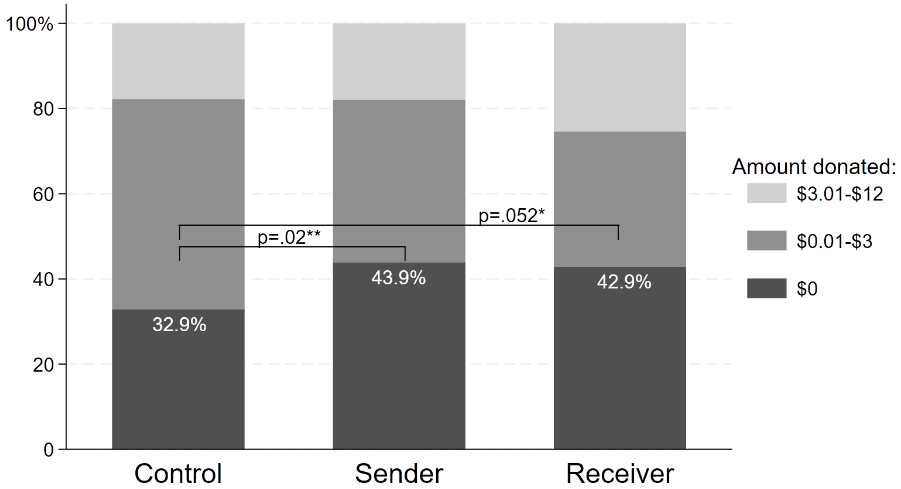
1.4.1 Slacktivism Does Occur

Control individuals give us a baseline level of donations, where one-third do not donate and the majority of people (79.2%) donate $3 or less (Figure 1.2). Having two-thirds of control individuals donate something is in line with the number of individuals who typically donate to charitable organizations (Philanthropy Roundtable n.d.). Among those who donate, almost all donate at salient whole dollar increments and most donate a small amount ($1, $2, or $3), although there is a large right tail with some individuals donating the maximum amount (Appendix Figure 1.C.1).

Senders are significantly less likely to donate than control individuals (Figure 1.2). The ability to show support for the cause induces this extensive margin crowd out effect, where 44% of senders do not donate, up from 33% of control individuals. According to logistic regressions, those with the ability to show support are twice as likely to abstain from donating (Table 1.1). Since role is randomized, the ability to show support has a causal effect on donation crowd out. The increase in non-donators among senders is coming from those who would have donated $1 or $2 if they had been in the control group. People who would have donated a large amount without a social media environment continue to do so. Since high-donators are relatively unaffected, the impact of being able to post on the average donation of senders is negative but insignificant (Appendix Figure 1.C.3).

The posting decision is endogenous, but we are particularly interested in senders who choose to send a post of support. Under our definition of slacktivism, someone sends a digital post of support, and in doing so, crowds out their individual total contribution.

The majority of senders send a digital post of support for racial justice (75%). Those who post support say they did so because they want others to know they support the cause, they want to encourage others to give, or because it is free. Those who don't post want to be more private



Note: P-values are from a logistic regression of all observations on the likelihood to donate zero. Control variables are demographics (age, gender), parameters (alpha, gamma), and self-reported past activism (offline, donations).

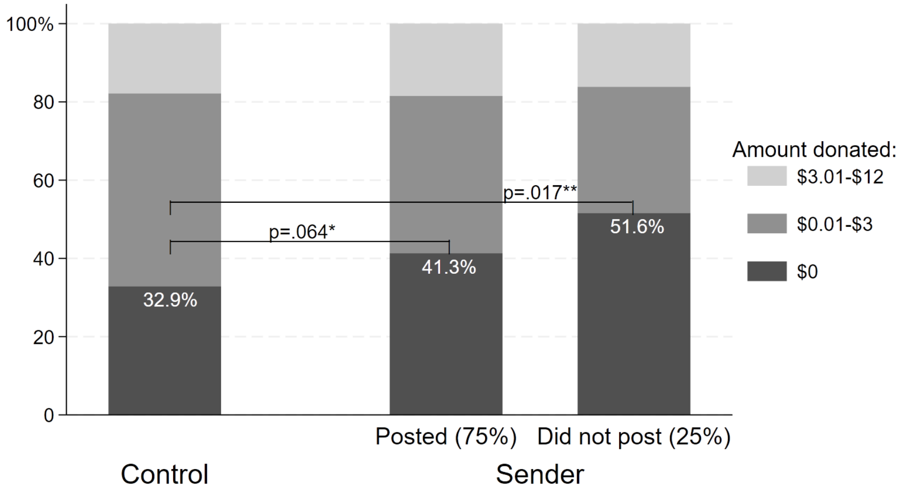
Figure 1.2: **Senders and receivers are significantly less likely to donate, with the change concentrated among those who would have donated a small amount**

and still report support for racial justice, but to a lesser degree.[31] Upwards of 82% of people post in the no information treatment. Fewer post in the information treatment, especially when they are informed that posts have a negative impact on others, although choosing to post remains the majority decision.

Both senders who post and senders who don't post donate less than two-thirds of the time, such that there appears to be crowd out even among the subsample of senders who want to show support (Figure 1.3). Posters and non-posters are selected subsamples, with non-posters being less strong supporters or prefering to be more private, while posters are stronger supporters or want others to know they support the cause. While only descriptive, having the option to post and taking that option generates a crowd out effect on donations of the extensive margin. This result is robust to using control individuals who hypothetically say that they would have posted if given the choice as the comparison group.[32]

[31] Participants answered a 7-point Likert scale question about how much they agree/disagree with "Racial justice is an important cause." No one disagreed with the statement. Those who didn't post were less likely to 'Strongly Agree' than senders who posted. The percent who strongly agree is similar across control, senders, and receivers. Free response answers corroborate this idea of weaker supporters.

[32] The exit survey asked control individuals if they would have posted if given the option. Less than half of control individuals said they would post, much less than the incidence of posts among senders. Therefore the hypothetical posting decision is likely a poor proxy for actual posting decisions. However if we compare senders who post to



Note: P-values are from a logistic regression of all observations on the likelihood to donate zero. Control variables are demographics (age, gender), parameters (alpha, gamma), and self-reported past activism (offline, donations).

Figure 1.3: **Both those who choose to post and not post donate less than two-thirds of the time**

The extensive margin crowd out is large. Logistic regressions show that those who can post are twice as likely to donate nothing relative to individuals without a social media environment (Table 1.1). Even the subsample of senders who post are nearly twice as likely to not donate. Senders often use the posting option, and both those who post and those who don't are less likely to donate.

For the second piece of the slacktivism puzzle, we turn to receivers to quantify how posts influence others' donations.

Receivers are also more likely to abstain from donating relative to control individuals (Figure 1.2 and Table 1.1). Receivers cannot post their support, so there is something else about the posting environment that is causing receivers to donate less often.

Since receivers are randomly matched with 5 senders, there is random variation in the number of posts observed by each receiver. Most receivers observe 3 to 5 posts of support, with the modal receiver observing 4 posts from their 5 matched senders.[33]

I exploit this variation to measure the marginal impact of observing one additional post on receivers' donations. I find that seeing an additional post does not encourage receivers to give more

control individuals who would post if given the chance, results hold since donation behavior is nearly identical between hypothetical posters and non-posters in the control.

[33] No one observes 0 or 1 post in this sample.

Table 1.1: **Both individuals with the option to post and those who see others' posts are twice as likely to donate $0**

	Donate Zero (Odds Ratio)				
	(1)	(2)	(3)	(4)	(5)
Sender	1.598	1.689*	1.980**	2.213**	
	(0.493)	(0.531)	(0.659)	(0.755)	
Posted					1.953*
					(0.706)
Did not post					3.157**
					(1.515)
Receiver	1.531	1.655	1.800*	1.934*	1.937*
	(0.471)	(0.519)	(0.600)	(0.656)	(0.657)
N	322	321	313	313	313
Pseudo R^2	0.006	0.020	0.083	0.110	0.113
Demographics		Y	Y	Y	Y
Parameters			Y	Y	Y
Past Activism				Y	Y

Exponentiated coefficients; Standard errors in parentheses. Relative to control individuals. Demographics are age and gender identity. Parameters are 'I want to contribute' (alpha) and 'I need the money' (gamma). Past activism behavior are answers to 7-point Likert scale questions of how often donated in the last two years and how often participated in offline activism. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

along the extensive nor intensive margin. Receivers who see more posts donate at similar rates as those who see fewer posts (Figure 1.4) and the average effect of posts on others' donation amount is not statistically different from zero (Table 1.2).[34] I can rule out posts having a positive effect of 27 cents or more. So in this lab experiment setting, the best estimate of the impact of a post is zero.

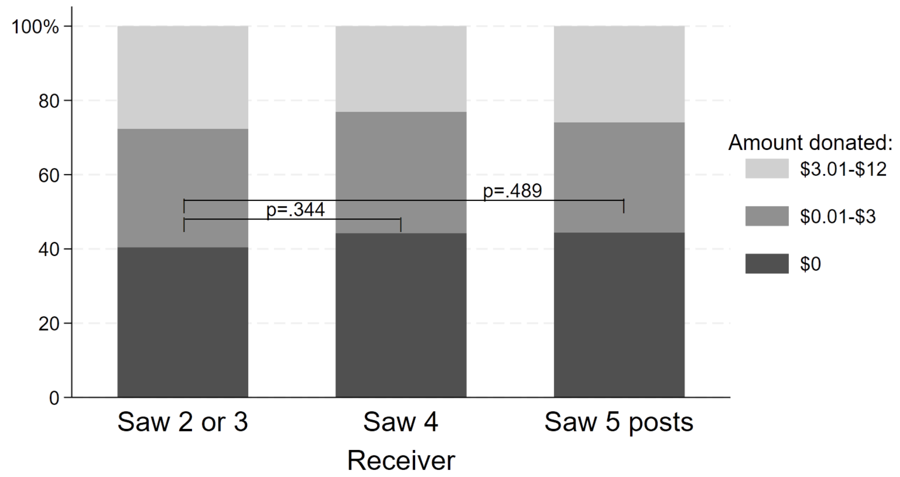
Table 1.2: **Each post has a statistically zero effect on the average donation of receivers**

	Donate Amount ($)			
	(1)	(2)	(3)	(4)
Number of posts observed	-0.349	-0.393	-0.375	-0.406
	(0.366)	(0.373)	(0.344)	(0.342)
N	126	126	121	121
Adjusted R^2	-0.001	-0.018	0.145	0.157
Demographics		Y	Y	Y
Parameters			Y	Y
Past Activism				Y

Standard errors in parentheses. Relative to other Receivers. Demographics are age and gender identity. Parameters are 'I want to contribute' (alpha) and 'I need the money' (gamma). Past activism behavior are answers to 7-point Likert scale questions of how often donated in the last two years and how often participated in offline activism. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

The null impact of a marginal post implies that when senders crowd out their donations,

[34] The regression that uses the random variation in number of posts observed on the average donation of receivers was preregistered as the way I would measure the impact of posts, see AsPredicted #90396.



Note: P-values are from a logistic regression of all observations (including senders, not shown) on the likelihood to donate zero. Control variables are demographics (age, gender), parameters (alpha, gamma), and self-reported past activism (offline, donations).

Figure 1.4: **Receivers donate similarly regardless of the number of posts observed**

they crowd out their total contributions since their post does not have an effect on its own.[35] Hence, senders who post and no longer donate are indeed engaging in slacktivism.

The harms of slacktivism are large. The social media environment changes both senders and receivers donation choices, leading to an aggregate reduction of the donor base of 15% (Figure 1.5). The fraction of people who donate drops from two-thirds in the control to just over half in the treatment group.

1.4.2 Mechanisms

I have established that slacktivism occurs. Next I explore the role of beliefs, whether people think they are making a difference by posting support, and what happens if we provide information. I then explore other mechanisms driving behavior, and find evidence of motivated beliefs, justification effects, and moral licensing. These behavioral heuristics combine to create a check-the-box attitude towards supporting racial justice.

A natural next question is whether senders know they are engaging in slacktivism and reducing their individual net contribution when they reduce their donations since their post is not

[35] Posts and participating in the experiment more generally also does not have an effect on how people think themselves or others will behave next week. See Appendix Figures 1.C.5, 1.C.6, and 1.C.7 for details.

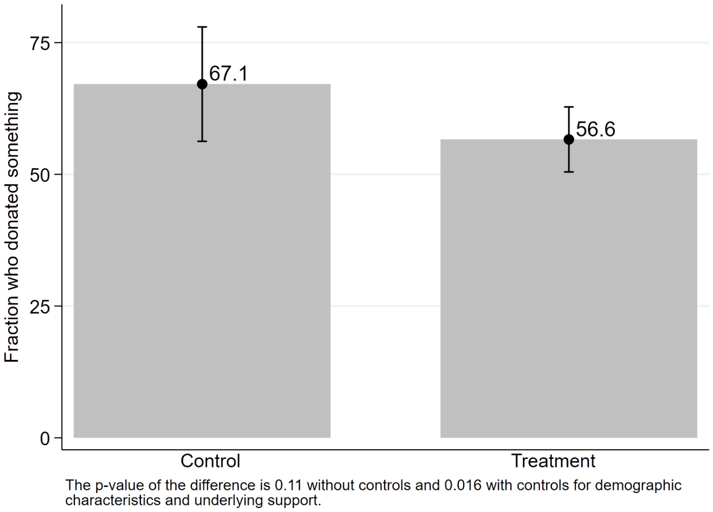


Figure 1.5: **The posting environment reduces the donor base by 15%**

impactful. In fact, one of the hypotheses of why we may see slacktivism is that people will feel like they are making a difference by posting. In this way, they think they are having an impact and so any individual total contribution crowd out is unintentional.

I do find that people vastly overestimate the impact of posting. Elicited beliefs show that most believe posts increase others' donations even though the true effect is zero (Figure 1.6). People who are not given information about the impact of a post report that they believe each additional post will increase every receivers' donation by $1.04 on average. Even though the true effect is zero, about 85% of senders in the no information sessions believe that the impact is positive. Overestimation persists among those who are told the real-time estimate of the impact of posts on receivers. Senders who are told that posts have a small positive effect on receivers' donations overestimate the impact of a post to a lesser degree ($0.41 on average), but a similar proportion think the impact is positive (86%). Senders who are told posts have a negative impact also revise their beliefs downward with fewer believing that the impact of the post is positive (64%), although still the majority.[36]

However, what people think about a post's impact is not a major factor in their donation

[36] Recall that the positive values senders could be told were 7, 16, or 56 cents. The negative values are -11 and -38 cents. For more details, see Appendix Table 1.C.4.

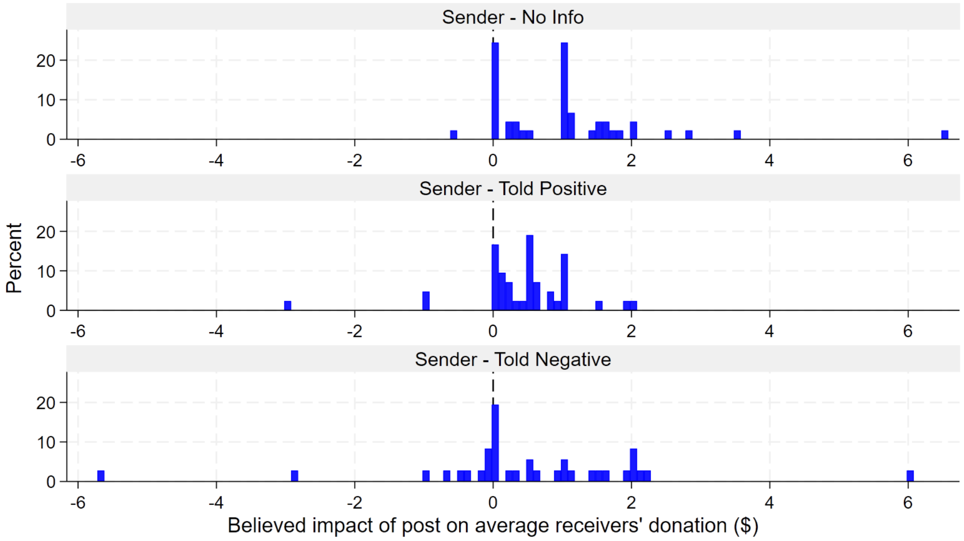


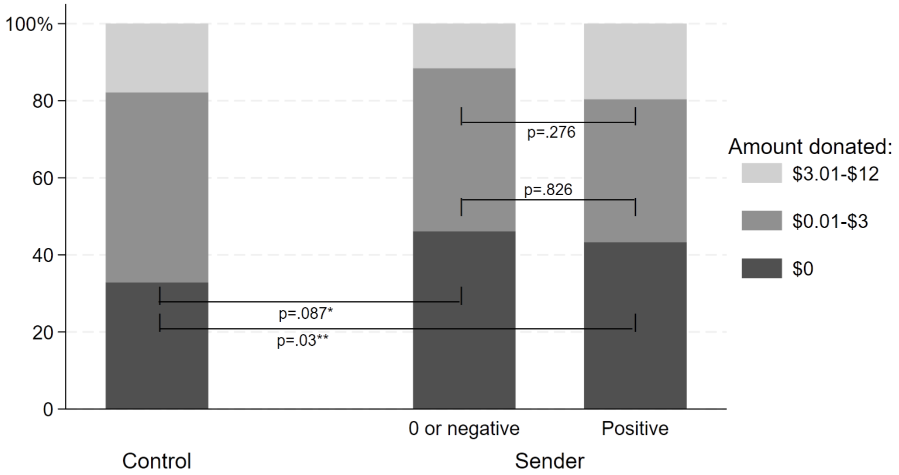
Figure 1.6: **The majority of senders overestimate the impact of a post, whose true effect is 0. Providing information reduces the amount of overestimation**

decision. Senders who think posts have a negative impact are less likely to post (2 out of 5 post) compared to those who think posts have a zero or positive impact (4 out of 5 post), but there is no such correlation among donation decisions. Senders who think the impact of posting is zero or negative have a similar propensity to not donate, donate a small amount, or donate a large amount relative senders who think the impact of posting is positive (Figure 1.7). If we separate senders with negative and zero views or focus on senders who sent a post of support, we see a similar story.[37] If we split senders by the information they were told instead of their reported beliefs, which doesn't suffer from endogeneity, we see the same lack of evidence of people using posts as a one-for-one replacement for donations, although this comparison has less power since there is variation in beliefs within each group of information told.[38] This means that people are not substituting a post for an equivalent dollar-value reduction in donations. Rather, there is something about posting that, regardless of how impactful someone believes it to be, they can use it to justify not donating anything.

Related, receivers observe posts but this doesn't make them think others are donating more.

[37] See Appendix Figure 1.C.8 and Appendix Figure 1.C.9, respectively.

[38] See Appendix Figures 1.C.10 and 1.C.11. Beliefs are influenced by the randomly assigned Information Treatment (Figure 1.6).



Note: Senders split by their beliefs of the average impact of posting on Receivers' donations. P-values are from a logistic regression of all observations on the likelihood to donate zero (or donate more than $3). Controls are demographics (age, gender), parameters (alpha, gamma), and self-reported past activism (offline, donations).

Figure 1.7: **Senders who believe posts have a non-positive impact crowd out donations to the same degree as those who think posts have a positive impact**

Receivers, senders, and control individuals all have similar estimates of how much others are donating (Appendix Figure 1.C.12). The intensive margin of the number of posts observed also doesn't move beliefs of the average amount donated by others in the room (Appendix Fig 1.C.13). This implies that receivers are not crowding out according to a simple free riding story.

While beliefs about the monetary value of a post do not correlate with donation behavior, I do find suggestive evidence of motivated reasoning impacting donation behavior. I find that roughly half of senders hold beliefs consistent with motivated reasoning about the impact of posts. The majority of senders recognize that posts don't influence their own donation, but continue to think posts increase others' donations (Figure 1.8).[39] These beliefs are consistent with self-interested motivated reasoning because they allow people to donate less while believing they are making more of a difference than they are. At the individual level, 43% of senders report guesses consistent with motivated beliefs.

Moreover, there is some correlation between holding beliefs consistent with motivated reasoning and donation crowd out. Those who believe that posts won't influence themselves but will

[39] Figure 1.8 shows both senders and receivers. The pattern holds if we look at senders and receivers separately (Appendix Figure 1.C.14).

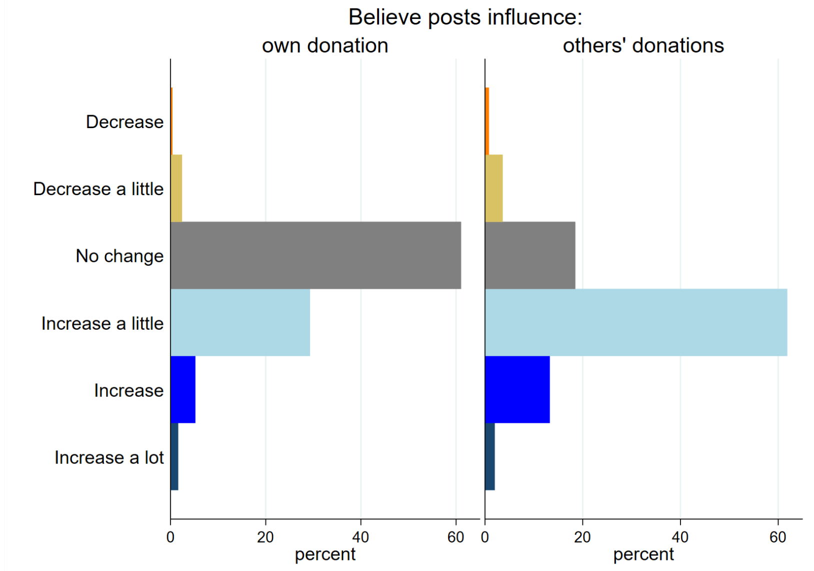


Figure 1.8: **Individuals in the social media environment hold motivated beliefs, where they believe posts influence others' donations more than their own**

influence others (i.e. those with motivated beliefs) donate less often, while those who are optimistic about posts influencing both themselves and others donate similarly to the control group (Figure 1.9, middle bars).[40] There is a third group of people who are pessimistic and recognize that posts wouldn't influence themselves nor others. This is a smaller group of people, but they don't hold motivated beliefs and are less likely to donate, suggesting that motivated beliefs are part, but not all, of the story.

Interestingly, receivers also hold motivated beliefs but the tie between these beliefs and their donation decisions is not as strong (Figure 1.9). This mechanism is logical in that people have the ability to convince themselves that posts will influence others in a certain way, such that they can justify not donating themselves. This idea is more relevant for people who can post support, and they exhibit a stronger motivated reasoning channel.

I also see evidence of motivated belief updating among senders. Senders who are told posts have a small positive impact are more than twice as likely to think posts influence others more than themselves relative to senders who were not given any signal about the impact of posts, while

[40] This figure shows all senders. Results are identical when focusing on senders who posted (Appendix Figure 1.C.15).

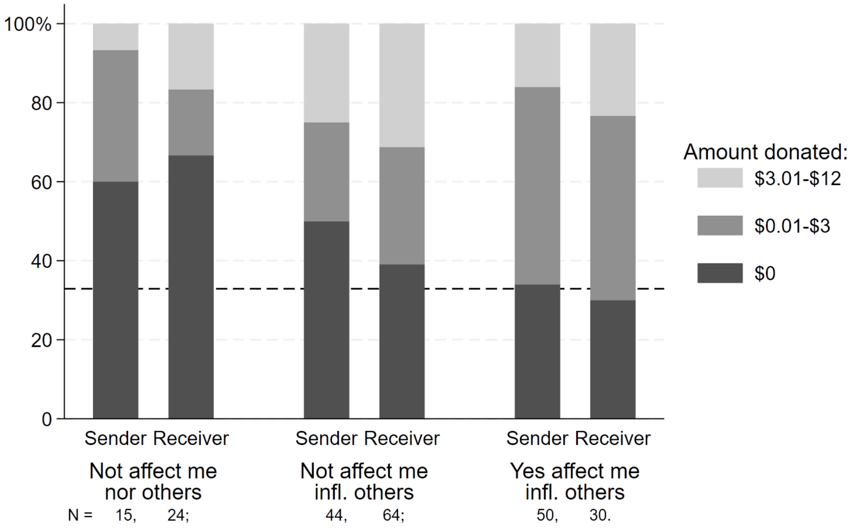


Figure 1.9: **The half of subjects who hold motivated beliefs that posts impact others more than themselves are more likely to be slacktivists than people who are optimistic about posts for both themselves and others**

those who are told a negative amount are not significantly less likely to hold this motivated belief (Appendix Figure 1.C.16). This means that when senders are given information that helps build their case for justifiably contributing less, they use it, but the same cannot be said for contradictory information.

Free response explanations of why they donated the amount they did gives suggestive evidence of additional justification effects for senders. First, senders write significantly shorter answers, suggesting that they feel less need to justify why they did or did not donate since they have already had the chance to post their support publicly (Appendix Figure 1.C.17).[41] Second, the treatment induces people who would have donated a small amount ($1 to $3) to not donate, and I observe a similar shift in mentions of financial need in their free response. This suggests that people with the same amount of financial need donate a little in the control, but don't donate when they can post. Since people were randomly assigned to role, it is not the case that there were simply more senders who needed the money.[42] But senders who don't donate are more likely to bring up

[41] Senders write 62 characters less on average (9 to 16 fewer words), which equates to 31% shorter answers.

[42] I find a similar amount of financial need overall within each session and across sessions, with very similar answers to the direct financial need question.

financial need in their free response explanation, with a proportional decrease among senders who donate a small amount.[43] Senders may feel more able to donate selfishly and say they need the money because they have already had the chance to show their support. Receivers cannot post, and accordingly, I do not see evidence of them giving shorter or different explanations of their donation choice.

Slacktivism is also concordant with moral licensing. Moral licensing is when "past good deeds can liberate individuals to engage in behaviors that are immoral, unethical, or otherwise problematic, behaviors that they would otherwise avoid for fear of feeling or appearing immoral." (Blanken et al 2015). Moral licensing is common in many contexts (Dütschke et al 2018; Monin & Miller 2001; Simbrunner & Schlegelmilch 2017), and charitable giving is no exception (Grieder et al 2021; Zhang & Peng 2022). While I don't have direct evidence of moral licensing, this literature helps rationalize my results. Posting support for racial justice is an example of an act that has a licensing effect on donations.

The social observability factor of posts of support is important. Rotella, Jung, Chinn, and Barclay (2023 forthcoming) show that moral licensing only occurs when participants are observed. They conclude that moral licensing is an interpersonal effect based on reputation, rather than an intrapsychic effect based on self-image. The fact that reputation matters has also been shown in the charitable giving and social media literatures. Kristofferson et al (2014) found that initial tokens of support that were public reduced subsequent more meaningful tasks, but initial tokens of support that were private actually increased subsequent meaningful actions of support. Lacetera et al (2016) ran a social media campaign field experiment that reached 6.4 million people but only garnered 30 donations. Fosgaard and Soetevent (2022) find that promising to donate later is a way for people to get out of the ask-situation with a positive image, but most do not donate later. Similarly, people avoid the ask when possible (Andreoni et al 2017; DellaVigna et al 2012). These

[43] An independent coder read all free response answers and categorized them into eight common types of reasoning. The coder did not know who was treated, who was a sender or receiver, what the treatment was, or how much they donated. The categorization is reasonable, with donators often saying they chose to donate because they care about the cause or they chose that amount because it was a fair or easily divisible amount. Non-donators more often state that they didn't donate because they need the money or they don't know or trust the charities (Appendix Figure 1.C.18). The shift in senders' free responses mentioning financial need can be seen in Appendix Figure 1.C.19.

highlight the desire to appear supportive.

A form of moral licensing fits for receivers as well. Vicarious licensing is moral licensing that uses the moral behaviors of others to justify a reduction in your own moral behaviors. Kouchaki (2011) and Meijers et al (2019) find evidence for vicarious licensing where people extend moral licensing behavior to the actions of others. Even though receivers cannot take action by sending a post, the presence and ability for others to post support leads receivers to be more likely to abstain from donating relative to control individuals. Meijers et al (2019) also finds that vicarious licensing occurs more often when others display supportive behavior. Since most senders post their support, the vicarious licensing effect may be larger than it otherwise could have been.

Overall, digital posts of support lead people away from donating through motivated reasoning, justification effects, and moral licensing. These mechanisms are not so surprising since they are present across many contexts. In line with the finding that donation crowd out is an extensive margin phenomenon, the mechanisms driving behavior also follow a check-the-box idea rather than an intensive margin shift.

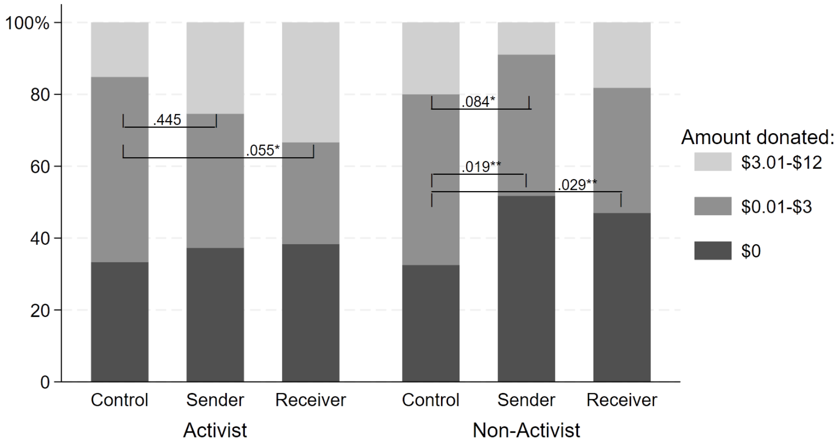
1.4.3 Differential treatment effects

The posting environment induces an increase in the proportion of individuals who do not donate, and there are differential treatment effects by history of activism and demographics.

The posting environment has opposite effects on activists and non-activists; only people who didn't participate in any offline activism in the last two years engage in slacktivism (Figure 1.10). For simplicity, I will refer to those who didn't participate in any offline activism in the last two years as non-activists.[44] The posting environment drastically reduces donations from non-activists but does not crowd out donations from activists, and may actually crowd-*in* donations

[44] When asked about offline activism, participants were given examples such as participating in protests and writing letters to Congress representatives. The last two years includes the summer of 2020 for most sessions, a period of time with many opportunities to engage in offline activism related to racial justice. I anticipated running all sessions before the end of the summer 2022, but due to challenges with the size of the EconLab participation pool, the sessions run in September and October did not technically include the summer of 2020. However, the distributions of the self-reported amount of past activism are nearly identical (visually and statistically) between the March to July and September to October sessions.

from some activists. Both senders and receivers who are non-activists are significantly less likely to give, with non-activist senders also being less likely to donate large amounts of money. Alternatively, those who have done some sort of activism in the last two years do not exhibit any slacktivist behavior. They are not less likely to donate nor do they donate smaller amounts of money. Even senders who post continue to donate at similar rates as they would have in the control environment (Appendix Figure 1.C.21). Further, receivers who are activists even marginally crowd-*in* and donate larger amounts of money. This is the only subgroup where I see any significant crowd-in of donations.[45]



Note: P-values are from a logistic regression of all observations on the likelihood to donate zero (or donate more than $3). Control variables are demographics (age, gender), parameters (alpha, gamma), and self-reported past donations. Activists by role are not statistically differently likely to donate $0.

Figure 1.10: **Non-activists (people who did not participate in offline activism in the past 2 years) engage in slacktivism; activists are unaffected and may even respond to others' posts by donating larger amounts**

The posting environment even has an effect on the average donation amount of activists and non-activists (Appendix Figure 1.C.22). When looking at all senders or all receivers, we did not see an effect on the average donation amount, but this masked important heterogeneity. Non-activist senders give significantly less on average, and activist receivers give larger amounts.

The differences between activists and non-activists is not present in the control group, meaning that it is the posting environment itself that impacts people with varying levels of past

[45] The crowd-in result is not strong enough to be robust. I do not see significant crowding-in among activist receivers in the nudge intervention (results shown in Section 1.4.4), so this result is interesting but secondary.

activism differentially. Activists and non-activists have nearly identical donation distributions in the control group. In the treatment group, however, the donation behavior of the two groups show very different patterns of giving.

In line with this, the mechanisms are more likely to be present among non-activists. Some activists and some non-activists have beliefs consistent with motivated beliefs (Appendix Figure 1.C.23), but non-activists act on them while activists do not (Appendix Figure 1.C.24). Non-activists also give even shorter explanations of why they didn't donate than similar activists (Appendix Figure 1.C.25). I interpret this as evidence that the 'check the box' mentality mechanism that turns on in the posting environment is more present for non-activists, making it easier for them to justify not donating relative to activists.

While non-activists are less central to social justice movements, we do care about their behavior. Non-activists are still supporters as some of them would have donated in the control environment. They are less active than others, but they remain an essential part of grassroots campaigns.

In addition to differences by past activism, I find that women are more affected by the posting environment, possibly because they are more generous in the control environment. Within each role, women are more likely to donate than men.[46] The posting environment pushes both men and women away from donating, but more so for women (Appendix Figure 1.C.26). Only 20% of women in the control don't donate, a number which jumps to 38% in the posting environment. Men are also less likely to donate in the posting environment, although the jump is smaller from 40 to 49% and statistically insignificant. These patterns are similar if we focus on senders who post only, meaning the results extend to slacktivist behavior (Appendix Figure 1.C.27).

Neither the lower baseline nor the stronger treatment effect for women is obviously driven by related factors. Self-reported support for the cause, financial need, and beliefs of the impact of posts are similar by gender. Women are a little more likely to be activists in the sense of having participated in some offline activism in the last two years.

[46] This is consistent with the literature, where many papers find that women are more generous, including Eckel & Grossman (1998).

I do not find significant heterogeneous treatment effects by other dimensions. Past donation behavior is correlated with the likelihood of donating, but not the size of the treatment effect; the posting environment induced a parallel shift of crowd out for people who didn't donate in the past two years, donated sometimes, or donated often (Appendix Figure 1.C.28).

I would love to be able to report heterogeneity results for other groups, such as by age, race/ethnicity, and whether they are part of a traditionally disadvantaged group (e.g. members of the LGBTQ+ population, African Americans). Unfortunately, my sample does not have enough variation by age and lacks enough data on the aforementioned subpopulations. Instead, I leave these questions to future work.

1.4.4 Policy Implications

Now that we have established that slacktivism occurs, I explore whether I can mitigate it and raise donations to their previous level even in the presence of public posts of support. This extension is important because social media and other low-effort ways of publicly showing support ways will continue to exist.

I first tried providing information about whether posts hurt or improve donations, and found no change in slacktivism behavior. Recall that in the Information Treatment arms, subjects who believe posts have a zero or negative impact crowd out to the same degree as those who think the impact of the post is positive. This suggests that simply saying posts don't contribute won't be enough to abate slacktivism behavior. Instead, any policy intervention should aim to affect reputation gains from posting, motivated reasoning, and people's ability to justify not donating to themselves.

Since there remains an open question about mitigation, I run additional treatment sessions that include a nudge treatment crafted using insights from my main analysis. The nudge intervention targets the mechanisms through which slacktivism occurs. The main mechanisms are motivated beliefs and moral licensing, both of which have a justification and a reputation component.

The language of the intervention is simple, such that charitable organizations would be able to copy the encouragement if it works.

I designed a nudge treatment in the posting environment that combines all of these considerations. I presented all treated individuals (senders and receivers) with a message from the charitable organizations asking them to consider donating regardless of their interaction with posts. The message includes facts about reduced donations in the presence of posts, i.e. slacktivism. It also uses phrases like ‘is not enough’ to reduce their ability to pat themselves on the back for posting. Subjects are informed that everyone else will see the same message so we may be able to attack the reputation piece.

Specifically, senders in the nudge treatment see the following message.

> A message from the charities to everyone in the room:
>
> If you only post or rely on others posting, *it is not enough*. Both people who send and observe posts of support are less likely to donate than they otherwise would have been. Donations are crucial for us as non-profits to fight for racial justice. So, **even if you choose to post, please consider donating**. Any amount is helpful.

Receivers saw an almost identical message, except the second to last sentence reads “So, **regardless of posts you see from others, please consider donating**.” Messages are shown when they are informed of their role. This occurs after all instructions and comprehension questions but just before they make their decisions and/or observe others’ posts.[47]

My hypothesis is that this intervention will partially mitigate donation crowd out by reducing the following channels of behavior: people’s sense of posting ‘being enough’, people thinking posts are influencing others positively, and reputation gains from posting. Combined with the intervention being a genuine ask from the charities, I hope to see more people donating even in the social media-like environment.

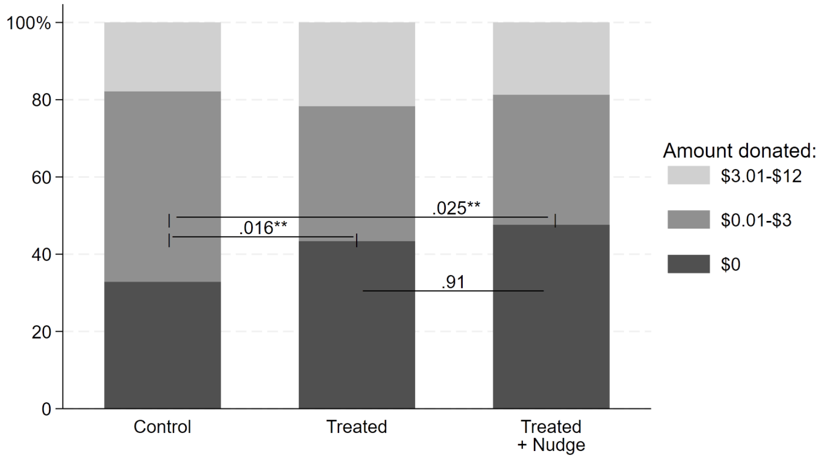
A benefit of this nudge treatment is that it can be translated to environments outside of the laboratory. It is standard for non-profit organizations to encourage donations, but telling people that the impact of posting is statistically zero in this lab experiment is not. Further, organizations

[47] This is at the same point in time that Information Treatment senders were told the impact of posts on receivers using data from previous sessions. This means I can compare across treatments easily.

want engagement. They would likely be hesitant to discourage posting support. Instead, this intervention encourages donating in addition to posting, and hypothesize that it will mitigate the justification and reputation components of slacktivist behavior.

Policy extension sessions were also run in the EconLab at UCSD, in the summer and fall of 2023. The sample of 107 individuals in the nudge treatment is largely balanced with the main sample of 322 individuals (Appendix Table 1.C.5).

Somewhat surprisingly, I find no mitigation from the nudge intervention. Individuals who read the nudge paragraph continue to donate at rates comparable to the main treated sample, meaning they are also significantly less likely to donate than control individuals (Figure 1.11). This pattern holds for both those who can post and those who see posts from others (Appendix Figure 1.C.29).



N = 73, 249, 107. P-values from a logistic regression of treatment on not donating, controlling for age, gender, past activism and donations, financial need, and desire to show support.

Figure 1.11: **The nudge intervention has no effect on donation behavior**

The nudge intervention is ineffective because it fails to impact any of the mechanisms. People continue to overestimate the impact of their post, crowd out regardless of whether they think the post is impactful, hold beliefs consistent with motivated reasoning, crowd out when they think posts won't influence themselves but will impact others, and posts give license to not donate as evidenced by senders giving shorter justifications of their donation choice.[48]

[48] See Appendix Figures 1.C.30, 1.C.31, 1.C.32, 1.C.33, and 1.C.34, respectively.

In terms of heterogeneity, non-activists continue to be the group of people who are most likely to engage in slacktivism (Appendix Figure 1.C.35). One thing to note is that in this sample, activist receivers no longer crowd in, suggesting that the result of posts encouraging activist receivers is less stable.

While I did not find an intervention that effectively mitigated slacktivism, there are other policy implications. First, we learned that slacktivism is real, pertinent, and persistent. Straightforward information and nudge interventions are unsuccessful in reversing donation crowd out. Second, we learn that in an environment with visible, low-effort actions, these public actions are the flexible margin. In the information and nudge interventions, subjects are slightly less likely to post support. Table 1.3 shows that 82% of people post when give the choice in the no information sessions, relative to about 71% in the information and nudge treatments. Together, this indicates that if the interventions have an effect on behavior, it is only on the low-effort visible action. The influence of the intervention does not trickle through to costly donation behavior.

Table 1.3: **Posting is the margin that people are willing to change in the information and nudge interventions**

Treatment	% who post
No Info	82.2
Information	70.5
told positive	76.2
told negative	63.9
Nudge	71.2

1.4.5 Discussion

This paper shows that slacktivism occurs and is more pervasive than previously thought. Activists and experts were right to be skeptical about simple forms of online support contributing to structural change, and wondering if it can even be counterproductive (Ho 2020).

Although I use a controlled environment, since the underlying mechanisms permeate other contexts I believe the results have a high degree of external validity. Focus groups find that people tend to overestimate the impact of posting and think they are doing more for the cause than they are when they show support on social media (Harlow & Guo 2014). Motivated beliefs explain why most people think they are better than average drivers and that their housing prices will always increase (Bénabou 2015). People avoid altruism by distorting beliefs of others and of norms (Di Tella et al 2015; Bicchieri et al 2023). Moral licensing is evident on and off social media when people have multiple possible actions. For instance, people who buy electric vehicles end up driving more miles than before (Dütschke et al 2018) and people tell themselves it's okay to not recycle this week if they usually do (Dodgson 2017). This behavior persists even when other individuals are directly affected. People who were given the opportunity to disagree with blatantly sexist statements were later more willing to select a man for a stereotypically male job (Monin & Miller 2001) and people who expressed support for president Obama were more likely to be racist or prejudiced afterwards (Effron et al 2012). Since these heuristics are commonplace,[49] it is extremely likely that donation crowd out extends beyond the laboratory environment.

One of the main counterarguments about slacktivism is that crowd out behavior might be unintentional. This line of thinking is evident in the model but empirical results are not purely unintentional. Donation behavior is not simply about beliefs about the impactfulness of posts or using posts to substitute for donations. Senders who donate a lot continue to donate a lot, but senders who would have donated a small amount crowd out their donation completely. Those who think posts do nothing or hurt the cause still crowd out their donations relative to the control. Even if slacktivism was unintentional, we would still care about it because engagement with social justice issues and donations to charitable organizations matter.

It is important to note that my setting is unique in that awareness levels are high. Experimental sessions are run in 2022 and 2023, years after the summer of 2020 surge in awareness for racial justice. Subjects are young undergraduate students who agree that racial justice is an important cause and have a high rate of social media use.[50] Laboratory posts are short-lived, lim-

[49] People may or may not be aware of the behavioral mechanisms at play in their actions. More research about how cognizant people are of motivated beliefs and moral licensing may shed light on this. Zimmermann (2020) shows that people have an asymmetry in the recall of negative feedback, further demonstrating the tenacity of motivated beliefs.

[50] Only a handful of subjects disagree with the statement 'Racial justice is an important cause'. Two-thirds of subjects have used social media to post in support of a cause they support, and 84% of 18-29 year olds use social media in general according to Auxier & Anderson (2021).

ited to the session time frame, and about racial justice broadly instead of raising awareness about specific charities. It is possible that posts would have had a larger impact on others' decisions if they were raising awareness about a cause or specific organization. In fact, the protests and wider societal movement of Black Lives Matter raised millions for racial justice charities (Goldmacher 2020). While this effect is not due to social media solely, 23% of users say social media led them to change their views on a matter, including online support for Black Lives Matter (Perrin 2020). Therefore my results are likely contingent on awareness levels already being high. Similarly, a non-profit organization running a social media campaign cares not only about the ultimate cause, but about awareness of their specific charity. I believe this factor of awareness raising capacity is important to consider when extrapolating this study's results to other situations. If low-effort visible forms of showing support can reasonably raise awareness, we could see less harmful aggregate effects of individual donation crowd out.

The posts of support used in the experiment are especially relevant for low-effort, generic forms of support. Posts that share deep, personal stories may have a positive impact on causes through empathy, and posts shared by celebrities may bring about large swathes of awareness. Relatedly, many say that social media may enable more voices to be shared, particularly voices of underrepresented groups (Anderson et al 2018). However, social media activism that is criticized as slacktivism are these simple posts of support like re-tweets or shares, posts with hashtags that lack information, and 'likes'. I chose to use these simple 'I support racial justice' posts in the experiment to allow for a strict test of slacktivism. Senders know they cannot personalize their posts of support and that they will only exist on receivers screens during the decision-making portion of the session. Yet, they crowd out their own donations and justify their behavior.

Related, posts in this study mimic social media use that is independent from donation decisions. One social media campaign that saw financial success was the ALS ice bucket challenge where donations were inherently tied to the social media campaign. Fazio et al (2023) find that more exposure to the challenge increased donations, although the bump lasted for a short duration. This suggests that another avenue for slacktivism mitigation involves intertwining social media campaigns with donations directly.

My crowd out results suggest that the self-image or consistency effects are relatively weak in the interaction of social media and offline actions. Self-image or consistency concerns would suggest that when people are given the option to post and choose to post, they would be encouraged to donate as well (Lee & Hsieh 2013). I don't see evidence of self-image or moral consistency, implying that these effects are outweighed by moral licensing, motivated beliefs, and other behavioral influences.

Not all forms of social media use are harmful, for example, Enikolopov et al (2020) and Zhuravskaya et al (2020) find that social media proliferation helped coordinate protests in Russia. However, this paper focuses on the low-effort, visible actions of support that enable slacktivism, a behavioral phenomenon that has economic consequences.

This paper provides a framework to think about slacktivism in other contexts. Key considerations include (i) the awareness raising capacity of low-effort, visible actions of support, (ii) the impact of the visible action of support, and (iii) whether the visible action of support is considered to be prosocial.

1.5 Conclusion

Social justice movements, charitable causes, unions, political parties, and other groups rely on actions of support. In the digital age, these actions extend beyond donations, protests, and volunteering, to include social media campaigns and blog posts. While digital forms of showing support are very visible and can have a wider reach, there is a salient concern that these low-effort ways of showing support may crowd out higher-effort actions of support that these groups depend on. Termed slacktivism, this phenomenon where someone posts to social media in support of a cause, but in doing so reduces their total contribution to that cause relative to if they had not posted, is the focus on this study.

Slacktivism is notoriously difficult to identify with certainty. People's actions of support are correlated, social media use is self-selected, data does not exist about all actions of support, and the impact of social media posts are immeasurable. This paper develops both theoretical and empirical tools, and provides the first causal evidence and quantification of slacktivism that includes the impact of posts.

I find that slacktivism does occur and is a large extensive margin phenomenon. Social media-like environments allow people to show support for racial justice in a low-effort way, which reduces their likelihood of supporting the cause with higher-cost actions. People with the option to post support are only half as likely to donate, and many of them post support. The public displays of support have no positive impact on the cause, leading not only to a causal reduction in donations, but also in total contributions. People who see posts from others are equally less likely to donate, meaning the social media-like environment led to a loss of 15% of the donor base.

The underlying behavioral mechanisms driving behavior are self-interested and widespread. The ability to send or observe posts of support gives people who would have donated a small amount an ‘out’ to justify not donating. In particular, I find evidence of motivated reasoning, justification effects, and moral licensing. These behavioral channels manifest in various daily behaviors, making this result unsurprising but expansive.

It is natural for charities and fundraising organizations to want to mitigate this harmful behavior, but unfortunately, simple information and nudge interventions are ineffective in reducing the incidence of slacktivism. The interventions of providing information about the impact of posting or encouraging donations in addition to visible, low-effort actions do not affect donation behavior, perhaps because mechanisms are self-interested. Instead, low-effort visible posts are the action that people are willing to change. More drastic measures of encouraging high-cost actions of support are needed to offset the harms of slacktivism. For example, future work could examine whether matching donations, which have been shown to encourage donating, reduce the extensive margin crowd out in the presence of public posts of support.

All in all, I find compelling evidence of slacktivism. This phenomenon cannot be ignored and has implications for individuals, non-profit organizations, social movements, public goods, and more. Social justice movements often exist at the intersection of digital and offline contributions, making them particularly susceptible to the harms of slacktivism. However, this behavior is not limited to social media, and can be extended to the interaction of any lower-effort visible action and higher-effort more private action, where these actions work towards a common goal.

Chapter 1 is currently being prepared for submission for publication of the material. The dissertation author, Amanda Bonheur, was the primary investigator and author of this material.

1.6 References

Abdi, Yomi (2020, Sep 5). A Tale of performative activism: How Black Lives Matter became just a trend. *Yale Daily News*. https://yaledailynews.com/sjp/2020/09/05/a-tale-of-performative-activism-how-black-lives-matter-became-just-a-trend/

Aksoy, Billur, Ian Chadd, and Boon Han Koh (2023). Sexual identity, gender, and anticipated discrimination in prosocial behavior. *European Economic Review*, Volume 154, 104427, ISSN 0014-2921. https://doi.org/10.1016/j.euroecorev.2023.104427

Anderson, Monica, Skye Toor, Kenneth Olmstead, Lee Rainie, and Aaron Smith (2018). Activism in the Social Media Age. *Pew Research Center*. https://www.pewresearch.org/internet/2018/07/11/activism-in-the-social-media-age/ and full report https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2018/07/PI_2018.07.11_social-activism_FINAL.pdf

Anderson, Monica, Michael Barthel, Andrew Perrin, and Emily A. Vogels (2020, June 5). #BlackLivesMatter surges on Twitter after George Floyd's death. *Pew Research Center*. https://www.pewresearch.org/fact-tank/2020/06/10/blacklivesmatter-surges-on-twitter-after-george-floyds-death/

Andreoni, James (1990). Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving. *The Economic Journal*, 100(401), 464–477. https://doi.org/10.2307/2234133

Andreoni, James (2006). Philanthropy. In S.-C. Kolm & J. M. Ythier (Eds.), Handbook of the Economics of Giving, Altruism and Reciprocity, Vol. 2, Ch. 18, pages 1201–1269. *Elsevier*.

Andreoni, James, and Abigail Payne (2013). Charitable Giving. *Handbook of Public Economics.* 5, 1-50. DOI: 10.1016/B978-0-444-53759-1.00001-7.

Andreoni, James, and Ragan Petrie (2004). Public goods experiments without confidentiality: a glimpse into fund-raising. *Journal of Public Economics*, Volume 88, Issues 7–8, Pages 1605-1623. ISSN 0047-2727. https://doi.org/10.1016/S0047-2727(03)00040-9

Andreoni, James, Justin M. Rao, and Hannah Trachtman (2017). Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving. *Journal of Political Economy*, Vol 125, Num 3. https://www.journals.uchicago.edu/doi/abs/10.1086/691703?journalCode=jpe

Ariely, Dan, Anat Bracha, and Stephan Meier (2009). Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially. *American Economic Review*, 99 (1): 544-55. DOI: 10.1257/aer.99.1.544

Auxier, Brooke, and Monica Anderson (2021, April 7). Social Media Use in 2021. *Pew Research Center*. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

Battigalli, Pierpaolo, and Martin Dufwenberg (2022). Belief-Dependent Motivations and Psychological Game Theory. *Journal of Economic Literature*, 60 (3): 833-82. DOI: 10.1257/jel.

20201378

Bénabou, Roland, and Jean Tirole (2006). Incentives and Prosocial Behavior. *American Economic Review*, 96 (5): 1652-1678. https://www.aeaweb.org/articles?id=10.1257/aer.96.5.1652

Bénabou, Roland (2015). The Economics of Motivated Beliefs. *Revue d'économie politique*, Vol. 125, p. 665-685. DOI 10.3917/redp.255.0665. https://www.princeton.edu/~rbenabou/papers/REDP_255_0665.pdf

Bicchieri, Cristina, Eugen Dimant, Michele Gelfand, and Silvia Sonderegger (2023). Social norms and behavior change: The interdisciplinary research frontier. *Journal of Economic Behavior & Organization*, Volume 205, Pages A4-A7, ISSN 0167-2681. https://doi.org/10.1016/j.jebo.2022.11.007

Blanken, Irene, Niels van de Ven, and Marcel Zeelenberg (2015). A Meta-Analytic Review of Moral Licensing. *Personality and Social Psychology Bulletin*, 41 (4), 540–558. https://doi.org/10.1177/0146167215572134

Boris, Elizabeth T., Erwin de Leon, Katie L. Roeger, and Milena Nikolova (2010). Human Service Nonprofits and Government Collaboration: Findings from the 2010 Nonprofit Government Contract and Grants. *The Urban Institute*. https://www.urban.org/sites/default/files/publication/29221/412228-Human-Service-Nonprofits-and-Government-Collaboration-Findings-from-the-National-Survey-of-Nonprofit-Government-Contracting-and-Grants.PDF

Bracha, Anat, and Lise Vesterlund (2017). Mixed signals: Charity reporting when donations signal generosity and income. *Games and Economic Behavior*, 104, issue C, p. 24-42. https://EconPapers.repec.org/RePEc:eee:gamebe:v:104:y:2017:i:c:p:24-42

Burdea, Valeria, and Jonathan Woon (2022). Online belief elicitation methods. *Journal of Economic Psychology*, Volume 90, 2022, 102496, ISSN 0167-4870. https://doi.org/10.1016/j.joep.2022.102496

Bursztyn, Leonardo, and Robert Jensen (2017). Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure. *Annual Review of Economics*, 9:1, 131-153. https://www.annualreviews.org/doi/abs/10.1146/annurev-economics-063016-103625

Chen, Daniel L., Martin Schonger, and Chris Wickens (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, Elsevier, vol. 9(C), pages 88-97.

Chon, Myoung-Gi, and Hyojung Park (2020). Social Media Activism in the Digital Age: Testing an Integrative Model of Activism on Contentious Issues. *Journalism & Mass Communication Quarterly*, 97(1), 72–97. https://doi.org/10.1177/1077699019835896

Danz, David, Lise Vesterlund, and Alistair J. Wilson. (2022). Belief Elicitation and Behavioral

Incentive Compatibility. *American Economic Review*, 112 (9): 2851-83.

DellaVigna, Stefano, John List, and Ulrike Malmendier (2012). Testing for altruism and social pressure in charitable giving. *Quarterly Journal of Economics*, 127(1):1–56.

Di Tella, Rafael, Ricardo Perez-Truglia, Andres Babino, and Mariano Sigman (2015). Conveniently Upset: Avoiding Altruism by Distorting Beliefs about Others' Altruism. *American Economic Review*, 105 (11): 3416-42. DOI: 10.1257/aer.20141409

Dütschke, Elisabeth, Manual Frondel, Joachim Schleich, and Colin Vance (2018). Moral Licensing—Another Source of Rebound? *Frontiers in Energy Research.* https://www.frontiersin.org/articles/10.3389/fenrg.2018.00038/full

Dodgson, Lindsay (2017, Nov 15). The way 'good' people explain away bad behaviour is called 'moral licensing' — here's what it means. *The Insider.* https://www.businessinsider.com/what-moral-licensing-means-2017-11

Eckel, Catherine C., and Philip J. Grossman (1996). Altruism in Anonymous Dictator Games. *Games and Economic Behavior*, Volume 16, Issue 2, 1996, Pages 181-191, ISSN 0899-8256. https://doi.org/10.1006/game.1996.0081

Eckel, Catherine C., and Philip J. Grossman (1998). Are Women Less Selfish Than Men?: Evidence from Dictator Experiments. *The Economic Journal*, 108(448), 726–735. http://www.jstor.org/stable/2565789

Effron, Daniel, Dale Miller, and Benoît Monin (2012). Inventing Racist Roads Not Taken: The Licensing Effect of Immoral Counterfactual Behaviors. *Journal of Personality and Social Psychology*, 103(6):916. DOI: 10.1037/a0030008. https://www.researchgate.net/publication/263924672_Inventing_Racist_Roads_Not_Taken_The_Licensing_Effect_of_Immoral_Counterfactual_Behaviors

Elias, Megan (2020). Armchair Activism: how social media changed the way we make change. *The Current.* https://thecurrentmsu.com/2021/01/16/armchair-activism-how-social-media-changed-the-way-we-make-change/

Enikolopov, Ruben, Alekey Makarin, and Maria Petrova (2020). Social Media and Protest Participation: Evidence from Russia. *Econometrica*, Vol 88, Issue 4, p. 1479-1514. https://doi.org/10.3982/ECTA14281

Erkal, Nisvan, Lata Gangadharan, and Boon Han Koh (2020). Replication: Belief elicitation with quadratic and binarized scoring rules. *Journal of Economic Psychology*, Volume 81, 102315, ISSN 0167-4870, https://doi.org/10.1016/j.joep.2020.102315

Evren, Özgür, and Stefania Minardi (2017). Warm-glow Giving and Freedom to be Selfish. *The Economic Journal*, Volume 127, Issue 603, Pages 1381-1409. https://doi.org/10.1111/ecoj.12351

Fazio, Andrea, Tommaso Reggiani, and Francesco Scervini (2023). Social Media Charity Campaigns and Pro-Social Behaviour. Evidence from the Ice Bucket Challenge. *IZA Discussion Paper*, No. 16046, http://dx.doi.org/10.2139/ssrn.4410222

Filiz-Ozbay, Emel, and Neslihan Uler (2019). Demand for giving to multiple charities: An experimental study. *Journal of the European Economic Association*, Volume 17, Issue 3, June 2019, Pages 725–753. https://doi.org/10.1093/jeea/jvy011

Fisher, Richard (2020, Sep 15). The subtle ways that 'clicktivism' shapes the world. *BBC Future*. https://www.bbc.com/future/article/20200915-the-subtle-ways-that-clicktivism-shapes-the-world

Fosgaard, Toke R., and Adriaan R. Soetevent (2022). I will donate later! A field experiment on cell phone donations to charity. *Journal of Economic Behavior & Organization*, Elsevier, vol. 202(C), pages 549-565.

Gee, Laura K, and Jonathan Meer (2019). The Altruism Budget: Measuring and Encouraging Charitable Giving. *National Bureau of Economic Research Working Paper Series*, Number 25938, doi 10.3386/w25938

Goldmacher, Shane (2020, June 16). Racial Justice Groups Flooded With Millions in Donations in Wake of Floyd Death. *New York Times*. https://www.nytimes.com/2020/06/14/us/politics/black-lives-matter-racism-donations.html

Grieder, Manuel, Jan Schmitz and Renate Schubert (2021). Asking to Give: Moral Licensing and Pro-Social Behavior in the Aggregate. Available at SSRN: http://dx.doi.org/10.2139/ssrn.3920355

Greijdanus, Hedy, Carlos A de Matos Fernandes, Felicity Turner-Zwinkels, Ali Honari, Carla A. Roos, Hannes Rosenbusch, and Tom Postmes (2020). The psychology of online activism and social movements: Relations between online and offline collective action. *Current Opinion in Psychology*, 35, 49-54. https://doi.org/10.1016/J.COPSYC.2020.03.003

Haberman, Clyde (2016, Nov 13). Philanthropy That Comes From a Click. *New York Times*. https://www.nytimes.com/2016/11/14/us/philanthropy-that-comes-from-a-click.html

Halupka, Max (2014). Clicktivism: A systematic heuristic. *Policy & Internet*, https://doi.org/10.1002/1944-2866.POI355

Harlow, Summer, and Lei Guo (2014). Will the Revolution be Tweeted or Facebooked? Using Digital Communication Tools in Immigrant Activism. *Journal of Computer-Mediated Communication*, Volume 19, Issue 3, 1 April 2014, Pages 463–478. https://doi.org/10.1111/jcc4.12062

Harwell, Haley, Daniel Meneses, Chris Moceri, Marc Rauckhorst, Adam Zindler, and Catherine Eckel (2015). Did the Ice Bucket Challenge Drain the Philanthropic Reservoir? *Economic*

Research Laboratory, Texas A&M University Working paper.

Ho, Shannon (2020, June 13). A social media 'blackout' enthralled Instagram. But did it do anything? *NBC News*. https://www.nbcnews.com/tech/social-media/social-media-blackout-enthralled-instagram-did-it-do-anything-n1230181

Hogben, Jasmine, and Fiona Cownie (2017). Exploring Slacktivism; Does The Social Observability of Online Charity Participation Act as a Mediator of Future Behavioural Intentions? *Journal of Promotional Communications*, 5 (2). http://eprints.bournemouth.ac.uk/24947/

Hossain, Tanjim, and Ryo Okui (2013). The Binarized Scoring Rule. *The Review of Economic Studies*, Volume 80, Issue 3, Pages 984–1001. https://doi.org/10.1093/restud/rdt006

Jones, Cat (2015). Slacktivism and the social benefits of social video. *First Monday*. https://firstmonday.org/article/view/5855/4458

Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler (1986). Fairness as a Constraint on Profit Seeking: Entitlements in the Market. *The American Economic Review*, 76(4), 728–741. http://www.jstor.org/stable/1806070

Karpf, David (2010). Online Political Mobilization: Looking Beyond Clicktivism. *Policy & Internet*. https://doi.org/10.2202/1944-2866.1098

Karpf, David (2012). The MoveOn Effect: Disruptive Innovation in the Interest Group Ecology of American Politics. In The MoveOn Effect: The Unexpected Transformation of American Political Advocacy. *Oxford Studies in Digital Politics, New York, online edn, Oxford Academic*. https://doi.org/10.1093/acprof:oso/9780199898367.003.0002

Kristofferson, Kirk, Katherine White, and John Peloza (2014). The Nature of Slacktivism: How the Social Observability of an Initial Act of Token Support Affects Subsequent Prosocial Action. *Journal of Consumer Research*, 40(6), 1149-1166. doi:10.1086/674137

Kouchaki, Maryam (2011). Vicarious Moral Licensing: The Influence of Others' Past Moral Actions on Moral Behavior. *Journal of personality and social psychology*, 101. 702-15. DOI: 10.1037/a0024552.

Lacetera, Nicola, Mario Macis, and Angelo Mele (2016). Viral Altruism? Charitable Giving and Social Contagion in Online Networks. *Sociological Science*. DOI 10.15195/v3.a11. https://www.sociologicalscience.com/download/vol-3/march/SocSci_v3_202to238.pdf

Lee, Yu-Hao, and Gary Hsieh (2013). Does slacktivism hurt activism? the effects of moral balancing and consistency in online activism. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13). Association for Computing Machinery*, New York, NY, USA, 811–820. https://doi.org/10.1145/2470654.2470770

Marwell, Gerald, and Ruth E. Ames (1981). Economists free ride, does anyone else?: Experiments on the provision of public goods, IV. *Journal of Public Economics*, Volume 15, Issue 3, 1981, Pages 295-310, ISSN 0047-2727. https://doi.org/10.1016/0047-2727(81)90013-X

Meier, Stephan (2006). A Survey of Economic Theories and Field Evidence on Pro-Social Behavior. *FRB of Boston Working Paper*, No. 06-6. http://dx.doi.org/10.2139/ssrn.917187

Meijers, Marijn H. C., Marret K. Noordewier, Peeter W. J. Verlegh, Simon Zebregs, and Edith G. Smit (2019). Taking Close Others' Environmental Behavior Into Account When Striking the Moral Balance? Evidence for Vicarious Licensing, Not for Vicarious Cleansing. *Environment and Behavior*, 51(9-10), 1027-1054. https://doi.org/10.1177/0013916518773148

Monin, Benoît, and Dale T. Miller (2001). Moral credentials and the expression of prejudice. *Journal of personality and social psychology*, 81 (1), 33–43. https://pubmed.ncbi.nlm.nih.gov/11474723/

Morozov, Evgeny (2009a, May 19). The Brave New World of Slacktivism. *Foreign Policy*. https://foreignpolicy.com/2009/05/19/the-brave-new-world-of-slacktivism/

Morozov, Evgeny (2009b, Sep 5). From Slacktivism to Activism. *Foreign Policy*. https://foreignpolicy.com/2009/09/05/from-slacktivism-to-activism/

NPT Charitable Giving Statistics (2022). *National Philanthropic Trust*, Accessed June 2022. https://www.nptrust.org/philanthropic-resources/charitable-giving-statistics/

Ogilvy & CSIC (2011). Slacktivists Doing More than Just Clicking. *Ogilvy Public Relations Worldwide and Georgetown University Center for Social Impact Communication*. https://csic.georgetown.edu/wp-content/uploads/2016/12/dce-slacktivists.pdf

Ottoni-Wilhelm, Mark, Lise Vesterlund, and Huan Xie (2017). Why Do People Give? Testing Pure and Impure Altruism. *American Economic Review*, 107 (11): 3617-33. DOI: 10.1257/aer.20141222

Perrin, Andrew (2020, October 15). 23% of users in U.S. say social media led them to change views on an issue; some cite Black Lives Matter. *Pew Research Center*. https://www.pewresearch.org/short-reads/2020/10/15/23-of-users-in-us-say-social-media-led-them-to-change-views-on-issue-some-cite-black-lives-matter/

Pew Research Center (2021). Social Media Factsheet. https://www.pewresearch.org/internet/fact-sheet/social-media/

Piper, Greg, and Schnepf Sylke (2007). Gender Differences in Charitable Giving. *Institute for the Study of Labor (IZA), IZA Discussion Papers*. https://www.researchgate.net/publication/5137646_Gender_Differences_in_Charitable_Giving

Philanthropy Roundtable (n.d.). Statistics on U.S. Generosity. Accessed May 2024, data through 2016. https://www.philanthropyroundtable.org/resource/statistics-on-u-s-generosity/

Robertson, Ashley Fish (2020). The Age Of Slacktivism: BLM Advocacy Beyond Keyboard

Crusading. *The Concordian*. https://theconcordian.com/2020/06/the-age-of-slacktivism-blm-advocacy-beyond-keyboard-crusading/

Rotella, Amanda, Jisoo Jung, Christopher Chinn, and Pat Barclay (2023, March 28). Observation moderates the moral licensing effect: A meta-analytic test of interpersonal and intrapsychic mechanisms. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/tmhe9

Samuelson, Paul A. (1954). The Pure Theory of Public Expenditure. *The Review of Economics and Statistics*, 36(4), 387–389. https://doi.org/10.2307/1925895

Scharf, Kimberley, Sarah Smith, and Mark Ottoni-Wilhelm (2022). Lift and Shift: The Effect of Fundraising Interventions in Charity Space and Time. *American Economic Journal: Economic Policy*, 14 (3): 296-321.

Schotter, Andrew, and Isabel Trevino (2014). Belief Elicitation in the Lab. *Annual Review of Economics*, Vol. 6:103-128, DOI: 10.1146/annurev-economics-080213-040927

Schumann, Sandy, and Olivier Klein (2015). Substitute or stepping stone? Assessing the impact of low-threshold online collective actions on offline participation. *European Journal of Social Psychology*, 45(3), 308–322. https://doi.org/10.1002/ejsp.2084

Simbrunner, Philipp, and Bodo B. Schlegelmilch (2017). Moral licensing: a culture-moderated meta-analysis. *Management Review Quarterly*, 67, 201–225. https://doi.org/10.1007/s11301-017-0128-0

Skidmore, Tessa, and Charles Sellen (2021). Giving while female: Women are more likely to donate to charities than men of equal means. *The Conversation*. https://theconversation.com/giving-while-female-women-are-more-likely-to-donate-to-charities-than-men-of-equal-means-141518

Thunström, Linda (2020). Thoughts and prayers – Do they crowd out charity donations? *Journal of Risk and Uncertainty*, Springer, vol. 60(1), pages 1-28.

Wilkins, Denise, Andrew Livingstone, and Mark Levine (2019). All click, no action? Online action, efficacy perceptions, and prior experience combine to affect future collective action. *Computers in Human Behavior*, 91. DOI: 10.1016/j.chb.2018.09.007.

Wilson, Alistair J., and Emanuel Vespa (2016). Paired-Uniform Scoring: Implementing a Binarized Scoring Rule with Non-mathematical Language. *Working Paper*. Available at https://sites.pitt.edu/~alistair/papers/PSR_November.pdf

WPI Women's Philanthropy Institute (2022). Women Give 2022 Report. https://philanthropy.iupui.edu/institutes/womens-philanthropy-institute/research/Women-Give-2022-Report-final.pdf

Zhang, Zhe, and Siyu Peng (2022). Licensing Effect in Sustainable Charitable Behaviors. *Sustainability*, 14(24), 16431. https://doi.org/10.3390/su142416431

Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov (2020). Political Effects of the Internet and Social Media. *Annual Review of Economics*. http://dx.doi.org/10.2139/ssrn.3439957

Zimmermann, Florian (2020). The Dynamics of Motivated Beliefs. *American Economic Review*, 110 (2): 337-61. DOI: 10.1257/aer.20180728

Chapter 1 Appendices

1.A Theoretical Framework Extensions

This appendix provides proofs and derivations used in the main text. Further, this appendix discusses ways one could expand the model to include more scenarios, including if the post had impact beyond encouraging others to donate. In this case, the direct impact of a post could be represented as g(P).

Let Y represent the total amount contributed to the cause from everyone's actions. In relation to social justice issues, you can think about this as contributions towards social impact. Total contributions or social impact are continuous, with higher values translating to more social justice. This outcome of interest is produced through donations and posts. Define D as the sum of all individual donations, $D = \sum_{j=1}^{n} d_j$, and P as the sum of all individual posts, $P = \sum_{j=1}^{n} p_j$. Then the total social impact of everyone's actions in dollar units is $Y = D + g(P)$ where $g(\cdot)$ maps posts to their monetary equivalent. The g function is generic with $g(0) = 0$ since no posts means there is no social impact, g'≥ 0 since more posts are weakly beneficial (e.g. raising awareness). How posts accumulate to social impact could be linear, concave, or convex, so g"$\gtreqless 0$. Intuitively, social impact comes about through monetary contributions towards causes, legislation, and societal change. Social media posts can influence societal change through ideas, norms, and social pressure. The more people post to social media, the more ideas spread. If so many people post that a cause or idea goes viral, this is an example of a convex $g(\cdot)$ function where the aggregation of everyone's posts creates a large social impact. Alternatively, if 1000 people post in support of a cause, the 1001st person's post is unlikely to amount to meaningful change. This scenario would be represented by a concave $g(\cdot)$ function.[51]

Following the above logic, person i's individual contribution to social change, denoted as

[51] In the experiment the posts have no direct impact in dollar terms, meaning the g function is zero. Letting $g(\cdot)$ be linear is a reasonable simplification due to the small number of people in each session and short duration of these laboratory-limited posts. This simplification helps identification.

y_i, is produced by their donation, d_i, and their visible post of support, p_i. This value is one outcome of interest. People get utility from contributing through altruism and warm glow. I derive the production function of an individuals' social impact as $y_i = d_i + f(p_i)$. The function $f(p_i)$ maps i's visible action to its social impact in dollars (to match the units of d_i) and is made up of a direct and indirect component. Person i's post directly affects social impact through the $g(P)$ function and indirectly by influencing the behavior of others. Mathematically, $f(p_i) = p_i\left[g(P_{-i} + p_i) - g(P_{-i})\right] + p_i\left[(D_{-i}|p_i = 1) - (D_{-i}|p_i = 0)\right]$. Notice that if someone doesn't post, $f(0) = 0$, so their individual contribution is only through their donation. When someone posts, the impact of their post is the marginal change in the impact of total posts and the marginal change in others' donations due to their post, $f(1) = \left[g(P_{-i} + p_i) - g(P_{-i})\right] + \left[(D_{-i}|p_i = 1) - (D_{-i}|p_i = 0)\right]$. This can be represented by $\triangle g(P)|p_i + \triangle D_{-i}|p_i$ for short.

Proof 1 ***Calculation of y_i and $f(p_i)$ from Y***

We can rewrite total contributions from all individuals into components from individual i and not from individual i.

$$\begin{aligned} Y &= D + g(P) \\ &= d_i + D_{-i} + g(P_{-i} + p_i) \\ &= d_i + D_{\{-i,i=\text{sender}\}} + (D_{\{-i,i=\text{receiver}\}}|(P_{-i}, p_i)) + g(P_{-i} + p_i) \end{aligned}$$

Individual i can contribute to Y through d_i and p_i. Receivers can see posts from senders, which may influence their behavior.

Donations are always private, meaning that d_i cannot influence others' actions. Individual i's donation contributes to social change directly.

On the other hand, individual i's posting choice is public and may influence others' (receivers') donations indirectly, while also contributing to social change directly through awareness, captured by the $g()$ function.

If i does not post, $p_i = 0$ and

$$Y = d_i + D_{\{-i,i=\text{sender}\}} + (D_{\{-i,i=\text{receiver}\}}|P_{-i}) + g(P_{-i} + 0) \tag{4}$$

meaning that the impact of no post is zero.

If i does post, $p_i = 1$ and

$$Y = d_i + D_{\{-i,i=\text{sender}\}} + (D_{\{-i,i=\text{receiver}\}}|P_{-i} + 1) + g(P_{-i} + 1) \tag{5}$$

Therefore the impact of i's post on social change is Equation 5 minus 4.

$$\begin{aligned} f(1) - f(0) &= d_i + D_{\{-i,i=\text{sender}\}} + (D_{\{-i,i=\text{receiver}\}}|P_{-i} + 1) + g(P_{-i} + 1) \\ &\quad - d_i + D_{\{-i,i=\text{sender}\}} + (D_{\{-i,i=\text{receiver}\}}|P_{-i}) + g(P_{-i} + 0) \\ f(1) &= \triangle_{\{p_i=1\}} D_{\{-i,i=\text{receiver}\}} + \triangle_{\{p_i=1\}} g(P_{-i}) \end{aligned} \tag{6}$$

where $\triangle_{\{p_i=1\}}$ is defined as the change resulting from p_i going from 0 to 1.

□

Proof 2 ***Derivation of Equation 1.1***

Consider the case with no visible actions. In this world $y_i = d_i$ and $Y = D = D_{-i} + d_i$. From i's perspective, $D = \hat{D}_{-i} + d_i$. Recall that $\hat{D_{-i}} = E_i(D_{-i})$.

$$\begin{aligned} max_{h_i}\, \alpha \ln(d_i) + \omega \ln(\hat{D}_{-i} + d_i) + \gamma \ln(x_i) \\ s.t.\; c^d d_i + x_i = I \end{aligned}$$

The first order condition with respect to d_i of the above problem is

$$\frac{\alpha}{d_i} + \frac{\omega}{\hat{D}_{-i} + d_i} - \frac{\gamma c^d}{I - c^d d_i} = 0$$

Simplifying gives

$$\begin{aligned} d_i^*(\hat{D}_{-i}) = &\frac{1}{2}\left(\frac{\alpha+\omega}{\alpha+\omega+\gamma}\right)\frac{I}{c^d} - \frac{1}{2}\left(\frac{\alpha+\gamma}{\alpha+\omega+\gamma}\right)\hat{D}_{-i} \\ &\pm \frac{1}{2}\left(\frac{1}{\alpha+\omega+\gamma}\right)\frac{1}{c^d}\left\{\left(-c^d\hat{D}_{-i}(\alpha+\gamma) + I(\alpha+\omega)\right)^2 + 4c^d\hat{D}_{-i}I\alpha\left(\alpha+\omega+\gamma\right)\right\}^{\frac{1}{2}} \end{aligned}$$

Notice that the above equation is almost Equation 1.1, except includes a ± from the quadratic formula. Let d^+ (d_i^-) denote the equation for d_i^* using the plus sign (minus sign) from the plus-minus sign. Donations must be greater than or equal to zero, since one cannot take out of the donation pot for themselves. I will show that $d^- \leq 0$. Using the fact that $d_i^* \geq 0$, we can conclude that $d_i^* = d_i^+$ which is Equation 1.1.

$$\begin{aligned} &d_i^- \leq 0 \iff \\ &d_i^- = \frac{1}{2}\left(\frac{\alpha+\omega}{\alpha+\omega+\gamma}\right)\frac{I}{c^d} - \frac{1}{2}\left(\frac{\alpha+\gamma}{\alpha+\omega+\gamma}\right)\hat{D}_{-i} \\ &\quad - \frac{1}{2}\left(\frac{1}{\alpha+\omega+\gamma}\right)\frac{1}{c^d}\left\{\left(-c^d\hat{D}_{-i}(\alpha+\gamma) + I(\alpha+\omega)\right)^2 + 4c^d\hat{D}_{-i}I\alpha\left(\alpha+\omega+\gamma\right)\right\}^{\frac{1}{2}} \\ &< \frac{1}{2}\left(\frac{\alpha+\omega}{\alpha+\omega+\gamma}\right)\frac{I}{c^d} - \frac{1}{2}\left(\frac{\alpha+\gamma}{\alpha+\omega+\gamma}\right)\hat{D}_{-i} \\ &\quad - \frac{1}{2}\left(\frac{1}{\alpha+\omega+\gamma}\right)\frac{1}{c^d}\left\{\left(-c^d\hat{D}_{-i}(\alpha+\gamma) + I(\alpha+\omega)\right)^2\right\}^{\frac{1}{2}} \\ &= \frac{1}{2}\left(\frac{\alpha+\omega}{\alpha+\omega+\gamma}\right)\frac{I}{c^d} - \frac{1}{2}\left(\frac{\alpha+\gamma}{\alpha+\omega+\gamma}\right)\hat{D}_{-i} - \frac{1}{2}\left(\frac{\alpha+\omega}{\alpha+\omega+\gamma}\right)\frac{I}{c^d} + \frac{1}{2}\left(\frac{\alpha+\gamma}{\alpha+\omega+\gamma}\right)\hat{D}_{-i} = 0 \end{aligned}$$

because $c^d > 0$, $I > 0$, $\alpha, \omega, \gamma \geq 0$. Therefore $d_i^- \leq 0$, and $d_i^* = d_i^+$. The equation for d_i^+ is that of Equation 1.1. □

Proof 3 ***Proof that $d_i^*(\hat{D}_{-i})$ is decreasing in $\hat{D}_{-i}$ as long as $\alpha > 0$ and $\alpha + \omega + \gamma > 0$.***

To prove that $d_i^*(\hat{D}_{-i})$ is decreasing in $\hat{D}_{-i}$ I will show that $\frac{\partial}{\partial \hat{D}_{-i}} d_i^* < 0$ as long as $\alpha(\alpha + \omega + \gamma) > 0$.

$$
\begin{aligned}
&\frac{\partial}{\partial \hat{D}_{-i}} d_i^* = \\
&-\frac{1}{2}\frac{\alpha+\gamma}{\alpha+\omega+\gamma} + \frac{1}{2}\frac{1}{c^d(\alpha+\omega+\gamma)}\frac{1}{2}\left((-c^d\hat{D}_{-i}(\alpha+\gamma)+I(\alpha+\omega))^2 + 4c^d\hat{D}_{-i}I\alpha(\alpha+\omega+\gamma)\right)^{-\frac{1}{2}} \\
&\quad * \left(2\left(-c^d\hat{D}_{-i}(\alpha+\gamma)+I(\alpha+\omega)\right)(-c^d)(\alpha+\gamma) + 4c^dI\alpha(\alpha+\omega+\gamma)\right) \\
&= -\frac{1}{2}\frac{\alpha+\gamma}{\alpha+\omega+\gamma} + \frac{1}{2}\frac{1}{c^d(\alpha+\omega+\gamma)}\frac{c^{d^2}\hat{D}_{-i}^2(\alpha+\gamma)^2 + c^dI(\alpha+\gamma)(\alpha+\omega) - 2c^dI\gamma\omega}{\sqrt{\left(-c^d\hat{D}_{-i}(\alpha+\gamma)+I(\alpha+\omega)\right)^2 + 4c^d\hat{D}_{-i}I\alpha(\alpha+\omega+\gamma)}} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad < 0
\end{aligned}
$$

$$
\begin{aligned}
\iff\ & \frac{1}{c^d}\frac{c^{d^2}\hat{D}_{-i}^2(\alpha+\gamma)^2 + c^dI(\alpha+\gamma)(\alpha+\omega) - 2c^dI\gamma\omega}{\sqrt{\left(-c^d\hat{D}_{-i}(\alpha+\gamma)+I(\alpha+\omega)\right)^2 + 4c^d\hat{D}_{-i}I\alpha(\alpha+\omega+\gamma)}} < \alpha+\gamma \\
\iff\ & \frac{c^d\hat{D}_{-i}^2(\alpha+\gamma)^2 + I(\alpha+\gamma)(\alpha+\omega) - 2I\gamma\omega}{\sqrt{\left(-c^d\hat{D}_{-i}(\alpha+\gamma)+I(\alpha+\omega)\right)^2 + 4c^d\hat{D}_{-i}I\alpha(\alpha+\omega+\gamma)}} < \alpha+\gamma \\
\iff\ & \left(c^d\hat{D}_{-i}(\alpha+\gamma) + I(\alpha+\omega) - 2I\frac{\gamma\omega}{\alpha+\gamma}\right)^2 < \\
& \left(-c^d\hat{D}_{-i}(\alpha+\gamma)+I(\alpha+\omega)\right)^2 + 4c^d\hat{D}_{-i}I\alpha(\alpha+\omega+\gamma)
\end{aligned}
$$

The LHS can be rewritten as

$$
\begin{aligned}
LHS &= \left(c^d\hat{D}_{-i}(\alpha+\gamma) + I(\alpha+\omega) - 2I\frac{\gamma\omega}{\alpha+\gamma}\right)^2 \\
&= c^{d^2}\hat{D}_{-i}^2(\alpha+\gamma)^2 + I^2(\alpha+\omega)^2 + 4I^2\frac{\gamma\omega}{\alpha+\gamma}\left(\frac{\gamma\omega}{\alpha+\gamma} - (\alpha+\omega)\right) \\
&+ c^d\hat{D}_{-i}I\left(2\alpha^2 + 2\alpha\gamma + 2\alpha\omega - 2\gamma\omega\right)
\end{aligned} \tag{7}
$$

The RHS can be rewritten as

$$
\begin{aligned}
RHS &= \Big(-c^d\hat{D}_{-i}(\alpha+\gamma) + I(\alpha+\omega)\Big)^2 + 4c^d\hat{D}_{-i}I\alpha(\alpha+\omega+\gamma) \\
&= c^{d^2}\hat{D}^2_{-i}(\alpha+\gamma)^2 + I^2(\alpha+\omega)^2 + c^d\hat{D}_{-i}I(2\alpha^2+2\alpha\omega+2\alpha\gamma-2\gamma\omega) \qquad (8)
\end{aligned}
$$

So, the LHS (Eqn 7) $<$ RHS (Eqn 8) when

$$
\begin{aligned}
4I^2\frac{\gamma\omega}{\alpha+\gamma}\left(\frac{\gamma\omega}{\alpha+\gamma} - (\alpha+\omega)\right) &< 0 \\
\frac{\gamma\omega}{\alpha+\gamma}\left(\frac{\gamma\omega}{\alpha+\gamma} - (\alpha+\omega)\right) &< 0 \\
\iff \left(\frac{\gamma\omega}{\alpha+\gamma}\right)^2 &< \frac{\gamma\omega}{\alpha+\gamma}(\alpha+\omega) \\
\left(\frac{\gamma\omega}{\alpha+\gamma}\right) &< (\alpha+\omega) \\
\gamma\omega &< \alpha^2+\alpha\omega+\alpha\gamma+\omega\gamma \\
0 &< \alpha(\alpha+\omega+\gamma) \qquad (9)
\end{aligned}
$$

Therefore, $\frac{\partial}{\partial\hat{D}_{-i}}d_i^* < 0$ as long as $\alpha > 0$ and $(\alpha+\omega+\gamma) > 0$. □

Proof 4 ***Derivation of Equation 1.2***

To prove the equation for the optimal donation decision, with the option to post support, $d_i^*(p_i.\hat{D}_{-i})$, consider the full utility maximization problem, with $c_p = 0$:

$$
\begin{aligned}
max_{d_i,p_i} \quad & \alpha\ln(y_i) + \omega\ln(D) + \gamma\ln(x_i) + \delta p_i \\
& \text{s.t. budget constraint } c^d d_i + x_i \leq I \\
& \text{production fn } y_i = c^d d_i + \kappa^* p_i \\
& \text{where } p_i = \{0,1\};\ D = \sum_{i=1}^{n} d_i = \hat{D}_{-i} + d_i
\end{aligned}
$$

Taking the first order condition gives

$$\frac{\alpha}{d_i + \kappa p_i} + \frac{\omega}{d_i + \hat{D}_{-i}} - \frac{\gamma c^d}{I - c^d d_i} = 0$$

Algebraic calculations to solve for d_i^* as a function of κ, p_i, I, and $\hat{D}_{-i}$ give

$$\begin{aligned}
d_i^* = &\frac{1}{2}\frac{I}{c^d}\frac{\alpha+\omega}{\alpha+\omega+\gamma} - \frac{1}{2}\frac{(\alpha+\omega)p_i}{c^d(\alpha+\omega+\gamma)} - \frac{1}{2}\frac{\gamma+\omega}{\alpha+\omega+\gamma}\kappa p_i - \frac{1}{2}\frac{\alpha+\gamma}{\alpha+\omega+\gamma}\hat{D}_{-i} \\
&\pm \frac{1}{2}\frac{1}{c^d(\alpha+\omega+\gamma)}[\left(\kappa p_i c^d(\gamma+\omega) + \hat{D}_{-i}c^d(\gamma+\alpha) - (\alpha+\omega)I\right)^2 \\
&- 4(\alpha+\omega+\gamma)c^d\left(\kappa p_i \hat{D}_{-i}\gamma c^d - \alpha I \hat{D}_{-i} - \omega I \kappa p_i\right)]^{\frac{1}{2}}
\end{aligned}$$

Further simplification and letting $c^d = 1$ gives Equation 1.2:

$$\begin{aligned}
d_i^*(p_i, \hat{D}_{-i}) = &\frac{1}{2}\left(\left(\frac{\alpha+\omega}{\alpha+\omega+\gamma}\right)I - \left(\frac{\alpha+\gamma}{\alpha+\omega+\gamma}\right)\hat{D}_{-i} - \left(\frac{\omega+\gamma}{\alpha+\omega+\gamma}\right)\kappa p_i\right) \\
&+ \frac{1}{2}\left(\frac{1}{\alpha+\omega+\gamma}\right) * \left\{\left[-(\alpha+\gamma)\hat{D}_{-i} + I(\alpha+\omega) - (\omega+\gamma)\kappa p_i\right]^2 \right. \\
&\left. + 4(\alpha+\omega+\gamma)\left(\hat{D}_{-i}I\alpha + I\kappa p_i\omega - \hat{D}_{-i}\kappa p_i\gamma\right)\right\}^{\frac{1}{2}}
\end{aligned} \tag{10}$$

□

Proof 5 ***Proof that $d_i^*(p_i, \hat{Y}_{-i})$ is decreasing in $\hat{Y}_{-i}$ as long as $\alpha > 0$ and $\alpha + \omega + \gamma > 0$.***

This proof is similar to Proof 3, and so is left to the reader.

□

1.B Instructions and Decision Screens

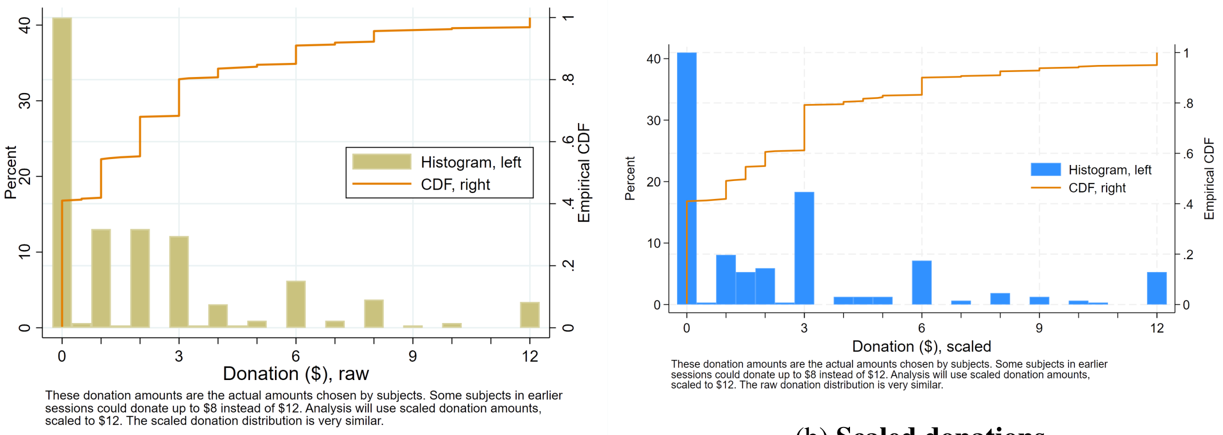
Please email the author for a copy of Appendix 1.B.

1.C Additional Figures and Tables

Table 1.C.1: **Participants match the UCSD undergraduate economics major population and treatment groups are balanced on observables (Main sample)**

	All	**Control**	**Sender**	**Receiver**
Demographics				
Age	20.4	20.3	20.4	20.4
Gender				
Male	47.5	50.7	48.8	44.4
Female	49.7	46.6	48.0	52.4
Non-binary / Other	1.6	0.0	2.4	1.6
Prefer not to say	1.6	2.7	0.8	1.8
Race				
Asian or Asian American	70.2	74.0	70.0	68.3
Black or African American	0.3	0.0	0.0	0.8
Hispanic or Latino	7.8	9.6	8.9	5.6
Native Amer. or Alaskan	0.6	0.0	0.0	1.6
White or Caucasian	11.5	6.9	11.4	14.3
Bi/Multiracial	4.4	4.1	4.1	4.8
Other	0.6	0.0	0.8	0.8
Prefer not to say	4.7	5.5	4.9	4.0
International student	47.4	55.6	47.2	42.9
Self-reports				
Racial justice is important (ω)	6.3	6.3	6.4	6.1
I want to contribute (α)	5.4	5.3	5.6	5.3
I need the money (γ)	5.3	5.4	5.4	5.1
Want others see I support (δ)	5.1	5.1	5.2	5.0
Past behavior				
Donate	2.3	2.1	2.4	2.2
Offline action	1.9	1.8	2.0	1.8
Social media	2.3	2.4	2.4	2.2
Online action	2.1	2.2	2.3	1.9
N	**322**	**73**	**126**	**123**

Note: Demographic variables are percentages, except age which is shown in years. Self-reports are the mean answer to 7-point Likert scale questions, see Appendix 1.B for details.



These donation amounts are the actual amounts chosen by subjects. Some subjects in earlier sessions could donate up to $8 instead of $12. Analysis will use scaled donation amounts, scaled to $12. The scaled donation distribution is very similar.

These donation amounts are the actual amounts chosen by subjects. Some subjects in earlier sessions could donate up to $8 instead of $12. Analysis will use scaled donation amounts, scaled to $12. The raw donation distribution is very similar.

(a) **Raw donations**

(b) **Scaled donations**

Figure 1.C.1: **Most donate at whole dollar increments and 80% donate $3 or less**

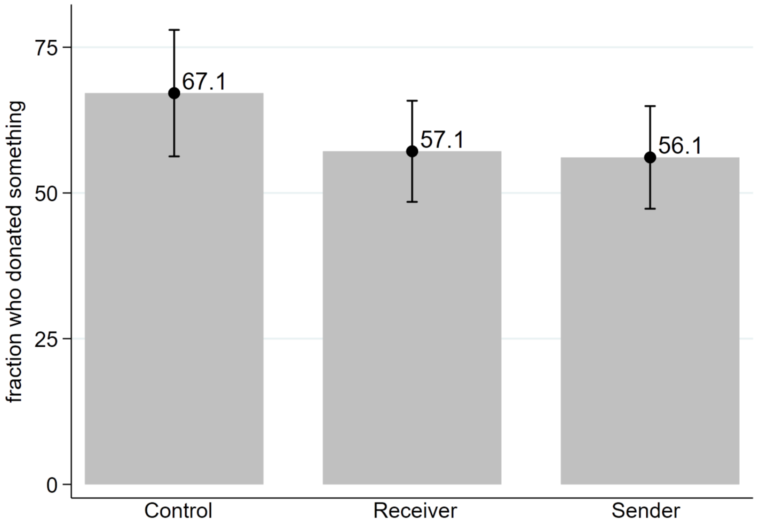


Figure 1.C.2: **Both senders and receivers are less likely to donate than control individuals**

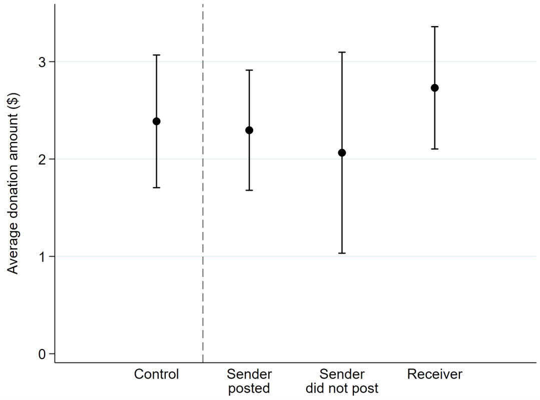


Figure 1.C.3: **Treated individuals, including senders who post, donate the same amount on average**

Table 1.C.2: **Treated individuals, including senders who post, donate the same amount on average**

	Donate Amount ($)			
	(1)	(2)	(3)	(4)
Sender, posted	-0.0913	-0.332	-0.416	-0.450
	(0.508)	(0.463)	(0.464)	(0.464)
Sender, did not post	-0.322	-0.723	-0.764	-0.890
	(0.695)	(0.636)	(0.634)	(0.636)
Receiver	0.344	0.149	0.124	0.0600
	(0.477)	(0.436)	(0.435)	(0.437)
N	322	314	314	313
Adjusted R^2	-0.00433	0.141	0.147	0.152
Parameters		Y	Y	Y
Past Activism			Y	Y
Demographics				Y

Standard errors in parentheses
Relative to Control individuals. Control variable parameters are 'I want to contribute' (alpha) and 'I need the money' (gamma).
Past activism behavior are answers to 7-point Likert scale questions of how often donated in the last two years and how often participated in offline activism. Demographics are age and gender identity.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

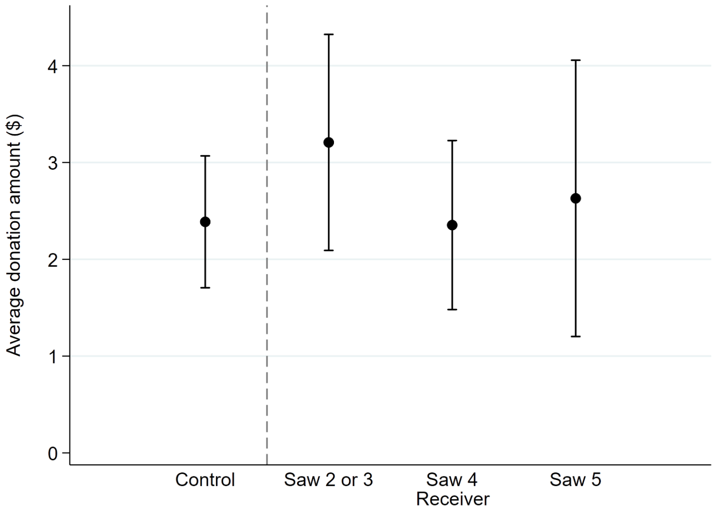


Figure 1.C.4: **Receivers do not change their donation amount on average depending on the number of posts observed**

Table 1.C.3: **Receivers don't change their likelihood of not donating depending on the number of posts they observe**

	Donate Zero (Odds Ratio)			
	(1)	(2)	(3)	(4)
Number of posts observed	1.082	1.104	1.145	1.141
	(0.223)	(0.244)	(0.266)	(0.268)
N	126	121	121	121
Pseudo R^2	0.000846	0.0621	0.111	0.112
Parameters		Y	Y	Y
Past Activism			Y	Y
Demographics				Y

Exponentiated coefficients; Standard errors in parentheses. Relative to other Receivers. Control variable parameters are 'I want to contribute' (alpha) and 'I need the money' (gamma). Past activism behavior are answers to 7-point Likert scale questions of how often donated in the last two years and how often participated in offline activism. Demographics are age and gender identity.
$^* p < 0.1$, $^{**} p < 0.05$, $^{***} p < 0.01$

Table 1.C.4: **Senders overestimate the impact of a post, whose true impact is $0, but less so when given information about the impact of a post**

	No Info	**Positive**	**Negative**
Believed avg impact	$1.04	$0.41	$0.49
Think impact positive	84%	86%	64%
N	45	42	36

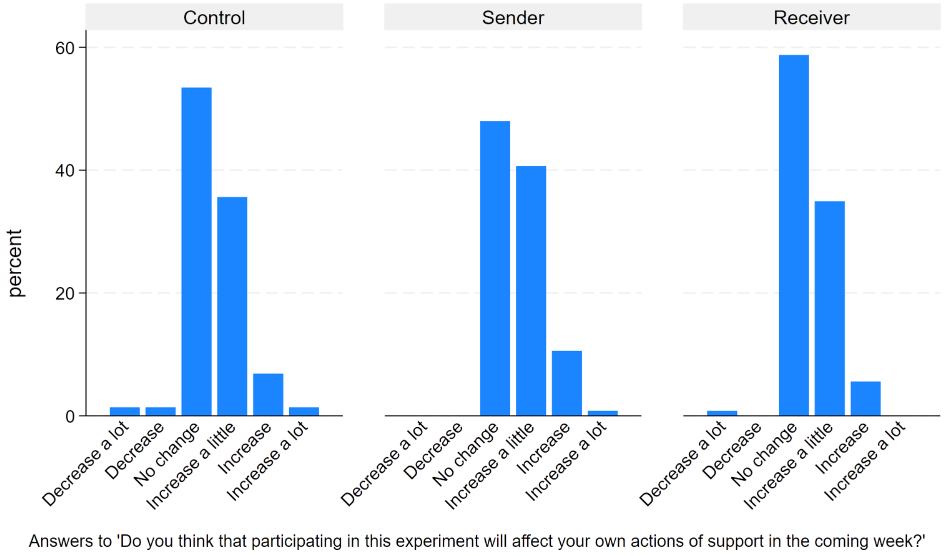


Figure 1.C.5: **Treatment doesn't change how people think they will behave next week**

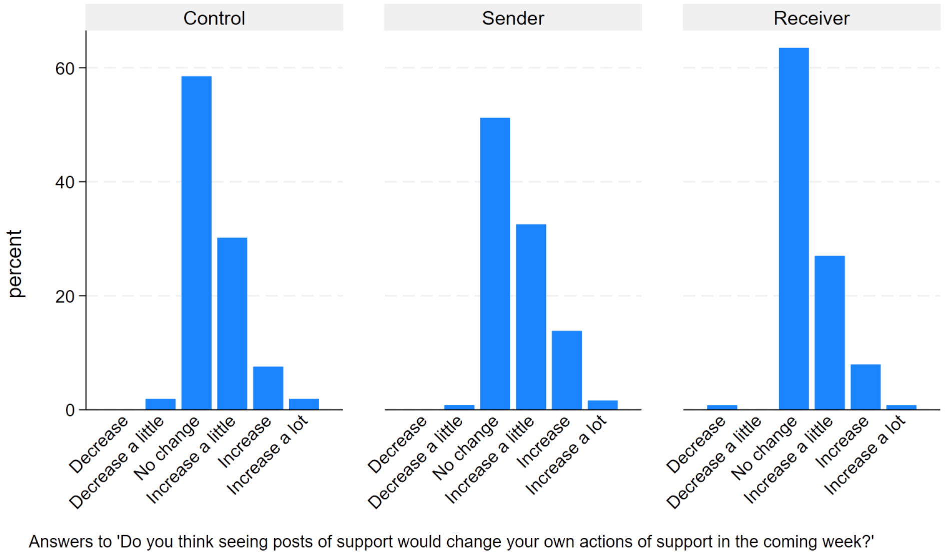


Figure 1.C.6: **Seeing posts (real or hypothetical) doesn't change how people think they will behave next week**

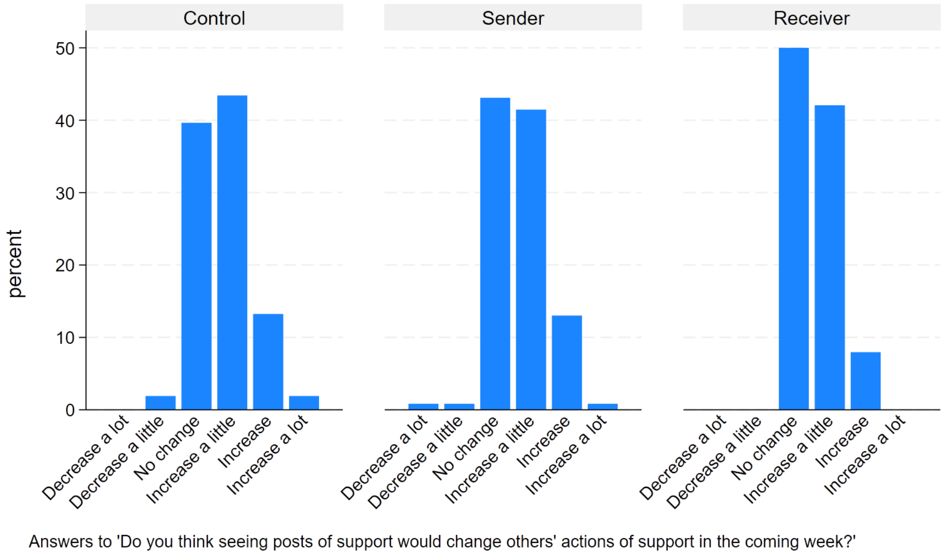
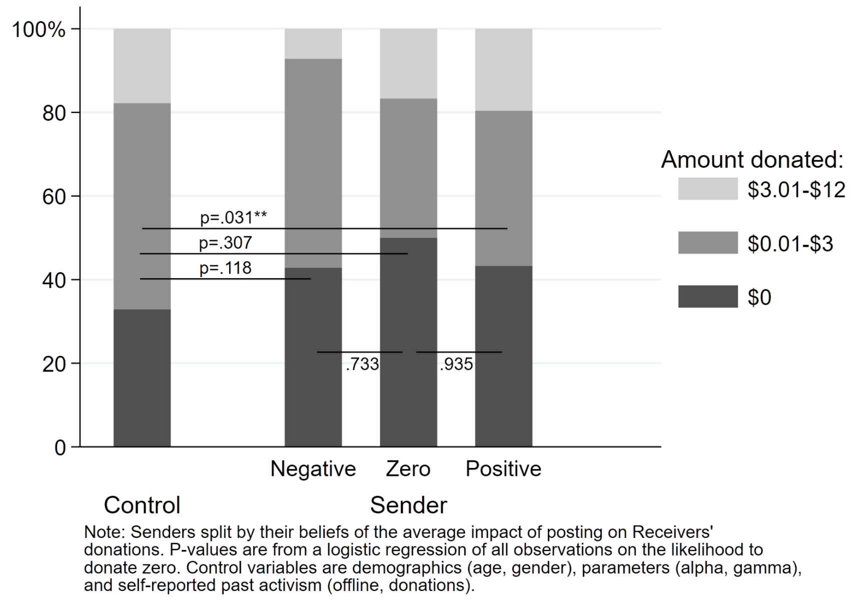
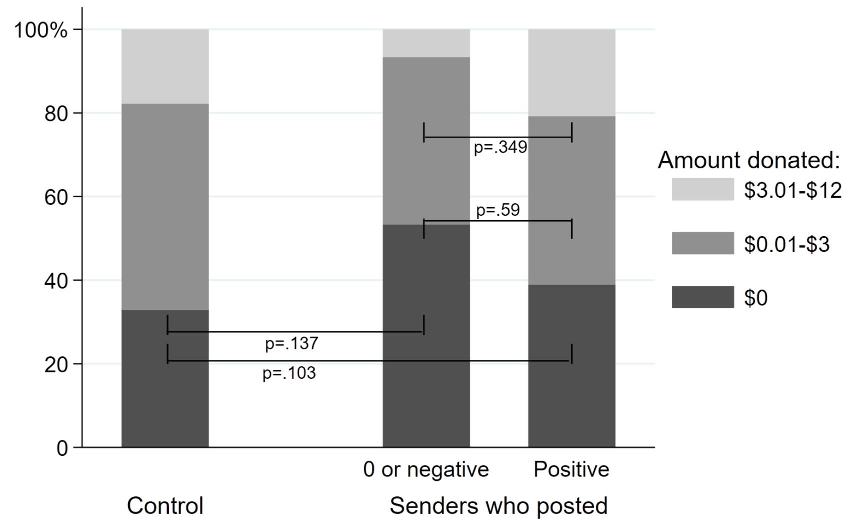


Figure 1.C.7: **Seeing posts (real or hypothetical) doesn't change how people think others will behave next week**



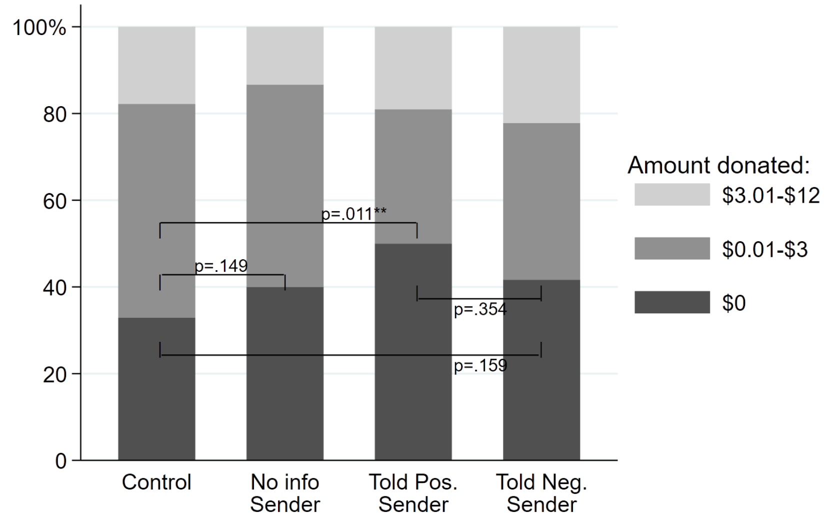
Note: Senders split by their beliefs of the average impact of posting on Receivers' donations. P-values are from a logistic regression of all observations on the likelihood to donate zero. Control variables are demographics (age, gender), parameters (alpha, gamma), and self-reported past activism (offline, donations).

Figure 1.C.8: **Senders who believe posts have a negative or zero impact crowd out donations to same degree as those who think posts have a positive impact**



Note: Senders split by their beliefs of the average impact of posting on Receivers' donations. P-values are from a logistic regression of all observations on the likelihood to donate zero (or donate more than $3). Control variables are demographics (age, gender), parameters (alpha, gamma), and self-reported past activism (offline, donations).

Figure 1.C.9: **Senders who post and believe posts have a non-positive impact crowd out donations just as much if not more than those who post and think posts have a positive impact**



Note: P-values from a logistic regression of all observations (including Receivers, not shown) on the likelihood of donating zero. Control variables are demographics (age, gender), parameters (alpha, gamma), and self-reported past activism (offline, donations).

Figure 1.C.10: **Senders who are told posts have a negative impact crowd out own donations nearly as much as senders who are told a positive impact, and not statistically differently so relative to other senders nor the control**

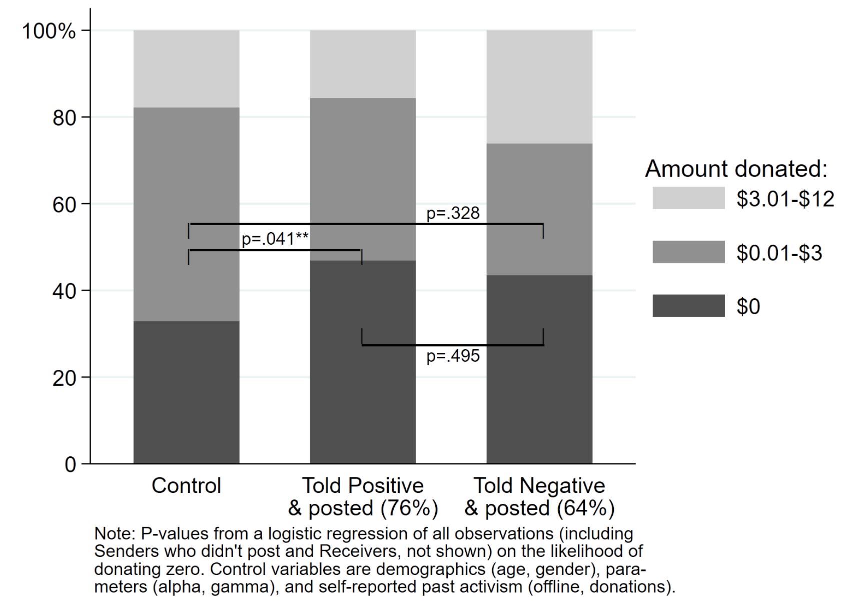


Figure 1.C.11: **Senders who are told posts have a negative impact and post crowd out own donations nearly as much as senders who are told a positive impact and post, but not statistically different from other senders nor the control**

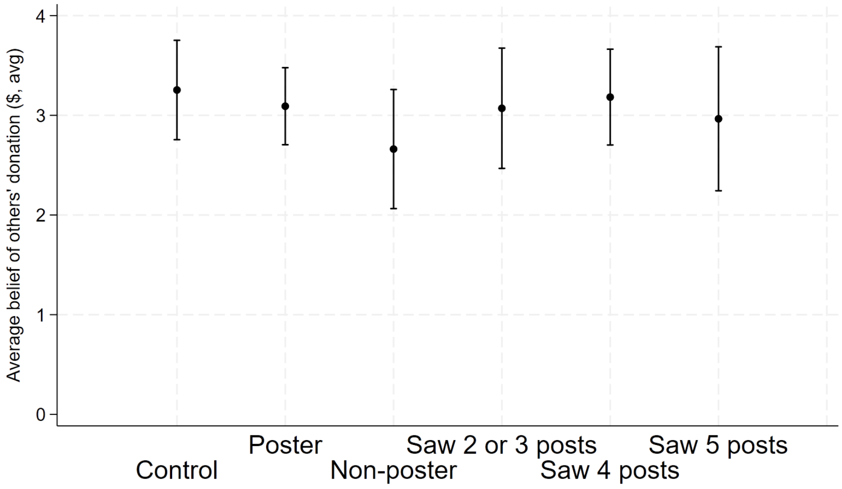
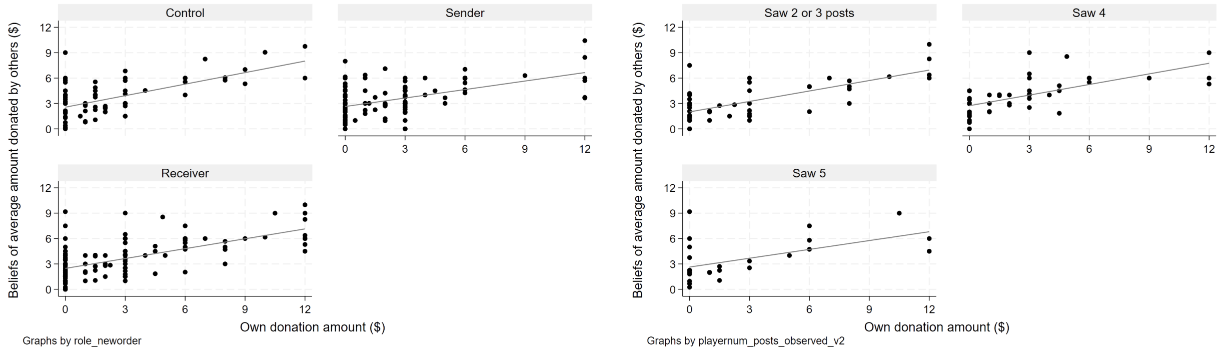


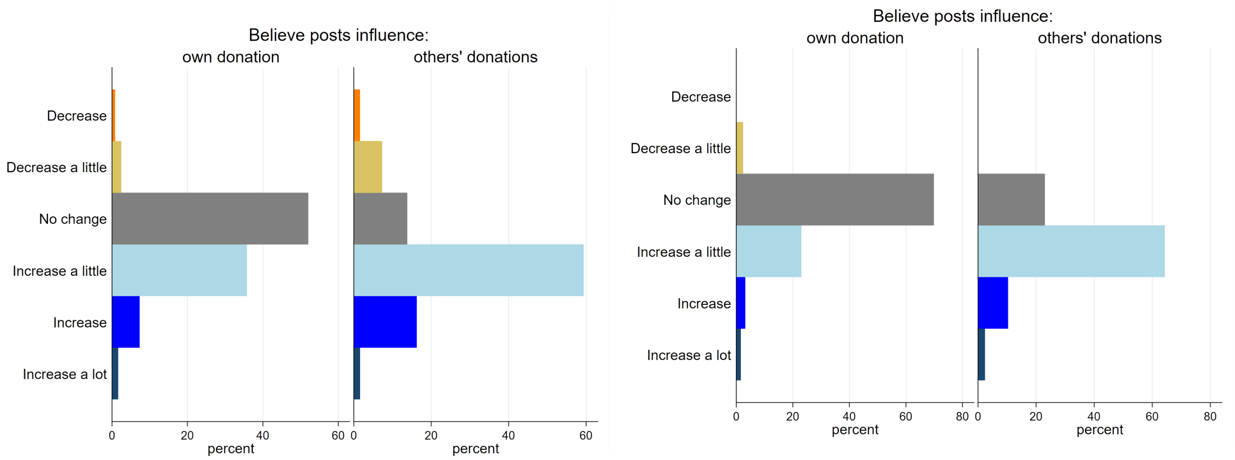
Figure 1.C.12: **Interacting with posts in any capacity does not change people's beliefs of how much others are donating on average**



(a) **All individuals guess others donate more when they donate more themselves**

(b) **Receivers who see more posts of support don't think others are donating more**

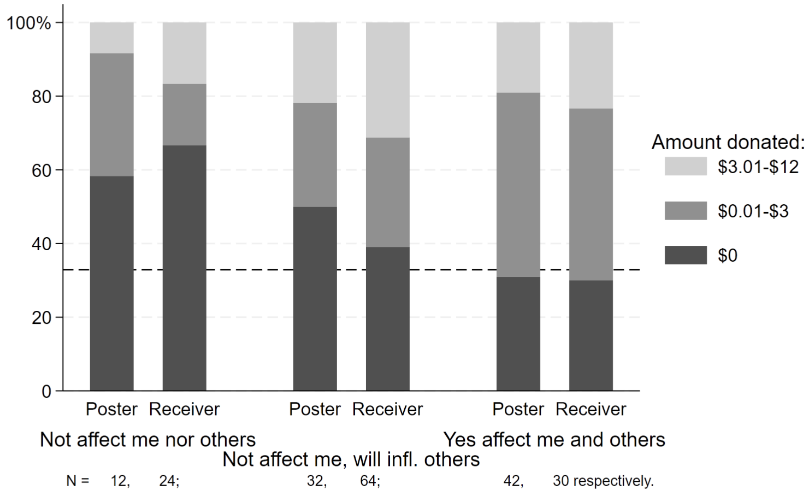
Figure 1.C.13: **People use their own donation decision as a basis for their best guess of others' donations, and this is not affected by the number of posts they see**



(a) **Senders**

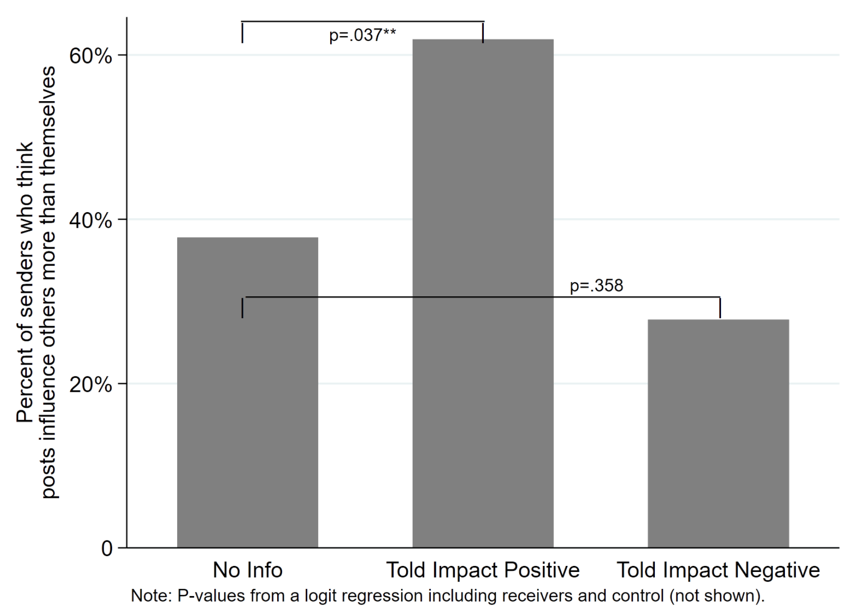
(b) **Receivers**

Figure 1.C.14: **Both senders and receivers have beliefs about the impact of posts on others versus themselves that are consistent with motivated reasoning**



N = 12, 24; 32, 64; 42, 30 respectively.

Figure 1.C.15: **Focusing on senders who post, we see a similar pattern where people who hold motivated beliefs or are pessimistic crowd out more than those who think posts have positive effects on themselves and others**



Note: P-values from a logit regression including receivers and control (not shown).

Figure 1.C.16: **Senders update their beliefs more strongly when they are given information that fits with prior beliefs/ motivated reasoning**

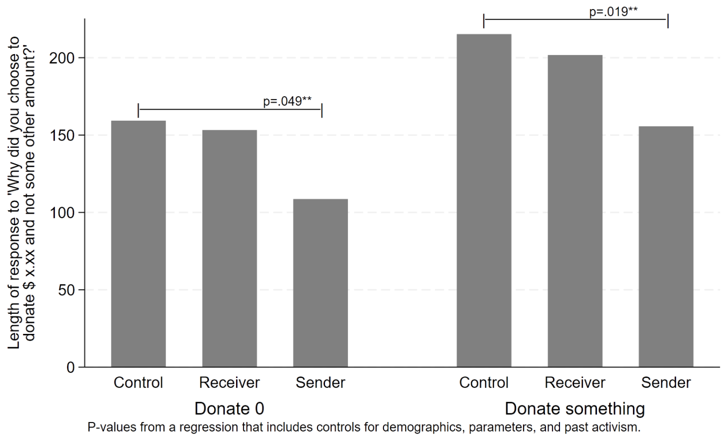


Figure 1.C.17: **Senders give significantly shorter explanations of their donation decision**

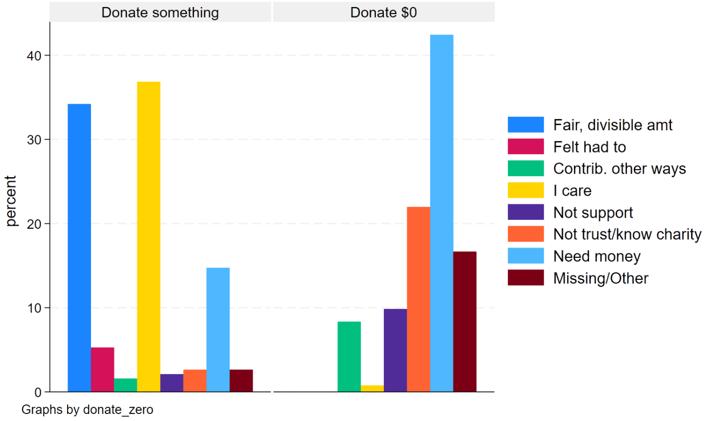


Figure 1.C.18: **Categories of donation decision explanations are reasonable and align with the donation amount**

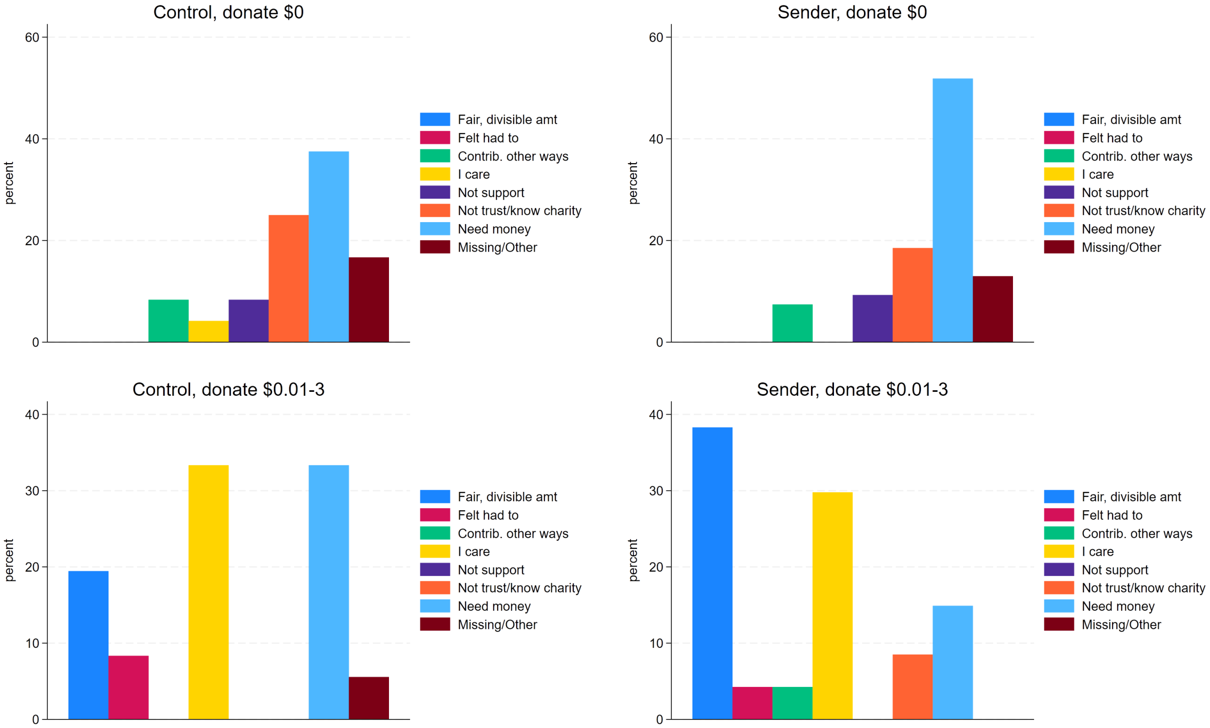


Figure 1.C.19: **Senders who don't donate are a little more likely to report needing they money relative to the control, with a similar reduction from those who donated a small amount (light blue bars)**

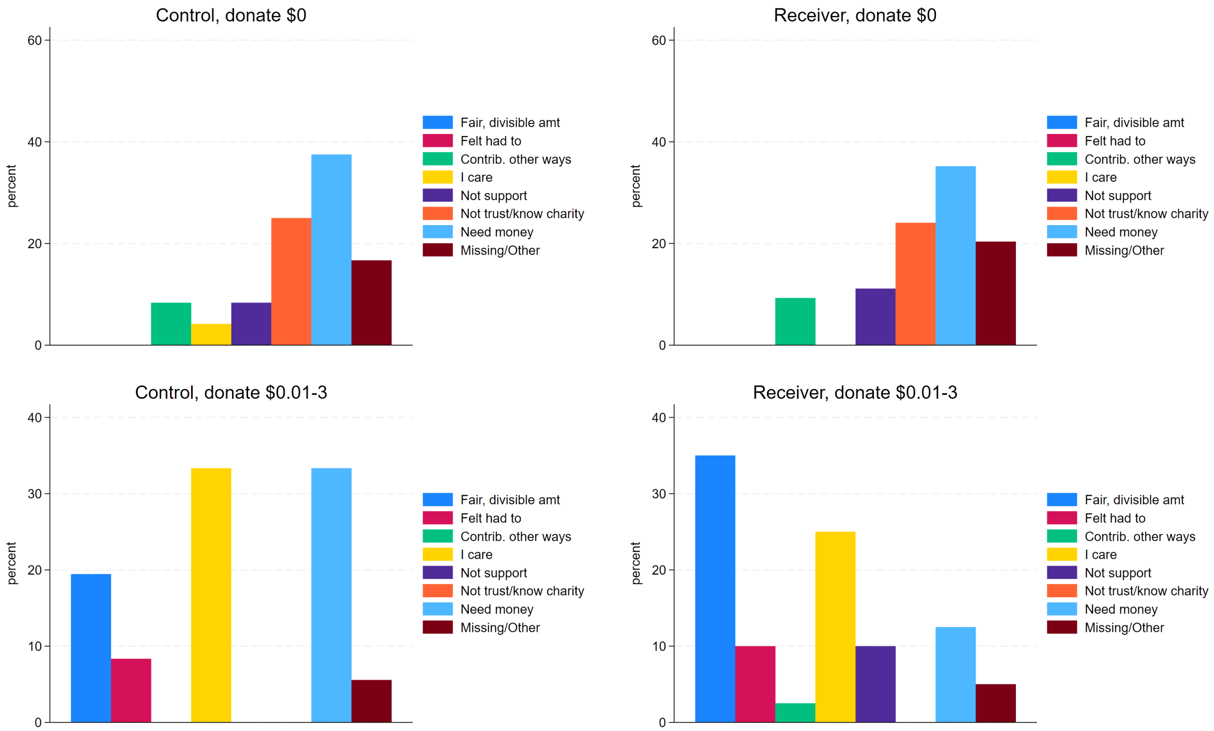
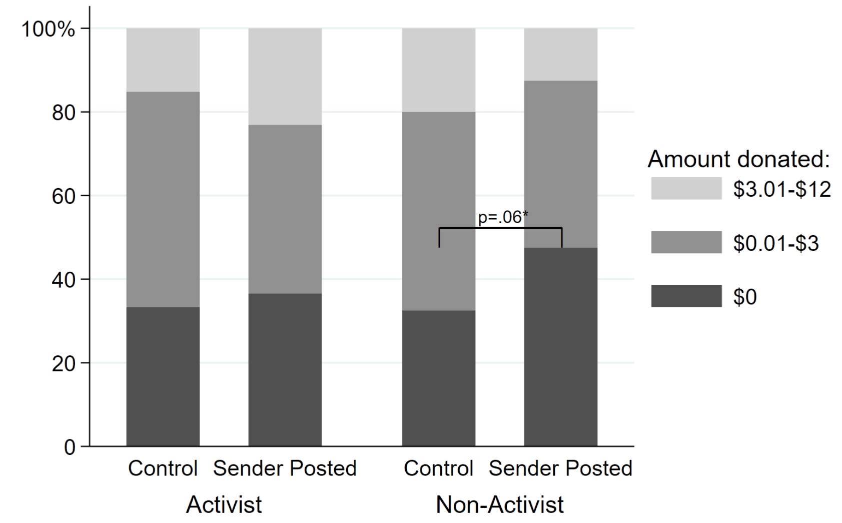
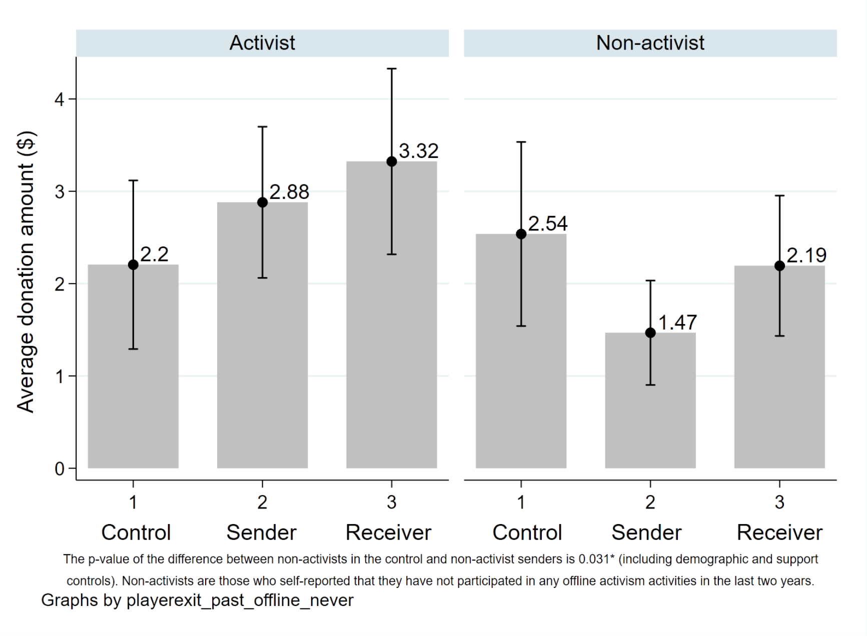


Figure 1.C.20: **Receivers give similar explanations for their donation decisions relative to the control**



Note: P-values are from a logistic regression of all observations (including Senders who didn't post and Receivers, not shown) on the likelihood to donate zero. Control variables are demographics (age, gender), parameters (alpha, gamma), and self-reported past activism (offline, donations).

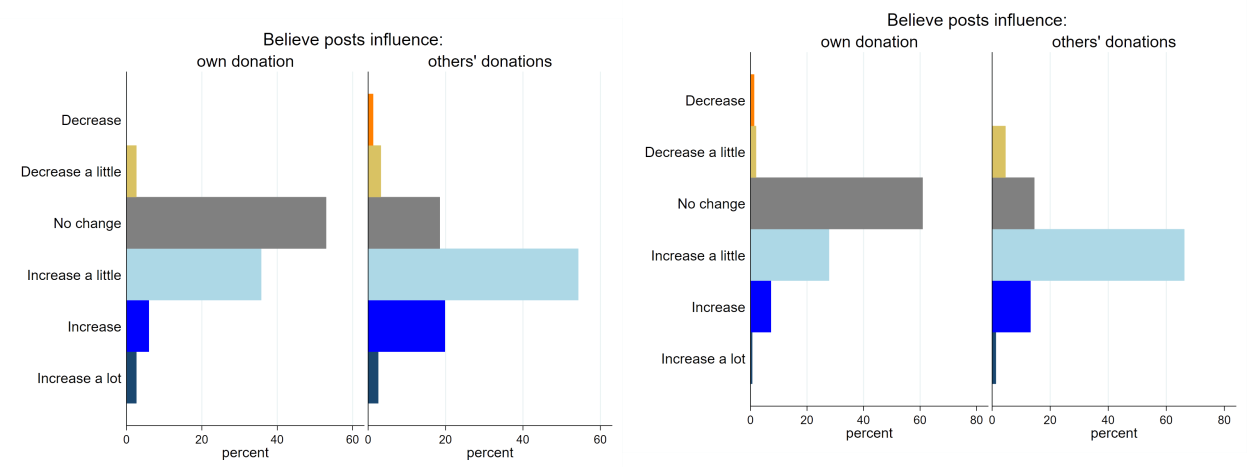
Figure 1.C.21: **Among activists, even senders who post donate similarly to control**



The p-value of the difference between non-activists in the control and non-activist senders is 0.031* (including demographic and support controls). Non-activists are those who self-reported that they have not participated in any offline activism activities in the last two years.

Graphs by playerexit_past_offline_never

Figure 1.C.22: **Along the intensive margin, it continues to hold that non-activists are those engaging in slacktivism; activists are unaffected or donate larger amounts**



(a) **Activists**

(b) **Non-Activists**

Figure 1.C.23: **Both activists and non-activists have beliefs consistent with motivated reasoning about the impact of posts on others versus themselves**

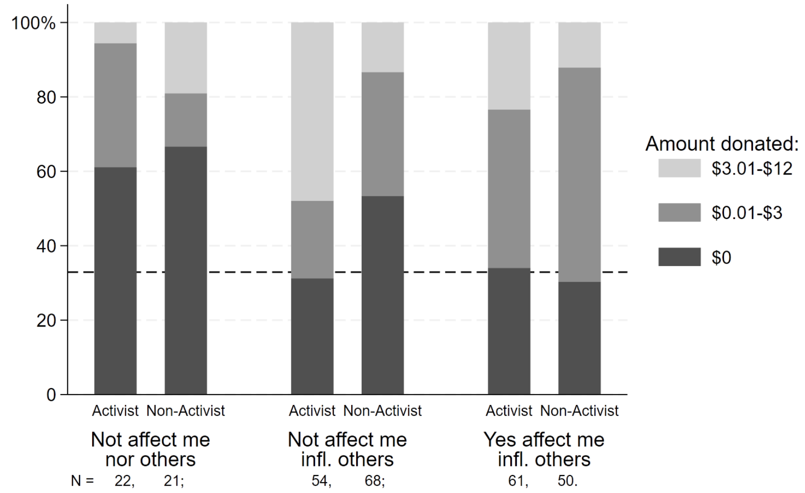


Figure 1.C.24: **The motivated beliefs channel of believing posts won't influence myself but will influence others is evident for non-activists but not activists (middle bars)**

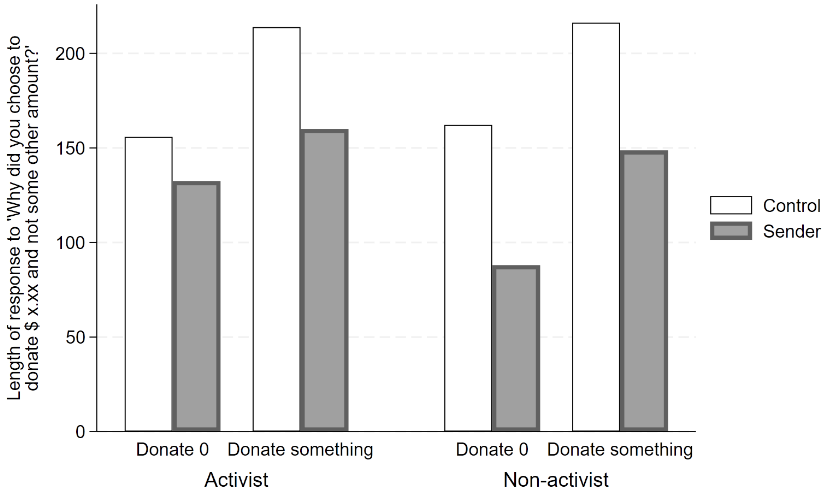
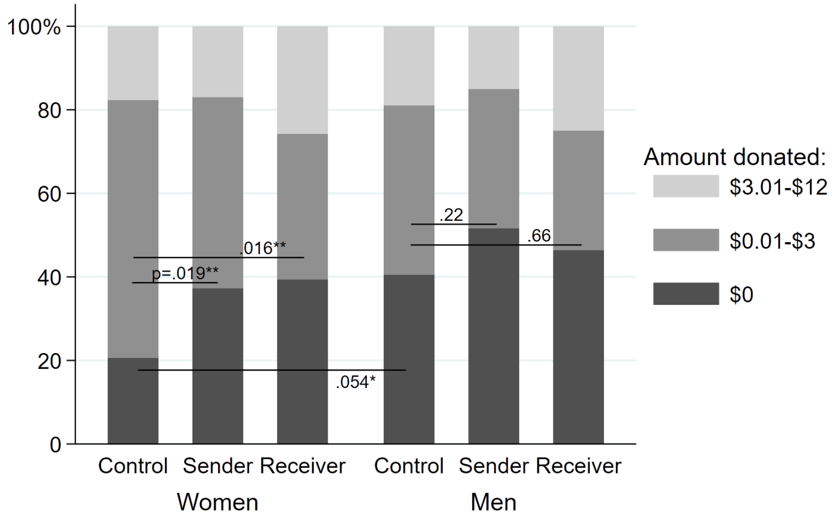
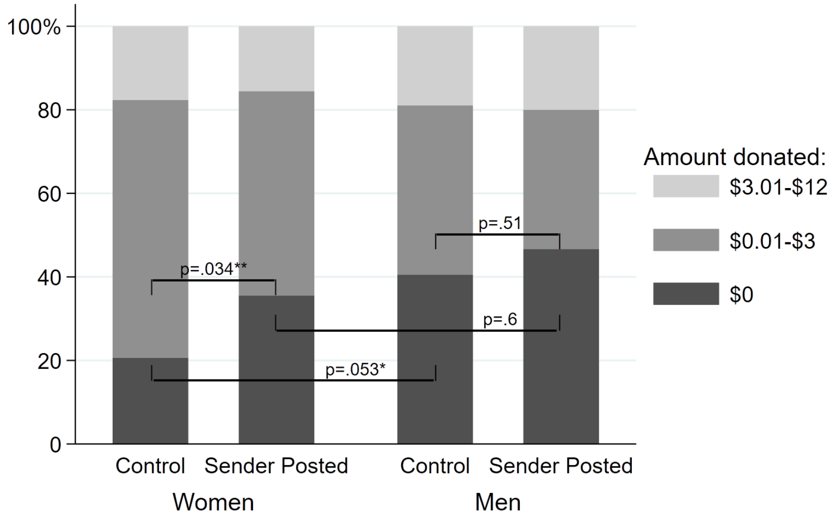


Figure 1.C.25: **Non-activist senders give even shorter explanations of their donation decision than activist senders**



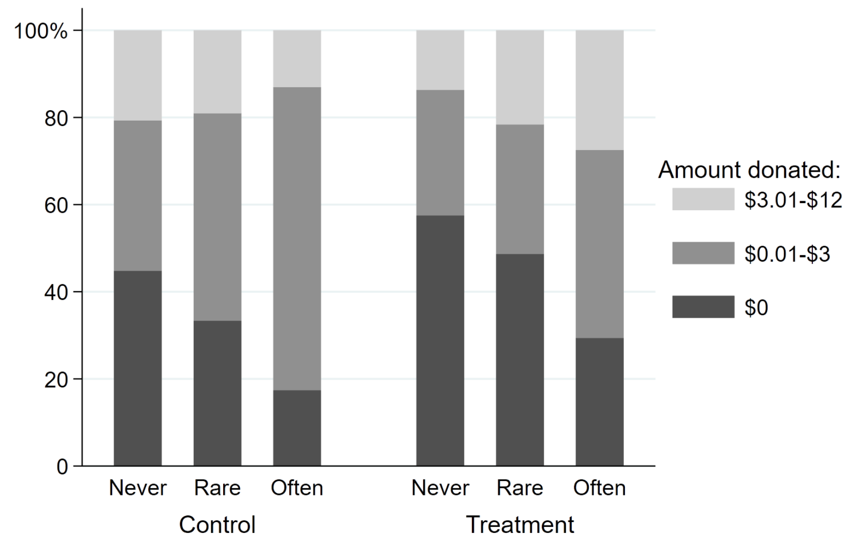
Note: P-values are from a logistic regression of all observations of men and women (excluding other genders but including Senders who didn't post and Receivers, not shown) on the likelihood to donate zero. Control variables are demographics (age), parameters (alpha, gamma), and self-reported past activism (offline, donations).

Figure 1.C.26: **Women are more likely to donate in the control group and women are more affected by the ability to post**



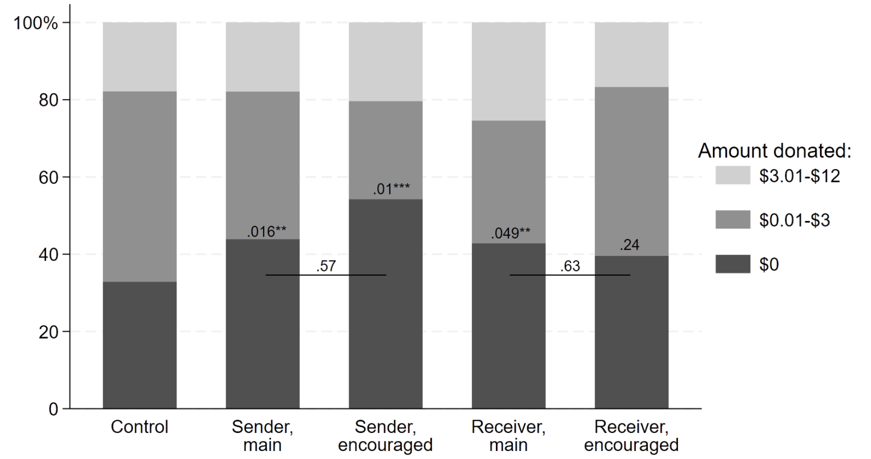
Note: P-values are from a logistic regression of all observations of men and women (excluding other genders but including Senders who didn't post and Receivers, not shown) on the likelihood to donate zero. Control variables are demographics (age), parameters (alpha, gamma), and self-reported past activism (offline, donations).

Figure 1.C.27: **The subgroup of women who post are also more affected by the ability to post since they are more generous in the control environment**



Note: P-values are from a logistic regression of all observations on the likelihood to donate zero (or donate more than $3). Control variables are demographics (age, gender), parameters (alpha, gamma), and self-reported past activism (offline, donations).

Figure 1.C.28: **Past donation behavior is predictive of how often they donate in the lab, but people crowd out donations to the same degree in the posting environment regardless of how often they donated in the past**



Note: Values above the bars are the p-values of the difference relative to control individuals from a logistic regression of all observations on the likelihood to donate zero. Values between bars are the policy treatment differences within that role.

Figure 1.C.29: **The nudge intervention has no effect on donation behavior for those who can send posts nor those who see posts**

Table 1.C.5: **Participants in the follow-on Nudge treatment are observably similar to participants in the main sample, albeit with a few more men and domestic students**

		Main			Nudge	
	All	**Control**	**Sender**	**Receiver**	**Sender**	**Receiver**
Demographics						
Age	20.3	20.3	20.4	20.4	20.3	20.1
Gender						
Male	48.7	50.7	48.8	44.4	54.2	58.3
Female	47.3	46.6	48.0	52.4	40.7	41.7
Non-binary / Other	1.9	0.0	2.4	1.6	5.1	0.0
Prefer not to say	1.2	2.7	0.8	1.8	0.0	0.0
Race						
Asian or Asian American	70.2	74.0	70.0	68.3	76.3	68.8
Black or African American	0.3	0.0	0.0	0.8	0.0	2.1
Hispanic or Latino	7.8	9.6	8.9	5.6	11.9	4.2
Native Amer. or Alaskan	0.6	0.0	0.0	1.6	0.0	0.0
White or Caucasian	11.5	6.9	11.4	14.3	6.8	8.3
Bi/Multiracial	4.4	4.1	4.1	4.8	3.4	4.2
Other	0.6	0.0	0.8	0.8	0.0	2.1
Prefer not to say	4.7	5.5	4.9	4.0	1.7	10.4
International student	47.4	55.6	47.2	42.9	32.2	39.6
Self-reports						
Racial justice is important (ω)	6.3	6.3	6.4	6.1	6.1	6.1
I want to contribute (α)	5.4	5.3	5.6	5.3	5.3	5.5
I need the money (γ)	5.3	5.4	5.4	5.1	5.8	5.6
Want others see I support (δ)	5.1	5.1	5.2	5.0	5.2	5.2
Past behavior						
Donate	2.3	2.1	2.4	2.2	2.3	2.4
Offline action	1.9	1.8	2.0	1.8	2.2	2.1
Social media	2.3	2.4	2.4	2.2	2.2	2.5
Online action	2.1	2.2	2.3	1.9	2.1	2.1
N	**429**	**73**	**126**	**123**	**59**	**48**

Note: Demographic variables are percentages, except age which is shown in years. Self-reports are the mean answer to 7-point Likert scale questions, see Appendix 1.B for details.

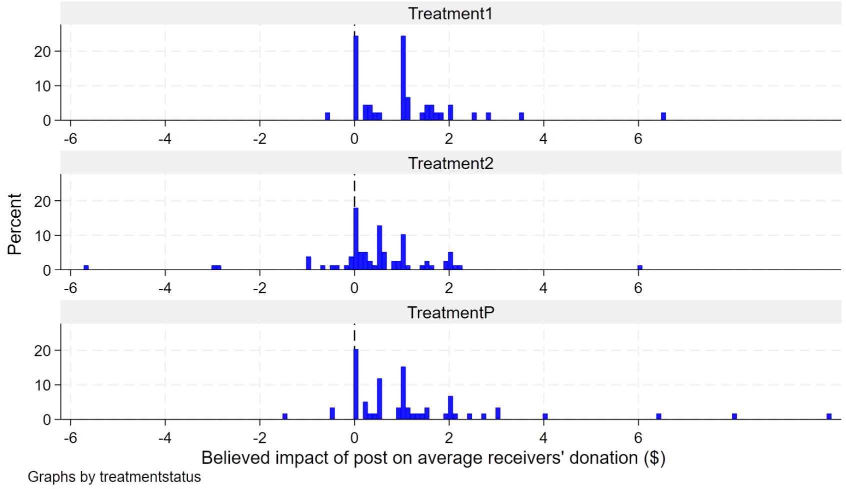


Figure 1.C.30: **Subjects in the Nudge treatment (TreatmentP) overestimate to the same extent as those in the No Information treatment (Treatment1)**

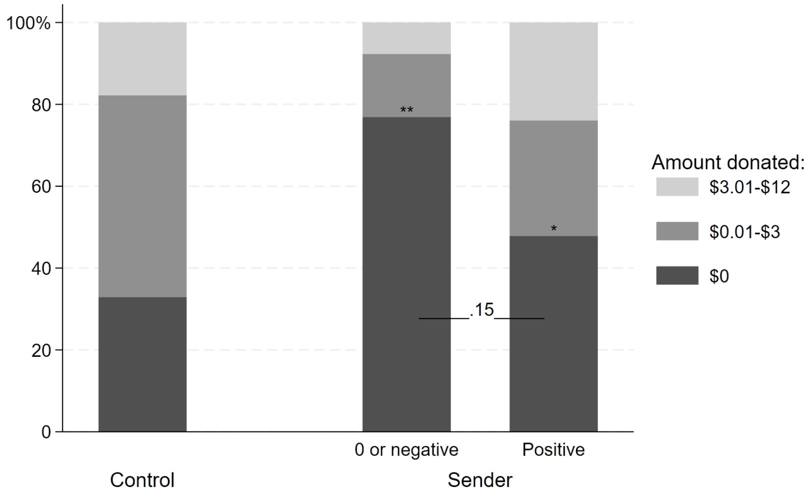


Figure 1.C.31: **Similar to the main sample, those who can post in the nudge treatment crowd out regardless of if they think posts have a positive impact**

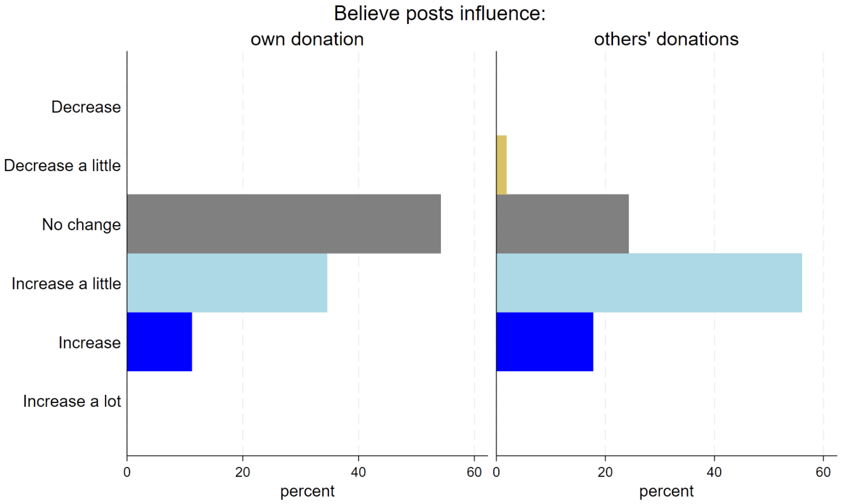


Figure 1.C.32: **Individuals in the Nudge treatment continue to hold motivated beliefs**

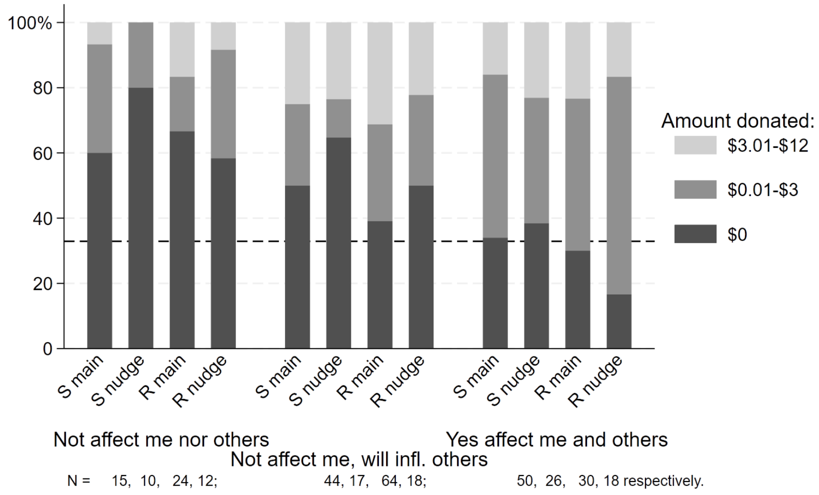


Figure 1.C.33: **Motivated beliefs continue to be correlated with donation behavior in the Nudge treatment**

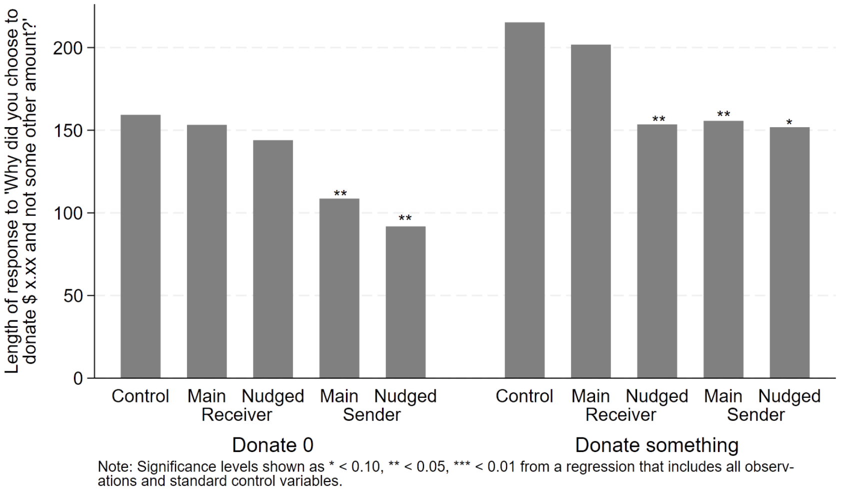


Figure 1.C.34: **Individuals in the Nudge treatment continue to give shorter explanations of their donation decision**

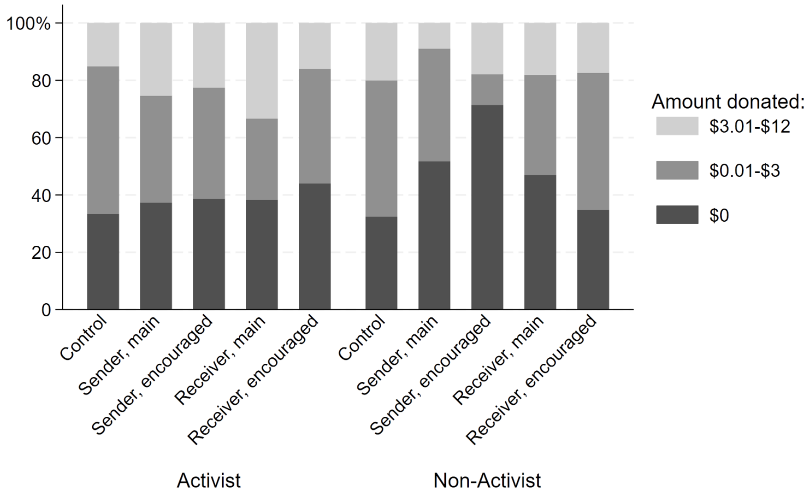


Figure 1.C.35: **Non-activists continue to be those who crowd out donations in the Nudge (encouraged) treatment**

Chapter 2

Disparate Impacts of Retaliation by Gender: Evidence from Airbnb

Amanda Bonheur

Abstract

Female whistleblowers face more retaliation in the workplace, and I provide novel evidence that retaliation also disproportionately disadvantages women on peer-to-peer review platforms. I leverage an exogenous policy change implemented by Airbnb on July 10, 2014 that made reviews simultaneous reveal, meaning that an individual could no longer see the comments the other party left before submitting their own review. The change removed the ability to retaliate to negative reviews and reciprocate extremely positive reviews, both of which encouraged honesty. Using review data for Airbnb listings, I find that both male and female guests respond by reducing the positivity of reviews they write for hosts. However, reviews towards male hosts experienced a larger negative shift of review sentiment, indicating that reviews of male hosts were twice as artificially elevated relative to those of female hosts. Reviews to male hosts show evidence of fear of retaliation in addition to reciprocity, while reviews of female hosts indicate reciprocity only. This demonstrates that peer-to-peer review systems, which are becoming increasingly common especially in non-traditional labor market settings, are not gender-neutral in implementation if retaliation is possible.

2.1 Introduction & Motivation

Gender gaps in academic and professional settings have been extensively studied (Boring 2017, Hengel 2022, Mitchell & Martin 2018, among many others), while gender gaps in peer-to-peer reviews are less understood. Peer-to-peer review systems are becoming increasingly common, especially in non-traditional labor market settings. This means that reviews often translate to income. Retaliation and honesty in review systems are beginning to be studied (Fradkin et al 2021, Mousavi & Zhao 2022, Bolton et al 2013), but how this interacts with gender is unknown.[1] Retaliatory behavior and reciprocity vary by gender in other settings (Dehdari et al 2019, Liyanarachchi & Adler 2011, Fatoki 2013, Rehg et al 2008, Kundro & Rothbard 2022, Chaudhuri & Gangadharan 2003, Dittrich 2015),[2] suggesting that interactions by gender is an important area to study. This paper provides the novel contribution of disparate impacts of retaliation and reciprocity by gender in a real-world, large scale context with economic consequences.

I leverage an exogenous policy change implemented by Airbnb on July 10, 2014 that removed the strategic interaction of reviews, thereby removing the ability to retaliate to negative reviews and reciprocate positive reviews. Airbnb is a peer-to-peer online marketplace to rent rooms/properties. Hosts and guests find each other using the Airbnb platform, and can leave reviews for the other party after the stay. Prior to mid-2014, reviews were posted immediately. The second person to write a review saw the review written for them before acting. Under this system that allowed for retaliation and reciprocation, people are deterred from reviewing honestly when the stay was subpar and encouraged to exaggerate when the stay was good. After the change, reviews are not posted until both parties review the other. Reviewing honestly is now incentivized

[1] I recognize that gender is not binary and that female and women mean different things. For this paper, I focus on men and women since I cannot identify non-binary and other gendered individuals (I use self-reported name to impute gender for commonly-gendered names). Since gender is based off of self-entered names, it is possible that my sample includes non-binary and transgender individuals. I use female as an adjective, e.g. ‘female guests’, to be less wordy than ‘guests who are women’, and I use this term to talk about every individual who identifies as a woman. For more details of how gender is imputed, see Section 2.4.2.

[2] Women are less likely to retaliate in a game show setting (Dehdari et al 2019) and women sometimes respond to potential retaliation differently in whistleblowing contexts (Liyanarachchi & Adler 2011, Fatoki 2013, Rehg et al 2008, Kundro & Rothbard 2022). Chaudhuri and Gangadharan (2003) find that women show higher levels of reciprocity, while Dittrich (2015) finds that men are more likely to engage in reciprocal behavior.

since there is no opportunity for the other person to retaliate or reciprocate reviews.

Detailed listing, host, and review data comes from InsideAirbnb, a data collection project that regularly scrapes listings on Airbnb. I have the universe of reviews for each active listing in 10 U.S. cities as of July 2015, one year after the policy change. Gender and review sentiment are imputed using text analysis methods (Bird, Klein, & Loper 2009; Hutto & Gilbert 2015; Pérez 2016).

I find new evidence of a male advantage created by strategic interactions in the review system. The incentives prior to July 2014 created artificially elevated reviews. This is evidenced by the average review becoming less positive after the policy change and previous research (Fradkin et al 2021; Mousavi & Zhao 2022). Crucially, I find that reviews towards male hosts experienced a drop in sentiment that was twice as large as the one towards female hosts, implying a male advantage prior to July 2014. The drop in review sentiment is true on average and across the distribution, with less positive reviews experiencing a greater shock from removing the ability to retaliate. All guests, regardless of gender, change review-giving behavior in the same way after the removal of retaliation and reciprocity channels. My results imply that the sequential review system induced a gender bias through differentially elevated reviews for men.

There are two mechanism interpretations consistent with these results. The first mechanism is retaliation: the policy removed the opportunity to retaliate and the associated fear of retaliation. The interpretation of this channel is that people fear retaliation from men more, and male and female guests are equally likely to be deterred by potential retaliation. This is consistent with previous literature that shows that men are more likely to retaliate, and that there is no consensus on gender differences in reaction to retaliation consequences (Dehdari et al 2019, Liyanarachchi & Adler 2011, Fatoki 2013). The second mechanism is reciprocity: the policy removed the opportunity to prod and reciprocate good reviews. For my results to be coming through this reciprocity channel, it must be that male hosts were more likely to reciprocate positive reviews, so there was a larger benefit to writing very positive reviews. As a result, all reviews are less exaggerated under simultaneous reveal, and this is especially true for male hosts. This is possible following Dittrich

(2015) but not Chaudhuri & Gangadharan (2003).

I show that reviews for male hosts were affected by both fear of retaliation and anticipation of reciprocity, while reviews for female hosts were affected by anticipation of reciprocity only. Reviews for male hosts become not only less positive but more negative, which implies that people were constrained in their ability to leave negative feedback. It follows that the retaliation story is at play. Alternatively, reviews for female hosts become less positive but not more negative, which suggests a reciprocity mechanism whereby guests used to exaggerate positive sentiment to prod a positive review in return.

These results are consistent with my model where host-guest interactions are represented as a sequential versus simultaneous decision game of how positive of a review to write, conditional on choosing to write a review. The model yields predictions about the effect of removing the retaliation and reciprocity components. First, making reviews simultaneous reveal will make reviews less positive and more truthful on average. Second, the entire distribution of review sentiment will shift to the left. Intuitively, a good stay will be reviewed positively regardless, but a so-so stay is more likely to be reviewed and reviewed honestly afterwards. This effect will vary by gender based on beliefs about the likelihood of retaliation. If guests believe that male hosts are more likely to retaliate, the new policy will induce a larger downward shift in review sentiment towards male hosts. If female guests are more deterred by the fear of retaliation or more encouraged by possible reciprocity, they would have been more constrained in their review-giving behavior, implying a larger adjustment.

This new evidence has policy implications for designing online peer-to-peer review systems. When reviews interact with each other, retaliation-honesty and reciprocity-honesty trade-offs are created. When people fear retaliation from men more, this creates a male advantage. Notice that fear of retaliation creates this advantage for men. Whether men do retaliate more or not isn't necessary for the advantage to occur. Even though actual retaliation may occur infrequently,[3] this subtlety means that consequences of retaliation extend beyond direct receivers of retaliation. Fear

[3] Fradkin et al (2021) finds that conditional on a host leaving a negative review, the guest responds with a negative review 7% of the time prior to the policy change (and 2.2% after).

of retaliation is experienced by all.

A gender gap in reviews is a gender gap in earnings since reviews translate to bookings and the majority of hosts use Airbnb to supplement their income. Women over age 60 are the fastest growing host demographic and one-sixth of U.S. hosts are teachers. Half of hosts are low to moderate income households using the earned income to pay for living expenses (Smith 2016). The median host earns \$440 per month (\$924 on average) (Morris 2021). Reviews affect future bookings, so gender differences in reviews are economically meaningful. Ninety-four percent (94%) of people say that a bad review online has deterred them from interacting with a company (Penaflorida 2020). The number of 5-star reviews is a determining factor in how high a listing is in a guest's search (How Search Results Work n.b., How to Get 5-Star Reviews on Airbnb 2024). Identifying any gender differences in Airbnb reviews is one form of income disparities by gender.

Others have studied the Airbnb platform, but I focus on a different set of outcomes, which occurs after the stay. Discrimination by race/ethnicity, disability, and gender has been found in acceptance rates (Edelman et al 2017; Hardonk 2020; Ameri et al 2020) and demand for rentals (Kakar et al 2018; Marchenko 2019). I study reviews, which are written only for selected interactions that resulted in stays. In other words, conditional on existing discrimination in the acceptance/booking stage and price charged, this paper focuses on a second stage of disparate impacts in reviews. Reviews affect reputation and can lead to further discrimination and/or income inequality.

Two-way review systems are staples in peer-to-peer marketplaces (e.g., Airbnb, Uber, eBay) and must be designed to be gender-neutral in implementation. For this to be true they should be designed without retaliation and reciprocity capabilities. This conclusion follows from the fact that reviews are elevated when retaliation and reciprocity are possible, with men having an advantage that is twice as large. Since reviews affect bookings, this exacerbates income inequality.

2.2 Background

Airbnb is a peer-to-peer online marketplace to rent rooms/properties that relies on its review system. Airbnb is typically used for short-term vacation rentals. Hosts list their place on the platform along with a short description. Guests book places to stay through the Airbnb website or app. Trust is crucial and is built through reviews (Bolton et al 2013; Cameron 2017; Shapiro 2017). The system is self-moderated and self-sustained. Honest reviews will weed out the bad listings and/or discourage poor behavior.

The process to create bookings involves hosts and guests making a match. Hosts generate a listing for their place that includes pictures, a description, the price per night, their name, checkout rules, location, and other characteristics. All previous reviews written for that listing are also included. Guests search through listings based on location, price point, size, quality, and reviews. Guests choose where they would like to stay and the host confirms.[4]

After a stay, both the host and guest can leave a review for the other. I define a stay as one host-guest interaction, which could be a one night or 3 week stint of the guest renting out the host's property. Either party can review the other first. Once written, reviews for an individual are public to every potential person interacting with them in the future. Someone thinking of staying with a host can see all previous reviews written for that host. Similarly, a host can see all previous reviews written for a guest who is asking to rent their place. In this setting, individuals rarely have repeated interactions with the same person, but reviews can influence others' decisions to interact with that individual.[5]

Airbnb announced new review rules on July 10, 2014, the day the changes became effective. This policy change was unanticipated (Building Trust 2014; Protalinski 2014),[6] providing us with a natural quasi-experiment.

Prior to July 10, 2014, reviews were posted immediately. This created a sequential game where either party could review first, and the other could respond to that review. There was a 30 day deadline to write reviews.

As of July 10, 2014, neither review is posted until both parties have written their review

[4] Hosts can choose whether they must approve each guest or if guests are able to Instant Book. The instant book feature was introduced in November 2010 in an effort to make Airbnb more favorable for guests and reduce discrimination. Hosts who use Instant Book see a large increase in number of bookings (The Airbnb Story (n.b.).

[5] It is possible for people to have repeat interactions with each other, but this is a tiny fraction of Airbnb use.

[6] Note that both of these articles were posted on the same day as the policy change.

(or the 14 day deadline has passed). The new 14 day deadline is not binding as previous research has found that the vast majority of reviews are written within a couple days of checkout. The simultaneous reveal of reviews removes the sequential nature and any possibility of retaliation for leaving a bad review or reciprocation for leaving a glowing review.

Not being able to read the review written for you before writing your own review incentivizes truth-telling. Consider the following scenario: I had a stay that was subpar since the shower was dirty and the host asked me to leave 2 hours before the agreed checkout time. If this occurred in 2013, and I write a truthful review, the host may leave a bad review for me saying that I was disrespectful. On the other hand, if this occurred in 2015, I could write an honest review without any repercussions on the review the host writes for me. The reverse is also true; it is also in the host's best interest to write honest reviews. Imagine that I rent out the spare bedroom in my apartment. I have a guest who was generally respectful but made unreasonable noise early in the morning, which is against my house rules. I need more positive reviews to secure future bookings. As a result, in 2013 I would omit these details from the review so that the guest leaves a good review for me. After July 2014, I could be honest without jeopardizing my feedback. Both of these scenarios could differ by my perception of the likelihood of the other party to retaliate or reciprocate. If I believe men are more likely to react, then I will modify my behavior to a greater extent when interacting with a male host or guest. I will be more likely to omit negative details and more likely to overstate positive details. Porges (2014) and Mousavi and Zhao (2022) confirm the improved honesty in reviews.

All meaningful channels of retaliation and reciprocity are removed with this policy change, since hosts cannot delete reviews. However, they are given an opportunity to respond to guest reviews within 30 days with a public comment (Penaflorida 2020). Since the responses are visible to everyone, they must be considered. For example, a typical comment reads "I'm so sorry you had a bad experience. Thank you for your feedback, it will help me improve this stay for others!". It would send the wrong message to reply to a negative review with attacks on the guest. Since the comments are public and appear on their own review page, they cannot be as retaliatory in nature.

The host is limited to apologizing and thanking them for their feedback.

At the same time, review behavior continues to be motivated by their economic importance. Positive reviews translate directly to increased bookings and income for hosts.[7] More than half of hosts use earned income to stay in their homes, a number that climbs to 72% in New York City (Our Community of Teacher Hosts 2020; HNN Newswire 2015).[8] Half of hosts use earnings to pay for typical expenses and 52% are low to moderate income households (Shared Opportunity 2015).[9] Teachers and retirees commonly host to supplement their income. One in ten hosts in the U.S. are teachers (Rosenberg 2018),[10] and women age 60 and up are the fastest growing host demographic (Airbnb's Growing Community 2016, Kovachevska 2020). The median host earned $440 per month ($924 per month on average) (Yates 2020, Morris 2021). For guests, positive reviews mean a higher chance of being accepted for future stays. Hosts are more active than guests on the website, but both parties have incentives to ensure they obtain positive reviews even if this requires modifying their own review away from their true experience.

I leverage this natural setting to observe perceptions of retaliation and reciprocity on behavior in an understudied part of the labor market. This paper analyzes one policy with a disparate impact on women, and provides evidence that review systems with retaliation are not gender-neutral in implementation.

2.3 Related Literature

This paper combines two areas of work. The first is gender differences in strategic settings, and the second is the study of two-way review systems.

Previous studies of gender differences in settings with evaluations find that women often

[7] Shatford (2018) found that superhosts, who have at least a 4.8 star average, have 50% more revenue and an 81% average occupancy rate, which is much higher than the overall average of roughly 48% (Average Airbnb Occupancy Rates n.b.). Although Airbnb did not introduce the superhost designation until 2016 (Shatford 2018), these statistics highlight how reviews impact bookings.

[8] Data based on the year 2019 for all hosts around the world and 2014 for New York.

[9] Data using years 2012-2015.

[10] Another source found that one-sixth of hosts are teachers; in 2019, 65,000 hosts were teachers, which is around 16% of all hosts (Teacher Hosts Earned 2020).

receive lower or biased assessments. In an academic context, Macnell et al. (2014) and Mitchell & Martin (2018) find that students give biased evaluations to women instructors. In online courses, instructors that are listed as female are given worse reviews regardless of the true gender of the instructor. Anne Boring (2017) shows that this result is driven mostly by male students having a preference for male instructors. In professional environments, women receive more vague performance reviews (Correll & Simard 2016) and are held to higher standards in peer reviewed papers (Hengel 2022).

Retaliation is common in strategic settings, and there is evidence that genders exhibit different extents of this behavior. Dehdari and co-authors (2019) find that women are roughly 23% less likely to retaliate than men in a quiz show setting. The amount men retaliate does not depend on the gender of the receiver. However, women retaliate depending on the gender of the other person; women more likely to seek revenge against men than they are to women. Liyanarachchi & Adler (2011) find that gender, age, and retaliatory consequences interplay in the likelihood of someone being a whistleblower. Using a sample of accountants, the propensity for older women (over age 45) to blow the whistle declines as the retaliation threat increases. The same increase in consequences does not change male accountants' propensity. Fatoki (2013) similarly finds that intent to whistleblow decreases with retaliation, but does not find gender differences. This suggests that context is important for the materialization of gender differences to the threat of retaliation.

Gender differences in reciprocity are similarly sensitive to context. Most studies of reciprocity show that women show higher levels of reciprocity (Chaudhuri & Gangadhuran 2003; Croson & Buchan 1999; Buchan et al 2008). These studies have undergraduates playing an investment trust game in the lab. Buchan et al (2008) adds that women likely feel more obligated to trust and reciprocate while men behave more strategically. In a real-effort dictator game where the size of the pot is determined by effort of the recipient, Heinz et al (2012) find that women behave more reciprocally. However, Dittrich (2015) finds that men are more likely to engage in reciprocal behavior using an online trust game with heterogeneous subjects representative of the German population. Men have an inverse U-shape relationship between reciprocity and age, but no such relationship exists for women. Similarly, Garbarino and Slonim (2009) report gender differences being sensitive to age and amounts received use an online experiment with U.S. individuals age 18-84. They add that men and women of all ages trust women and older people more than men and younger people. Given that the population of hosts and guests on Airbnb are generally older than undergraduate students, reciprocal behavior by gender is ex-ante ambiguous.

Contrary to the studies aforementioned, Airbnb is a large-scale, real-world context with relatively large stakes and multiple interactions. Many individuals use Airbnb regularly, meaning that interactions with the app are not one-shot. Gender-specific responses to reciprocity and retaliation will result in discrepancies in reviews, bookings, and income. These high stakes affect many individuals, and magnify the importance of understanding gender differences in retaliation and reciprocity in this setting.

The other broad literature that this research builds upon is that of behavior in two-way review systems. Tadelis (2016) provides a review of online marketplace feedback systems, including discussing reputation and trust, common feedback systems, two-way vs one-way systems, and problems of bias. In particular, Dellarocas & Wood (2008) find that there is reporting bias such that people are less likely to write reviews when the experience is not positive. Nosko and Tadelis (2015) provide further evidence of reporting bias. They use eBay internal data and find more complaints than negative reviews. Calabral & Hortaçsu (2010) study the dynamics of seller reputation on Ebay. Once a seller receives negative feedback, weekly sales drop by 13 percentage points, subsequent negative reviews are more common, and sellers with lower reputation are more likely to exit. Bolton et al (2013) study reciprocity and retaliation on eBay when reviews were two-sided. They find that sellers wait to get feedback before giving their feedback to buyers, and much of sellers' negative feedback is retaliatory. Feedback being blind reduced the correlation between sellers' and buyers' reviews, increased negative feedback, and provided more informative reviews as they can better signal sellers' quality. Interestingly, when one party gave a negative review, the retaliatory negative review was posted much quicker; retaliatory reviews were posted mostly the same day, whereas positive reciprocity had an average gap between reviews of 2.5 days.

This highlights an honesty-retaliation tradeoff as fear of retaliation disincentivizes honesty.

Other authors have studied the July 2014 Airbnb policy change and found evidence of the retaliation-honesty tradeoff. Mousavi & Zhao (2022) analyze the effect of the policy change on review characteristics. They find lengthier, less positive reviews with more variety in content after reviews are simultaneous reveal. Reviews become more objective temporarily, and reviews permanently become less about personal opinions.They find that without fear of retaliation, people are less likely to pretend to be nice. This is evidenced with fewer positive words with no change in negative emotions. Fradkin et al (2021) ran an experiment two months prior to the policy change where they randomly assigned simultaneous reveal reviews status to one-third of Airbnb hosts. Their pilot-like experiment yielded small effects on increasing the negativity of reviews and decreasing retaliatory behavior. They find that guests decrease retaliatory behavior; conditional on a host leaving a negative review first, guests responding with a negative review falls from 7% to 2.2%. Guests also write more reviews (1.2 pp), write more reviews with negative sentiment (1 pp), and leave more 3 and 4 star reviews and fewer 1 star reviews than before. Hosts substantially increase the amount of reviews given (7 pp) but don't leave statistically more negative reviews than before. They also found an increase in private comments and shorter review times (which they attribute to curiosity; one is notified when the other party reviewed, but could not read the review until they submitted a review or 14 days passed). Overall, previous studies find that making reviews blind induces more honest and more negative reviews from a reduction in the fear of retaliation and a reduction in the incentives to overstate positive sentiment.

Importantly, no prior research about the Airbnb 2014 policy change investigates disparate impacts by gender. In turn, I replicate their overall findings on the negativity of reviews but highlight how the policy differentially impacts individuals by gender.

Although no papers have studied the retaliation-honesty tradeoff by gender, many have found evidence of Airbnb discrimination. A Harvard study found that guests with African-American names are 16% less likely to be accepted compared to identical guests with distinctly White names (Edelman et al 2017). Hardonk (2020) uses the data from the aforementioned study and further breakdown the results by gender, showing that female guests are more likely to be approved due to having more favorable stereotypes about their behavior. Ahuja & Lyons (2017) find that men in same-sex relationships are 20 to 30% less likely to be accepted, which is mostly through non-responses than outright rejections. Ameri et al (2020) find that individuals with disabilities are less likely to be accepted, and this effect is only partially attenuated among hosts who advertise as wheelchair accessible. Hosts also face booking discrimination. Marchenko (2019) finds that despite charging lower prices, Black and Asian hosts face lower demand. Kakar et al (2018) find that Asian and Hispanic hosts charge 8 to 10% lower prices relative to White counterparts with equivalent properties. These places have similar occupancy rates, which suggests that Asian and Hispanic hosts charge lower prices in anticipation of discrimination. Airbnb guests face discrimination in acceptance rates while hosts face discrimination in bookings and so charge lower prices.

This discrimination is important to consider when interpreting my results. Reviews are written only for interactions that resulted in stays. That is, reviews are a selected subset. Conditional on the current amount of discrimination at the guest-host matching stage, disparate impacts in reviews will only compound these effects. For this reason, removing retaliation-inducing bias is even more critical.

2.4 Data

The data comes from Inside Airbnb, a data collection project that compiles publicly available Airbnb listings information for public use.[11] This site scrapes Airbnb every so often to create a detailed snapshot of listings, hosts, and reviews for a large number of cities around the world. I have detailed data on all listings in a given city that are active at the point in time that it is scraped. The listing data includes information about the host and all guest reviews written for that host/listing.

Listing data includes description information of the place such as the listing title, neighbor-

[11] Inside Airbnb was developed by Murray Cox and is independent from Airbnb (Inside Airbnb n.b.). The data is available at `http://insideairbnb.com/get-the-data.html`

hood, zipcode, and ID number; price; characteristics such as room type (entire home/apartment, private room, or shared room), property type (condo, house, apartment, etc), and number of bedrooms; and other factors such as house rules, cancellation policy, and amenities (Wi-Fi, cable, air conditioning, fireplace, etc).

Information about the host such as their self-entered (first) name, a short description of themselves, host response time, host response rate, how long they have been a host, and the total number of listings by that host is included.[12]

The review data contains information on individual reviews, which includes the self-entered name of the reviewer, the date of the review, and the text of the review itself. The reviews can be matched to the listings using host and listing ID numbers.

There are a couple of things that guests see about the host that is not in my dataset, namely pictures (of the place and one of the host) and star ratings out of 5 (overall rating along with component ratings for cleanliness, communication, check-in, accuracy, location, and value).[13]

Inside Airbnb has been used by researchers in Economics and other fields (Kakar et al 2018; Li 2018; Marchenko 2019; Dann et al 2019; Barron et al 2021). Alsudais (2021) investigates the data quality and finds some incorrect review-to-listing matching resulting from the introduction of Airbnb Experiences in 2016.[14] My data ends before Airbnb Experiences was introduced, so all listing ID's are unique and this issue is not present in my sample.

The benefits of using this data are that I have a large, detailed sample, with the universe of reviews for each listing. There are a few limitations. I only observe reviews that guests write for hosts. This means I only see one side of the two way reviews, meaning I cannot observe retaliation directly. This is not a large concern, since only the fear of retaliation is needed to see behavioral modifications. Further I do not observe the star rating or gender, so I impute them using text analysis methods. In this way, I rely on proxies for these values. See sections 2.4.2 and 2.7.1 for

[12] In files at later dates, it includes an indicator for whether they are a Superhost. The superhost designation was introduced in 2016 after the time period I study.

[13] Similarly, the hosts can see the picture, description, and past reviews of the guest.

[14] Alsudais (2021) also mentions potential reproducibility issues resulting from different scrapes in time. This is not a concern since InsideAirbnb provides the date of the scrape and I detail this in Appendix Table 2.A.1.

further discussion of the imputation methods and robustness respectively.

2.4.1 Sample

Due to the collection method of InsideAirbnb, my sample consists of all listings that were active as of Fall 2015 and reside in one of 10 cities/ counties in the United States. InsideAirbnb scrapes Airbnb listings for a number of cities every so often. These scraped datasets contain the universe of listings (and reviews) for hosts that are active at that point in time.[15] For example, an August 5, 2015 report contains data on all hosts/listings that are active as of August 5, 2015. A listing that was active from February 2011 to February 2017 would be included in the August 2015 report, along with the entire history of reviews from 2011 to the 'present' August 2015. On the other hand, this listing and associated reviews would not be included in the March 2017 report, since they stopped being active in February 2017. Due to this format, the August 2015 file contains information on hosts who were active before and after the 2014 simultaneous reveal policy change. However, the further the scrape is from 2014, the more likely it is that there are listings that have stopped being active and so are excluded. These excluded listings could include ones that hosted across the policy change.[16] As such, it does not make sense to use the 2019 file, for example, as only hosts that have continued to host since before 2014 can be used. Therefore, I use scrapes after, but as close to, July 10 2015. This ensures one year of post-policy data with minimal loss of the sample.[17] In other words, for hosts to be included in my sample, that person must have continued to host for a year following the policy change. Unfortunately, this is a limitation of the data. If there is differential drop-out by gender, such that those who are more affected by the policy change are more likely to drop-out, then I estimate a lower bound.

Ten cities form my sample. Between July 25, 2015 and October 15, 2015, InsideAirbnb

[15] Hosts may have multiple listings, although the majority of hosts have one listing.

[16] Only two large cities in the US were scraped before the 2014 policy change. Using these, I can compare the universe of listings pre-July 2014 to those in fall of 2015, to get an idea of drop-out rates. This analysis is ongoing but I expect to find minimal drop-out rates that are equal by host gender. There was a large increase in take-up of Airbnb after the policy change, but those observations are excluded from my analysis using the 'host since' variable.

[17] Using July 2016 files does equate to substantially fewer observations of hosts that were active across the policy change. Even so, I perform robustness checks using July 2016 (two years post-policy change) data and find similar results.

scraped Boston, Chicago, Los Angeles, Nashville, New Orleans, New York City, Portland, San Francisco, Santa Cruz County, and Washington DC.[18] Appendix Table 2.A.1 provides the exact dates for reproducibility purposes.

Using these cities, I construct a sample of all reviews for a fixed set of hosts who were active before and after the policy change. In addition to being located in one of the cities that was scraped in fall of 2015, the listing must have been active for the entire estimation window of ± 1 year and have at least one review.[19]

My analysis requires gender and review sentiment, so observations that cannot be imputed are excluded. More specifically, only reviews where I can impute gender for both the host and the guest, reviews written in English, and reviews written by a human (i.e. not an automatic cancellation review) are included. See section 2.4.2 for further discussion of imputations.

In total, I have 243,123 listing-review matched observations written between July 2013 and July 2015.

2.4.2 Data Methods

The InsideAirbnb data does not include information on gender or review ratings (such as the number of stars out of 5). Therefore I perform the following imputation methods to create proxies that allow for estimation.

I predict the gender of individuals using the python package gender-guesser made available by Pérez (2016). The program works by comparing first names to its database of more than 40,000 names from 54 countries. These countries come from the United States and all across Europe and Asia.[20] The program returns one of female, mostly female, mostly male, male, unknown, or androgynous. Androgynous means that half of the people with that name are male and half are

[18] These are the 10 cities that were scraped around one year after the policy change.

[19] This method provides a large sample but excludes those who began hosting after - or for only a short period of time before - the policy change. Since Airbnb grew in popularity for many years after its start in 2008, conditioning on being a host since before July 2013 excludes more than half of the active hosts in July 2015. Given the popularity Airbnb had already reached in 2013, along with the size of the 10 cities, there is still a substantial sample size.

[20] Central and South America, Canada, the Caribbeans, Africa, and Oceania are not represented in the database. Underlying data is from the program 'gender' by Jörg Michael, see Gecko (2007) for more details.

female; unknown means that that name was not found in the database.[21] The package takes country into account the country when assigning gender, for instance Sascha is a boy's name in Germany while Sasha is often a girl's name in the US. Since my sample is people living in US cities and I don't have nationality, I give preference to the US when predicting gender.Given the origin of names in the database, I am identifying off of individuals with common American, European, or Asian names. Individuals with Hispanic and African-sounding names are mostly excluded due to the algorithm being unable to match their name to a gender. For my analysis, I define female as {female, mostly female} and male as {male, mostly male}. I exclude couples and observations with unknown or unpredicted gender from the main analysis.[22] See Table 2.1 for examples of gender imputations.

Table 2.1: **Predicted gender examples show accuracy and relative strength for American & European names**

Host		**Guest**	
Name	**Gender**	**Name**	**Gender**
Jared & Katherine		Nathan	male
Michael	male	Cory	mostly male
Michelle	female	Kim	mostly female
Shera	unknown	Mary	mostly female
Robin	mostly female	Joe	male
SoHoFlat	unknown	Gracia	andy
Sam	mostly male	Brian	male
Chris	mostly male	Jessica	female
Tiffany	female	Bat-Enkh	unknown
Jeffrey	male	Mei-Lyn	unknown
Mike	male	Elizabeth	female
Dima	andy	Merrin	unknown
Dosco	unknown	Rafael	andy
Whitney	mostly male	Skye	mostly female
Constantine	mostly male	Brian	male
Navin	andy	LisaRoxanne	unknown
Christie	mostly female	Heather	female

[21] Some of the non-classified names are words or phrases, like "Jm", "SoHoFlat" or "New-Yorker", since hosts self-enter their Airbnb name.

[22] I exclude couples from the main estimation because gender differences in behavior from and towards couples are likely different than that of individuals. Gender perceptions are important and in opposite-sex couples both genders are present and I cannot differentiate who is the 'main' host. In same-sex couples, there may be discrimination and the presence of a second person is likely to change dynamics regardless. Note that I am more likely to exclude couples that host together since they would create the listing and their name together (e.g. Jared & Katherine). Guests on the other hand often have their own account even if they stay as a couple. For this reason, I am more accurate in identifying couples who host together than couples who stay together. However, the guest who booked the stay is more likely to be the one to write the review, so gender dynamics may still be present for guests who are couples.

The gender-guesser package has the lowest misclassification rate among commonly used methods to predict gender from first name. Santamaría & Mihaljević (2018) compare 4 established gender inference by name tools and find that gender-guesser has the lowest misclassification rate. This means gender-guesser has the smallest gender measurement error. However, they also find that it has the highest amount of non-classifications.[23] Since gender-guesser is unlikely to assign male to women and vice versa, the gender-guesser package is used as the main gender imputation method. Other ways to compute gender are used for robustness checks which are discussed in Section 2.7.1.

Of the universe of 1 million host observations in the scraped data, 12% are couples (have two first names in the host name variable), 31% are female, 31% are male, and 26% are unknown or androgynous. In the sample I have 243,123 review level observations for hosts who joined Airbnb more than one year prior to the policy change and both the gender of the host and guest are predicted. Of these observations, gender is roughly 50% women and 50% men for hosts and guests (Tables 2.2 and 2.3). Furthermore, host-guest gender pairs are distributed evenly (Table 2.4).

Table 2.2: **Equal Representation of Host Gender in Sample**

		No.	%
Female		121,600	50.02
	female	103,750	42.7
	mostly female	17,850	7.3
Male		121,523	49.98
	male	104,948	43.2
	mostly male	16,575	6.8
Total		243,123	100

Table 2.3: **Equal Representation of Guest Gender in Sample**

		No.	%
Female		126,878	52.2
	female	115,353	47.5
	mostly female	11,525	4.7
Male		116,245	47.8
	male	106,766	43.9
	mostly male	9,479	3.9
Total		243,123	100

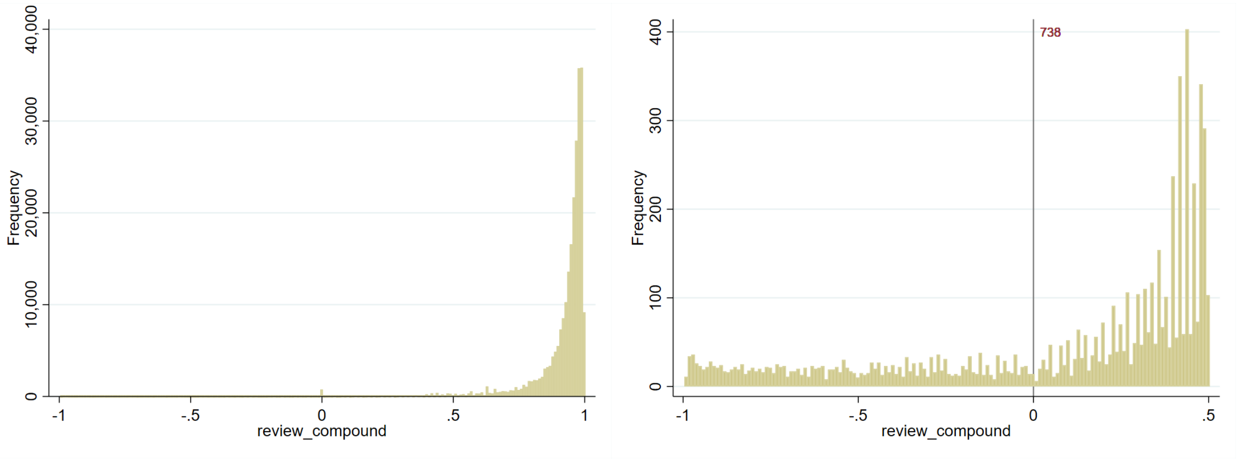
I use text analysis to assign a sentiment index to each review. I do not see star ratings, but I use the full text comments of reviews to quantify sentiment. I classify reviews using the sentiment polarity index set forth by Bird, Klein, and Loper (2009) and modified for online, social

[23] The other three commonly used packages are Gender API, Name API, and Genderize. Gender API is the overall best in the sense of combined low inaccuracies and low non-classification rates. Observations that are unknown using gender-guesser might be imputed by Gender API, but this is unlikely to change my estimates unless commonness of name significantly affects gender dynamics.

Table 2.4: **Host-guest gender pairs are distributed evenly**

		Guest	
	%	Female	Male
Host	Female	27.2	22.8
	Male	24.9	25.0

media contexts by Hutto & Gilbert (2015).[24] Using this natural language toolkit (NLTK) python package constructs a compound review score that ranges from -1 to 1. Reviews are overwhelmingly positive. Figure 2.1 shows the distribution of review scores is heavily skewed, but there is a long tail of reviews across the entire range. The average sentiment index in the pre-period is 0.906 (Table 2.5) and matches previous research.[25] For ease of interpretation, I use a normalized sentiment index multiplied 100 for the regression analysis.[26]



(a) All

(b) Lower tail (< 0.5)

Figure 2.1: **Review Sentiment Index is overwhemingly positive, with a long left tail**

In addition to the compound index, which is the main outcome variable, I use component

[24] Their book is titled Natural Language Processing with Python. It includes an open source python library and package, see `http://www.nltk.org/api/nltk.html`. I use Hutto & Gilbert's (2015) VADER Module. The NLTK python package is one of the most popular packages to conduct sentiment analysis (I.V. 2016).

[25] Iakubovskyi (2018) finds that 95% of reviews are positive or neutral, with a range of how positive. Fradkin et al (2021) finds only a small amount of negative reviews. Three-quarters of guests leave a 5 star overall rating with 48% having fives for every category. Similarly, 82% of hosts leave reviews for guests with all five stars. Mousavi & Zhao (2022) documents mostly positive reviews, with a drop in review sentiment accompanied by longer, more detailed reviews after the July 10, 2014 policy change. Zervas et al (2021) finds that 94% of reviews in 2015 (91% in 2018) were 4.5 or 5 stars.

[26] The normalized sentiment index is mean 0, standard deviation 100. I multiply by 100 to avoid excessive zeros in the regression tables.

indices to study the composition of reviews. The NLTK toolkit creates individual component indices that are ratios for the proportion of text that falls in each {positive, neutral, negative} category. In other words, the three component indices sum to 1. These indices offer another way to analyze the sentiment of reviews by looking at the fraction of the review that contains negative wording. When performing analysis with the component indices I control for review length to pick up a compositional effect conditional on length.[27] The average review has a positive compound index (Figure 2.1). This is reflected in the component indices as well, with the average review being 68% neutral, 30.6% positive, and 1.4% negative (Table 2.5). The overwhelming positivity is also evidenced by the fact that nearly two-thirds of reviews have no proportion of the review text that is negative. This means that not only is the average negative component index small, the mode is 0.

Table 2.5: **The average review has a compound index of 0.9, and is 30.6% positive words, 68% neutral, and 1.4% negative**

Index	**N**	**Mean**	**Std. Dev.**	**Min**	**Max**
Compound	243,123	0.906	0.179	-0.9966	0.9996
Positive	243,123	0.306	0.126	0	1
Neutral	243,123	0.68	0.12	0	1
Negative	243,123	0.014	0.028	0	1

NLTK's sentiment index performs quite well for reviews written in English, but not for reviews written in other languages (Table 2.6). Iakubovskyi (2018) documented the same phenomenon. He found that only a small fraction of reviews were incorrectly classified by the model,[28] and most misclassified reviews were the result of the review being in a language other than English. For this reason, I limit my sample to reviews in English.[29]

It is reasonable that reviews are mostly positive. Iakubovskyi (2018), Mousavi & Zhao

[27] It is important to control for review length when using the component indices. Otherwise we would pick up a combination of length and composition of review effect. Controlling for review length allows identification of review composition changes conditional on length.

[28] His article does not specify what "a small fraction" is.

[29] I classify the language of the review using Google Translate's language detect tool (automated through Google Sheets). In the future, I could do a robustness check where I translate reviews to English using google translate before imputing sentiment using google translate. This process is really slow and it isn't clear if the sentiment index would be as accurate, depending on the accuracy of the translation.

Table 2.6: **Imputed review sentiment is largely accurate for reviews written in English**

	Compound	**Composite index**			
Review Text	**index**	**Positive**	**Neutral**	**Negative**	**Accurate**
In sample					
I can't express enough what a wonderful experience I had. . . !	0.9965	0.369	0.606	0.024	y
WOW. Stunning is right. . . exactly what I needed. . .	0.9711	0.223	0.752	0.024	y
Host was not available. . . entered a house that was occupied. . .	0.9165	0.091	0.855	0.054	too high
Anissa was delightful and took very good care of us. . .	0.9165	0.34	0.66	0	y
Everything was great... It was easy to come in and out of the house.	0.8591	0.302	0.698	0	y
Jenny and her family are great. They were precise and helpful	0.7845	0.434	0.566	0	y
Please be careful... question integrity... not same apartment... mouse...	0.5393	0.065	0.877	0.057	too high
We had a mixed experience. . . underwhelming. . . cat urine. . .	0.3438	0.091	0.822	0.087	y
Totally painless and on one of the most walkable blocks. I'd stay again	0.081	0.121	0.774	0.105	too low
Kristen was helpful. . . problem, difficult to get ahold of... not clean	0.0821	0.14	0.714	0.146	y
It was convenient.	0	0	1	0	y
The place is in a good location. I had problems checking in. . .	-0.0459	0.124	0.767	0.109	y
Ok stay but a few problems encountered. . . wifi signal was weak. . .	-0.6057	0.131	0.701	0.168	y
Excluded					
Reservation was canceled 28 days before arrival... automated posting.	0	0	1	0	NA
Muy buena ubicacion! No mesa. . . todos tuvimos una buena estadia!!	-0.5216	0	0.914	0.086	too low

(2022), Zervas et al (2021), and Fradkin et al (2021) all find this pattern.[30] Intuitively, some of the neutral and less positive reviews are subtle hints at a stay being a sub par experience, without having to write an outwardly negative review.[31] Reviews with a sentiment index within the interquartile range are often short with no positive word (e.g. close to convention center) or muted (e.g. fine, functional). Neutral reviews are often shorter, and without emotion-filled positive or negative words. Only extremely negative reviews use unambiguously negative language. Therefore sentiment indices need not be below zero to represent less satisfying stays. In other words, even though the distribution of the sentiment index is heavily skewed, the dispersion and composition are informative (see Figure 2.1b and Table 2.6).

[30] Iakubovskyi (2018) uses the same NLTK sentiment index, while Fradkin et al (2021) and Zervas et al (2021) use the number of stars.

[31] This idea is expressed in multiple articles and blogs, including Porges (2014).

2.4.3 Data Visualization

Overall, I have a regression sample of 243,123 review-level observations. These matched listing-review observations are reviews that were posted between July 2013 and July 2015 that meet a few other requirements. Namely, that the host was active for the entire time period, the host was reviewed more than once, the listing is in one of the 10 cities, the gender of the host and guest can be imputed, the review is written in English, and the review is not an automatic cancellation review.

Table 2.7 describes the data at a glance. Men and women are equally likely in our dataset, for both guests and hosts. In these samples, there are roughly 14,400 unique listings, 10,750 unique hosts, and 220,000 unique reviewers.[32] The average listing has 21 reviews between July 2013 and 2015. The normalized and scaled review index has mean zero and standard deviation of 100.

Table 2.7: **Sample has 14,000 unique listings with an average of 21 reviews each**

Unique Listings	N	14,372
Unique Hosts	N	10,742
Unique Guests	N	219,567
# reviews per listing	Mean	21
# reviews per listing	p5	1
# reviews per listing	p95	77
Review index (z)	Mean	0.01
Review index (z)	Std. Dev.	99.98
Total	N	243,123

Airbnb experienced a general trend of increasing popularity during my estimation window, along with busy and quiet periods. Figure 2.2 illustrates that the number of reviews per week has been increasing since 2013 and experiences Summer and New Years Eve peaks.

The 250,000 observations included in the sample are representative of all Airbnb users as of 2015. Appendix Table 2.A.2 shows that the composition of observable characteristics of the sample resemble those of excluded observations on location and listing type. The only noticeable difference is in the range of review dates, since the sample focuses on reviews in the window of one year before and after the policy change.

[32] The number of unique hosts is less than that of unique listings because hosts can rent out multiple locations.

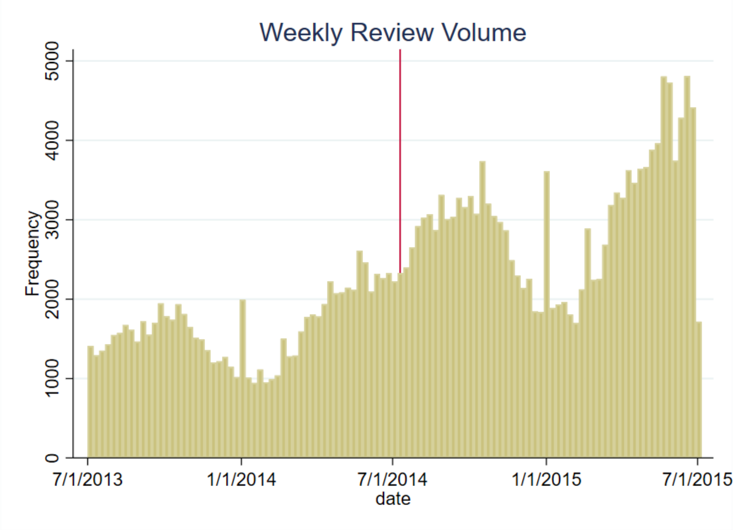


Figure 2.2: **Airbnb's increase in popularity and seasonality is reflected in the volume of reviews**

Hosts in my sample started hosting on Airbnb prior to July 2013 ("early"-joining hosts) have a similar gender composition to hosts who joined Airbnb later. Appendix Table 2.A.3 shows that the gender distribution is similar between early- and late-joining hosts. The number of reviews per month on average is larger for those who have been hosting for a shorter period of time. This is mostly mechanical and is expected. Any month where someone went on vacation and so did not rent is counted as a zero in the average. Additionally, the policy change does not seem to induce different types of hosts to join Airbnb after July 2014.[33]

Guests who leave reviews before and after the policy change are balanced across available measures. While the policy change may encourage review writing overall, guests who review one year before and one year after July 2014 are alike in terms of gender and types of stays reviewed (Appendix Table 2.A.4). If the type of person who reviews hosts changed as a result of the simultaneous reveal process, this wouldn't necessarily bias my results if it was a consequence of the policy change, but it would change interpretation. It appears that the composition of reviewers is relatively stable according to the observable characteristics.

With this data, I turn to the theoretical and empirical analysis.

[33] All of Table 2.A.3 is conditional on being an active Airbnb host as of Fall 2015.

2.5 Model

Hosts and guests face a decision problem of whether to write a review and if so, how honest to be. This 2-person decision game is one of cost-benefit analysis. Benefits of writing review for others includes expressive joy and other regarding preferences. Potential retaliation is a cost of writing a negative review. There is also a small time cost of writing a review. Utility is also affected by what the other person writes about you; everyone prefers to receive positive reviews. Writing a positive review could increase the likelihood of receiving a positive review in return, creating an added incentive to give glowing reviews.

The policy change is modelled as a change in the format of the decision game - from sequential to simultaneous. Retaliation and reciprocity appear as probabilities, that may or may not condition on your own action. Gender plays a role through differing expected probabilities of behavior. Before July 10, 2014 the game is sequential, resulting in first and second movers. The second mover sees what the first mover chose, and can react accordingly. The first mover anticipates this and makes their decision based on expected probabilities. The game becomes simultaneous on July 10, 2014. This means there is no ability for one person's action to influence the other.

The crux of the model is that potential retaliation distorts truth-telling. If your choice of review-writing did not influence the other person, you would review honestly. You get joy from sharing your experience, and this joy is the greatest when what you share matches the experience you had. If it was a good stay, you get utility from helping out the host. These other regarding preferences also work in the other direction; if it was a bad stay, you get utility from helping out potential future guests of that host. However, if what you write influences what is written about you, you will consistently write reviews that are more positive than the actual experience warrants. The reason is two-fold. The second mover will punish non-positive reviews and this detrimental effect outweighs the private benefits. The second mover may also reward positive reviews with a reciprocally positive review.

Let i and j represent the two players. Let person i is the guest. For now, the guest will review first. This simplification allows the model to match the empirical analysis of focusing on reviews written for hosts, and is used since guests often review first (Astaire 2017; Anders 2017).[34]

The quality of the stay for the guest is θ (i.e. the quality the host provides) and the quality of the stay for the host is ϕ (i.e. how well the guest behaved). An honest review is one where the review written matches the true quality. A guest (host) writes an honest review if the sentiment of the review equals the true quality, θ (ϕ). Quality is a continuum, but for simplicity I let $\{\theta, \phi\}$ take on one of three values: bad or negative ($-$), so-so or neutral ($\sim$), and good or positive ($+$).

Naturally, review sentiment is also a continuum but for ease has the same three categories. Each individual chooses to write a negative review ($-$), neutral review ($\sim$), or positive review ($+$).[35] I denote this set of actions by $A = \{-, \sim, +\}$. The guest (player i)'s review decision is x while the host (player j)'s decision is y.

The general decision problem is the following.

$$\begin{aligned} i\text{ (guest) : }\max_x U_i\Big(\ b(x,\theta),\ y(x,\theta,\phi),\ E_i(pr(y|x,\theta,\phi,g_i,g_j))\Big) \\ j\text{ (host) : }\max_y U_j\Big(b(y,\phi),\ x(y,\theta,\phi),\ E_j(pr(x|y,\theta,\phi,g_i,g_j))\Big) \end{aligned} \tag{2.1}$$

where $b(\cdot)$ are private net benefits of writing a review (expressive joy, other regarding preferences, time cost), x and y are the reviews written by or for you, and $E(pr(\cdot))$ are the expected probabilities of the other party's action conditional on your own review choice. Gender is denoted by g.

I assume that private net benefits of writing a review $b(\cdot)$ are maximized when the review sentiment (x, y) matches the true experience (θ, ϕ) (Assumption 1). This claim is reasonable based on previous literature on review writing behavior (Ledesma 2020; Fradkin et al 2021; Bad reviews 2019; Why Would They Write That 2018). See Appendix 2.B for more details.

Assumption 1. *Private net benefits are maximized when review sentiment matches true experience.*

I also assert that the sentiment of the review written by the second mover is correlated with

[34] Having guests review first is a simplification that could be loosened to allow for the host to review first. The implication is that my empirical results would be attenuated. See Appendix 2.B for further discussion.

[35] Individuals can also choose not to write a review. For simplicity and to focus on the retaliation and reciprocity mechanisms, I focus on the choice of sentiment conditional on choosing to write a review.

the sentiment of the first person's review to some degree. When reviews become simultaneous reveal, the reviews become less correlated with each other (Assumption 2). This was proven in Fradkin et al (2021). This idea is explicated in Appendix 2.B.

Assumption 2. *Review sentiment of guest and host reviews are more correlated when reviews are sequential reveal.*

We can breakdown the general decision problem (Equation 2.1) into 2 major components of reviews: utility from giving and utility from receiving reviews. I will refer to the first term as utility from direct action and the second as 'indirect' utility from others' actions.

Sequential Game, First Mover

Prior to July 2014, the first mover must account for how their action impacts the review they receive, which is modeled through expected utility with conditional probabilities.

$$\max_{x\in A}\left(U_i\left(b(x,\theta)\right)\right)+\left(\sum_{a\in A}E_i(pr(y=a|x,\theta,\phi,g_j))U_i(y=a)\right) \tag{2.2}$$

By assumption 1, the first term is maximized by writing a truthful review. However, the indirect utility does not incentivize truthtelling unless the truth is positive ($\theta = +$). The guest wants to receive a positive review; $U_i(y = a)$ is maximized when $y = (+)$. By assumption 2, the probability the host leaves a positive review is maximized when the guest writes a positive review. More generally, the host revises their review towards the sentiment of that written by the guest. In other words, as x decreases,[36] indirect utility from writing a review also decreases.

If the true experience is positive ($\theta = +$), the guest faces no tradeoff between direct and indirect utility, and will write a positive review. When $\theta \neq +$, the guest's decision depends on the magnitude of direct utility of writing a review and indirect utility of how your review impacts the one written for you. It is very reasonable that people care more about the reviews written about them since they stay with them forever and impact their future interactions. The direct utility from reviewing is more temporary and may mimic a warm glow feeling.

[36] Meaning as x moves from $+$ to $\sim$ to $-$.

Therefore, it is the first mover's best response to adjust their review up towards the positive end of the sentiment scale. This behavior is a result of the guest's expected probabilities of how the host will react to their review. In other words, this upward revising behavior is due to the possibility of detrimental retaliation and/or advantageous reciprocity. Note, however, that if the stay is horrendous, the expressive joy and other regarding preferences may outweigh the consequences of receiving a retaliatory bad review. For this reason we still expect to see some very negative reviews before the policy change.

Sequential Game, Second Mover

The second mover's decision in the sequential game does not involve probabilities since they observe the first mover's action, denoted by $\hat{a}$.

$$\max_{y \in A} \left(U_j \left(b(y, \phi, x = \hat{a}) \right) \right) + \left(U_j (x = \hat{a}) \right) \tag{2.3}$$

The second term is now determined. The review this individual writes has no impact on the review they receive.

The first term now also depends on the review written about them. In other words, expressive joy is the utility from expressing their experience with the guest's stay in addition to expressing feelings about the review through reciprocity or retaliation.

If the guest wrote a negative review $x = \hat{a} = (-)$, the host can gain by responding with a negative review regardless of the true experience θ,ϕ. Alternatively, if the stay was not great for the guest ($\theta = \sim$), but they leave a positive review, the host can express their appreciation by also writing a positive review. This is an example of reciprocity. This model is for a one-shot game, but the intuition can be extended over time. Hosts want to encourage positive review-giving behavior through reciprocity and retaliation, both of which ensure that guest's nudge review sentiment up.

Therefore the host's best response is to review with sentiment that is between being truthful of their experience and revising towards the sentiment of the guest. This is a type of tit-for-tat strategy that creates correlation in review sentiment.[37]

Simultaneous Game

Now let's move to the simultaneous reveal game that represents the post-policy change Airbnb setting. Regardless of who reviews first, the contents are blind until both have reviewed the other. People still care about the review given to them, but this cannot be influenced by their review decision. Therefore the expectation over the probability distribution does not condition on one's own choice.

$$\max_{x\in A}\left(U_i\left(b(x,\theta)\right)\right)+\left(\sum_{a\in A}E_i(pr(y=a|\theta,\phi,g_j))U_i(y=a)\right) \tag{2.4}$$

Since action does not cause reaction, there is no opposite pull on the choice of review sentiment. The guest chooses sentiment to maximize $b(\cdot)$, which is when sentiment matches the true experience $x=\theta$. Therefore reviews are utility maximizing when telling the truth. The best response (BR) is to be honest.

This is also true for hosts, whose decision problem mirrors the guest's.

$$\max_{y\in A}\left(U_j\left(b(y,\theta)\right)\right)+\left(\sum_{a\in A}E_j(pr(x=a|\theta,\phi,g_i))U_j(x=a)\right) \tag{2.5}$$

When reviews are blind, it is incentive compatible for everyone to write honest reviews. Therefore, the model predicts that reviews will become more negative on average and shift to the left following the policy change. This idea is also explicated in Porges (2014).

Model Prediction 1. *Reviews become more negative and more truthful on average.*

Model Prediction 2. *The distribution of review sentiment shifts to left.*

Additionally, some people may review more now than they used to. While I have not made assumptions on the exact utility function, the counteracting motivations for writing non-positive

[37] This intuition is confirmed by the active presence of Airbnb hosts online (What are some tips that new hosts on AirBnB can use n.b., Airhosts Forum n.b.) and previous literature (Fradkin et al 2021; Mousavi & Zhao 2022). See Appendix 2.B for more.

honest reviews could have induced people to not write a review in the sequential setting.

So far we have modeled how guests revise up and hosts revise towards the guest review. We now understand this component of the chronic over-positivity of reviews. We also know that removing the retaliation mechanisms induces honesty in reviews.

Gender

Gender of the host and guest play a role. In the model, gender interacts with the expected probabilities of review sentiment through differential likelihoods of retaliation and reciprocity. This interaction is only present in the sequential, pre-policy change environment. In the post-policy simultaneous game, there is no strategic interaction.

If men are more likely to retaliate, then guests would have heightened fear of retaliation from male hosts. In the model, this is represented by $E_i(pr(y = -|x = -, g_j = \text{male})) > E_i(pr(y = -|x = -, g_j = \text{female}))$.[38] In other words, when the guest is a first mover, they will be more deterred from writing negative reviews by potential retaliation from male hosts. This means that guests will be either more likely to revise sentiment upwards when reviewing male hosts or modify the sentiment more dramatically. Both of these lead to the prediction that the removal of the policy change will result in a larger negative shift for male hosts than female hosts. Since previous literature has found that men are more likely to retaliate,[39] this is the third model prediction.

If men are more likely to reciprocate positive reviews, then guests would have a larger incentive to revise reviews upwards towards male hosts. In the model, this is represented by $E_i(pr(y = +|x = +, g_j = male)) > E_i(pr(y = +|x = +, g_j = female))$. This would similarly lead to the prediction of the policy change resulting in a larger leftward shift in review sentiment for male hosts. As such, finding evidence to support Model Hypothesis 3 does not indicate whether the underlying mechanism is retaliation, reciprocity, or both.

Following this line of thinking, if women are expected to be more likely to reciprocate, this could attenuate the differential effect of the policy change by gender. Previous literature has

[38] Excluding θ, ϕ for brevity.

[39] See Dehdari et al 2019.

found that who is more likely to reciprocate is sensitive to the context. Chaudhuri & Gangadhuran (2003) and others find that women have higher levels of reciprocity while Dittrich (2015) finds that men are more likely to engage in reciprocity. Age plays a role, such that older men are more likely to reciprocate than younger men. Garbarino & Slonim (2009) documented different gender-reciprocity interactions across a range of contexts.

Model Hypothesis 3. *Larger negative (leftward) shift for male hosts' reviews. The effect will be due to the combination of male retaliatory and reciprocal behavior being more likely than the combination of female retaliatory and reciprocal behavior.*

If guests who are women are retaliated against more or are more pessimistic about the probability of retaliation, then they will be more deterred by retaliation. Actual or perceived retaliation will be more common, and women will augment their reviews towards the positive end of the scale to a greater extent. Mathematically, $E_i(pr(y = -|x = -, g_i = \text{male})) < E_i(pr(y = -|x = -, g_i = \text{female}))$ where g_i refers to their own gender. If this is true, then female guests will react to the policy change by increasing the amount of negative reviews given to a larger degree than male guests. Previous literature has mixed thoughts on gender differences in modifying behavior when retaliation is possible, but we can test this hypothesis.[40]

Model Hypothesis 4. *Larger shift towards negative reviews given by female guests, due to being more deterred by (fear of or actual) retaliation or less optimistic of reciprocity.*

Before turning to the empirical methodology, it is important to discuss the model limitation in identifying other things that could be a function of gender.

The utility obtained from expressing oneself, helping out others, or loss utility could vary by gender. If this is the case, then even when retaliation is possible, women have more incentives towards truthtelling. Due to women not modifying their reviews upwards as much, the reviews they write would not be as impacted by the policy change. In this way, this explanation competes with the expected probabilities of being retaliated against for explaining review-giving behavior differences by gender. Additionally, the utility loss experienced from a bad review could be corre-

[40] See Liyanarachchi & Adler (2011), Fatoki (2013), and Rehg et al (2008).

lated with gender. If so, retaliation deters negative review-giving behavior through 'indirect' utility $U_i(y = -)$ as well.

Overall, I can identify the net shift in review sentiment by gender. There are certain mechanisms that I cannot decompose. However, any differential impact by gender is still tied to retaliation and reciprocity opportunities. The net effect determines reviews and any gender advantage through the design of review systems.

2.6 Empirical Methodology

I exploit the natural experiment to identify differential effects of removing the retaliation and reciprocity channels by gender. I utilize listing fixed effects with the panel data to identify gender differences on average and across the distribution. There is no control group; men are also affected by the policy change, so it is an analysis of the differential impact of treatment. The majority of hosts rent one listing, so the listing fixed effects can be thought of as host fixed effects.

I begin with analysis on the compound (overall) sentiment index of reviews, then turn to compositional analysis using component positive, neutral, and negative indices. Average effects are identified using panel with FE methods and averages across time are calculated using an event study design. Logit and quantile regressions identify distributional shifts. Reviews for each listing are likely to be correlated since they are for the same place and host, so standard errors are clustered by listing.

The advantage of the host fixed effects (FE) specification is that it accounts for unobserved heterogeneity at the host level, and identifies effects of the policy change for each host. One potential downside is that we are identified off hosts with reviews before and after the policy change. Two-thirds of the hosts in my sample have fewer than 20 reviews and one-third have 5 or fewer reviews. This means that only a fraction of hosts have enough reviews on which the estimate is highly powered. For this reason, I also perform OLS regressions with additional controls for time-invariant factors like city and property type.[41] The results are almost identical, implying that

[41] With listing fixed effects, all variables that do not change across time for an individual host are collinear. City,

the host FE method is robust.

Equation (2.6) shows the main FE specification for hosts.

$$ReviewSentimentIndex_{l,t} = \alpha_l + \beta_2 Post_t + \beta_3 (FemaleHost_l \times Post_t) \ + \tau Time_t + \epsilon_{l,t} \tag{2.6}$$

where l represents the listing and t represents time. Listing fixed effects are represented by α_l. Female host $= 1$ if the host's name is imputed as a common woman's name and 0 otherwise. Reviews are observed at a daily frequency, while time fixed effects are defined at the month level. Post $= 1$ if the review is written on or after July 1, 2014, and 0 otherwise. Recall that the policy change took place on July 10th, 2014, so July is partially treated.[42] I define post using the beginning of the month so that the identification does not rely on the assumption of no mid-month effects, although results are almost identical using different cutoffs or excluding the entire month of July.[43] The coefficient β_2 gives the effect of removing the possibility for retaliation on male hosts, while β_3 gives the differential effect of removing retaliation for female hosts relative to male hosts.

I run separate regressions by gender for hosts and guests. Recall that I only observe the reviews that guests write and hosts receive, so these regressions analyze the same review data. Separating the regressions eases interpretation of guest review-giving behavior and host review-receiving experience. The same specification as Equation 2.6 is used, with Female Host replaced with Female Guest. For the guest regressions, β_2 gives the change in review-giving behavior for male guests and β_3 quantifies the differential effect of removing retaliation and reciprocity on female guests' review-giving behavior.

property time, and room type are controlled for through the listing fixed effects. I do not control for price since I only observe the price at the time the data was scraped (around July 2015). I do not want to control for factors ex-post as these prices could have changed as a result of the review policy change.

[42] Reviews are subject to the new or old rules depending on check out date. I observe the day the review was written, which is 0 to 30 days after the checkout date. A review written and posted on July 9th must be from a stay with a checkout date before July 9th and so is subject to the old review rules. However, a review posted on July 20th could be from a stay where the guest checked out on July 8th and is subject to the old rules, or from a stay with a checkout date of July 16th, subject to the new rules. Because I only see when the review is posted and not the checkout date, there is some ambiguity for reviews written in July.

[43] Results defining post as on or after July 10, 2014, July 1, 2014, or August 1, 2014 are nearly identical (see Appendix Tables 2.A.8 and 2.A.9). Results excluding the month of July 2014 also yield similar results (Columns 4).

Estimating gender differential effects with panel data methods relies on two identifying assumptions. The first is that nothing except the policy change affected men and women differentially simultaneously that isn't captured by the control variables. Since no other Airbnb policies occurred at this time, this is a reasonable assumption. The second is that the composition of individuals providing reviews does not change. If the probability of reviewing changes differentially by gender as a result of the policy change, then I identify a combination of effect per review and additional reviews. As discussed previously, if there is different drop out rates for hosts of each gender of hosts as a result of the policy change then I identify a lower bound.

Any shift in review sentiment can be interpreted as a result of the removal of retaliation and reciprocity. The only thing that changes is the elimination of the ability to react to the review that was written first. This change incentivizes honesty; regardless of the experience people need not consider the other party in what they write. If the experience was negative, the review can now be negative as there is no possibility for repercussions. Before, people may have opted to not write a review or just say that the stay was good. If the experience was neutral, the review can equally be neutral instead of inflating the review to be positive. If the experience was positive the review can be positive but need not be exaggerated. If reviews become more negative, as previous literature has found, the possibility of retaliation and reciprocity constrained truthful review-giving behavior. Similar logic follows for gender differentials. If guests who are women shift their review-giving behavior to a larger magnitude than men, then women were more constrained by retaliation and reciprocity incentives. If hosts who are men experience a larger drop in review sentiment, this implies that guests were more constrained when writing reviews for men by fear of retaliation and likelihood of reciprocity. Guests either believed that men retaliate more or were more likely to give reciprocally positive reviews and reacted accordingly.

This interpretation assumes that hosts do not change their hosting behavior as a result of the policy. However, it is possible that hosts know that they cannot use retaliation to hinder honest reviews, and so need to enhance their experience. For example, a host may know the kitchen is not very clean but that people won't include that in their review. Now that reviews are honest, they may begin to provide a higher level of cleanliness. If this is the case, then reviews do not become as negative as they otherwise would. In other words, the effect of retaliation alone is larger than what I would identify. For this additional reason, my findings are attenuated and I consider them underestimates.

To understand the timing of effects I estimate a dynamic event study specification.

$$Y_{l,e} = \alpha_l + \sum_{e=-E}^{E} \delta_e 1\{T = e\} + \sum_{e=-E}^{E} \theta_e 1\{T = e\} \times Female\_Host_l + \epsilon_{l,e} \tag{2.7}$$

where T is the number of months since July 2014. E is also in event study time, equal to 12 months. Coefficients δ_e and θ_e are of interest for the change in review sentiment that month for men, and the differential effect for women, respectively. The outcome $Y_{l,e}$ continues to be the sentiment index of the review. This specification shows whether review outcomes change immediately following the policy change.

The event study framework relies on the assumption that the policy was unexpected and that no other factors changed across the July 2014 threshold. It is reasonable that the review structure change was unexpected; no news articles appear before July 10, 2014 and no articles mention any announcement in advance. Many articles have the tone of surprise and explain 'how to react to this all-of-a-sudden new feature' (Building Trust 2014; Protalinski 2014). As for other factors, Airbnb changed their logo and redesigned their website and mobile app (Baldwin 2014). It is possible that this made the website and mobile app more user friendly, inducing people to join the website. However, since I am interested in behavior across the policy change and I limit my sample to hosts using Airbnb before July 2012, this concern is diminished. As seen in Table 2.A.4, the types of guests who submit reviews before and after the policy do not differ on observable measures.

I use quantile and logistic regressions to understand impacts beyond the average. The quantile regressions provide a holistic understanding of distributional effects without requiring sentiment cutoffs.[44] Logistic regressions require cutoff decisions but enable us to look at key

[44] I use standard Quantile Regression methods. I am unable to perform Quantile Regression with FE due to my setting of a fixed T being incompatible with current methods (Koenker et al 2017, e.g. Ch. 19 by Galvao & Kato 2017;

sections that map to intuitive concepts of review sentiment.

Recall that the sentiment index is heavily skewed (Figure 2.1). The 1st percentile of the distribution has a value of 0, which is halfway along the range of -1 to 1. The 5th percentile is at 0.66. This means that 95% of the distribution is relatively positive. Further, the bottom of the distribution is sparse making the effects difficult to identify. For this reason I estimate effects at the 5th, 25th, 50th, 75th, and 95th percentiles.

The logit method identifies distributional effects as the probability of an overall "negative", "neutral" or "positive" review, where these terms align with what a human categorizing the review would use to describe it.[45] Logit regression analysis allows me to use colloquial terminology to interpret effects in a way that directly ties to outcomes. It is easier to understand how a negative review impacts future revenue than what a review at the x-th percentile does. Following this, I define cutoffs in the review index that associate most closely with what we think of as 'negative', 'neutral', and 'positive'. The associated cutoffs are 0 and 0.8; reviews are defined as negative if the sentiment index is below zero $[-1, 0)$; neutral if the index is between 0 and 0.8 $[0, 0.8)$; and positive if the index is 0.8 or above $[0.8, 1]$.[46] I control for city, month, property type, and room type in the logit regressions.

I repeat the host FE and quantile regression methods for the component indices that measure the fraction of review text that fits within a $\{$negative, neutral, positive$\}$ category.

2.7 Results

Overall I find that male hosts differentially benefit in their reviews from the ability to retaliate and reciprocate. Reviews are artificially elevated prior to July 2014 due to the incentive

Machado & Santos Silva 2019). Since the main FE and OLS regressions give similar estimates, this limitation is likely small.

[45] Note that these intuitive concepts are in line with but not exactly the same as the component indices, which are what a computer can detect as negative, neutral, or positive.

[46] These cutoffs were picked by eye. Reviews classified by the NLTK given a score of -0.001 are more negative than those classified as 0 or 0.001 in general. Similarly, reviews scored 0.8 or above are very positive while those scored 0.798 are less so. A few examples of reviews with indices around these cutoffs can be found in Appendix Figure 2.A.4. I am currently working on a randomized, automated way to select these cutoffs. I will randomly pick 30 reviews in every 0.1 index bin, assign tone manually, and see where there is a marked jump (i.e. from mostly negative to mostly neutral reviews).

structure of sequential reveal reviews. When reviews become simultaneous reveal, male hosts experience a larger drop in review sentiment. In other words, guests had been modifying their review-giving behavior upwards to a larger extent when writing reviews for male hosts. The interpretation is that guests believed men to be more likely to retaliate to negative reviews, more likely to reciprocate positive reviews, or both. This average decrease comes from both a decrease in the fraction of each review that is positive, but also an increase in the fraction that is negative. This suggests a retaliation story, such that guests were constrained in their willingness to write negative reviews prior. On the other hand, reviews towards female hosts become lower on average and less positive, but they do not become more negative. This is consistent with a reciprocity story, whereby guests write less exaggerated reviews when their review cannot prod a good review in return. The lack of an increase in negative sentiment is indicative that guests were not held back from writing bad reviews to female hosts when they could retaliate.

Reviews become less positive when retaliation is no longer possible. Both male and female hosts receive less positive reviews after the policy change, as seen by the negative coefficient on "Post" and the sum of "Post" and "Female Host * Post" remaining negative in Table 2.8. Both male and female guests give lower reviews as seen in Table 2.9. This downward shift in average compound review sentiment is expected. Once reviews are simultaneous reveal, there are no longer incentives to exclude negative or exaggerate positive aspects of the stay. People are able to express their experience more honestly. This confirms Model Hypothesis 1; reviews become less positive and more truthful on average.

Importantly, reviews towards male hosts experience a larger drop in overall review sentiment than those towards female hosts. Table 2.8 shows that male hosts reviews become 8.4% of a standard deviation less positive while female hosts reviews only become 4.8% of a standard deviation lower.[47] This means that men had reviews that were twice as artificially elevated due to guests fearing retaliation or anticipating reciprocity from them more. Since hosts use Airbnb to supplement their income, this retaliation- and reciprocity-induced gender bias is significant.

[47] Female hosts reviews change by $-8.4 + 3.6 = -4.8\%$ of a standard deviation.

Table 2.8: **Removing retaliation and reciprocity reduced sentiment of reviews, impact is twice as large for male hosts**

z_review_compound	(1)	(2)	(3)
post	-7.952***	-8.415***	
	(0.694)	(0.697)	
female host * post	3.577***	3.671***	3.619***
	(0.958)	(0.959)	(0.958)
N	242,947	242,947	242,947
adj. R^2	0.001	0.001	0.001
Listing FE	Yes	Yes	Yes
Time controls	No	Month FE	Month x year

Standard errors in parentheses
* $p < .10$, ** $p < 0.05$, *** $p < 0.01$

Counter to Model Hypothesis 4, gender differences are not evident in the review-giving stage; both male and female guests change their review-giving behavior in the same way. Table 2.9 shows that both male and female guests give reviews that are 6.5 to 7% of a standard deviation less positive; guests who are women do not change their review-giving behavior differentially. This implies that all guests boosted their reviews towards men more, either from being more afraid of retaliation from men and/or from anticipating more reciprocal reviews from men. That is, women are not more deterred by potential retaliation or more encouraged by reciprocity in this setting.

Table 2.9: **Male and female guests react similarly to removal of retaliation and reciprocity**

z_review_compound	(1)	(2)	(3)
female guest	10.84***	10.83***	10.83***
	(0.681)	(0.682)	(0.681)
post	-6.475***	-6.920***	
	(0.665)	(0.669)	
female guest * post	0.632	0.680	0.693
	(0.853)	(0.853)	(0.852)
N	242,947	242,947	242,947
adj. R^2	0.004	0.004	0.004
Listing FE	Yes	Yes	Yes
Time controls	No	Month FE	Month x year

Standard errors in parentheses
* $p < .10$, ** $p < 0.05$, *** $p < 0.01$

The coefficient on female guest in Table 2.9 shows that there is a female guest positivity premium. Guests who are women tend to write reviews with more words associated with positive sentiment used by the index. The magnitude of the positivity premium is the same before and after the policy change, so it is not a result of retaliation or reciprocity.

These results from Tables 2.8 and 2.9 are robust to the exclusion of listing fixed effects; OLS results are shown in Appendix Tables 2.A.5 and 2.A.6. OLS results suggest that female hosts may offer better stays on average. Female hosts in the pre-period receive reviews that are 1.5 to 2% of a standard deviation higher on average. Since women are not much more likely to stay with women,[48] the difference is likely not driven by the female guest positivity premium.

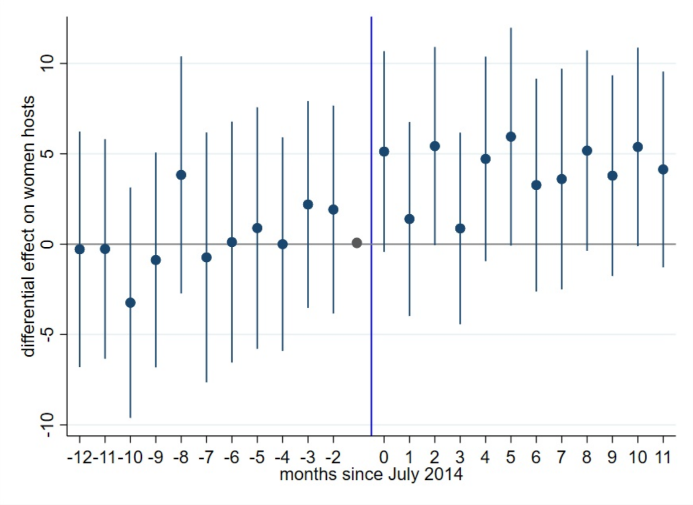
I complement this analysis with host-guest gender pairs, to analyze which types of interactions are affected the most by the policy change. I find that average results are not driven by any particular pairing. Male and female guests behave similarly towards female hosts, and male and female guests behave similarly towards male hosts. This implies that female guests are not more likely to respond to incentives in the review structure than male guests. Male host-female guest pairs and male host-male guest pairs yield same results across the policy change (See Appendix Table 2.A.7, columns 1 and 5).

Gender differences for hosts appear immediately after the policy change and are persistent. Female hosts are less affected on average relative to men in all months after July 2014 as seen in Figure 2.3. While no month is significant on its own, the jump is evident and the coefficient is in line with the differential effect coefficients found in Table 2.8.

Figure 2.4 shows how male and female guests do not significantly differ in their behavior at any point within 6 months before and after the policy change. Comparing to the previous year, guest behavior by gender fluctuates over time with no indication of differential reactions to the simultaneous reveal policy.

Beyond the average, I analyze distributional implications. As previously discussed, positive stays are likely to be reviewed positively regardless of the review system in place. On the other hand, sub par stays are likely to be reviewed as neutral-positive beforehand and more honestly

[48] Table 2.4 shows that the host-guest gender pair breakdown is nearly 25% in each category.



Note: Recall that men experience a drop in review sentiment around 8 percent of a standard deviation.

Figure 2.3: **Women hosts, relative to men, experience an immediate and sustained smaller negative impact on reviews**

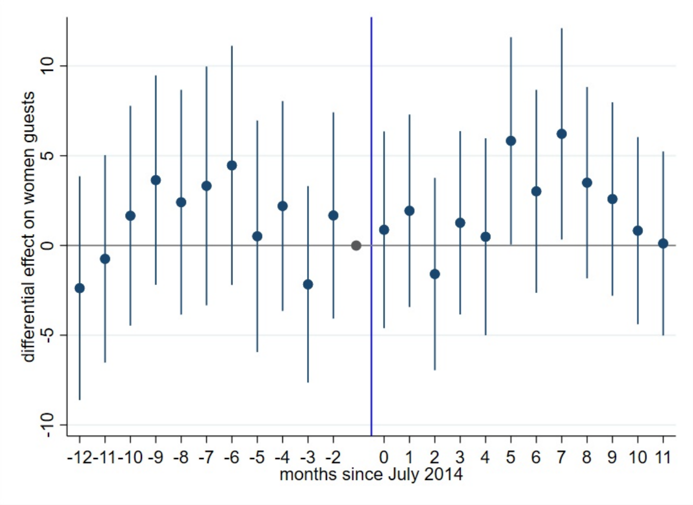
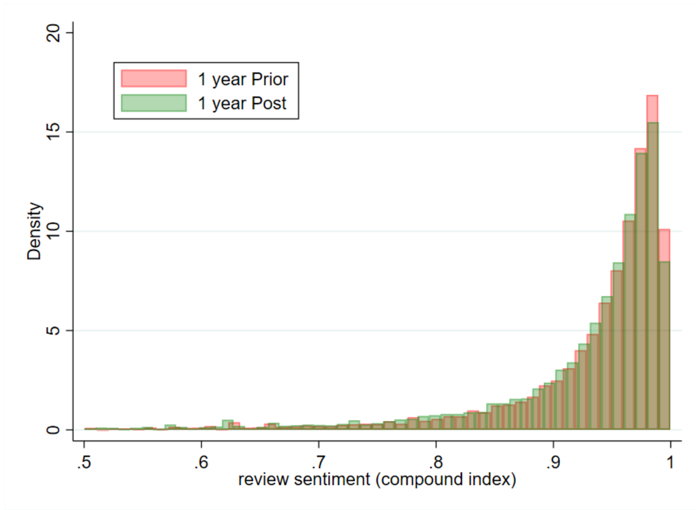


Figure 2.4: **Guests do not differentially react to policy by gender**

negatively after the removal of retaliation and reciprocity channels. Poor stays may be reviewed negatively before, but are almost surely reviewed negatively after. This intuition is confirmed by the raw data shown in Figure 2.5, with more mass at the top of the distribution under the sequential reveal policy.

Quantile regressions confirm that a leftward distributional shift does occur, meaning that Model Hypothesis 2 is true. Guests give and hosts receive more negative reviews at every quantile, with lower quantiles more affected (Figures 2.7 and 2.6, respectively). Reviews at the high end of



Note: This figure only shows reviews with sentiment above 0.5 for scale purposes.

Figure 2.5: **The policy makes reviews more honest, as evidenced by a leftward shift in the review sentiment distribution**

the distribution are nearly unaffected, as expected.

Further, male hosts are twice as affected at every quantile (Figure 2.6). This implies that people's additional fear of retaliation and/or anticipation of reciprocation from men is consistently twice as high across the range of stays.[49] There is no significant difference between male and female guests at any part of the distribution, as seen in Figure 2.7.[50]

Focusing on the base gender gap that exists before the policy change, Figure 2.8 shows that female hosts receive higher reviews and female guests give higher reviews overall. This is especially true at the lower end of the distribution. Reviews in the lowest quantile of reviews written for female hosts had higher sentiment than reviews for male hosts in that quantile (left panel). Female guests give more positive reviews in general, but the largest difference occurs for very negative reviews (right panel). Female guests giving reviews at the 5th percentile write reviews that are 60% of a standard deviation more positive than a male guest would. The female guest premium disappears at the top of the distribution; both men and women write a review with

[49] The gender gap in Figure 2.6 is shown separately in Appendix Figure 2.A.1a. We see that female hosts receive statistically higher review sentiment at all reviews between the 10th and 95th percentiles. Female hosts receiving reviews below the median are 2 to 10% of a standard deviation less affected by removing retaliation and reciprocity channels.

[50] Appendix Figure 2.A.1b graphs the gap in effects between guests who are men and women. The gender of the guest makes little to no difference on review-giving behavior.

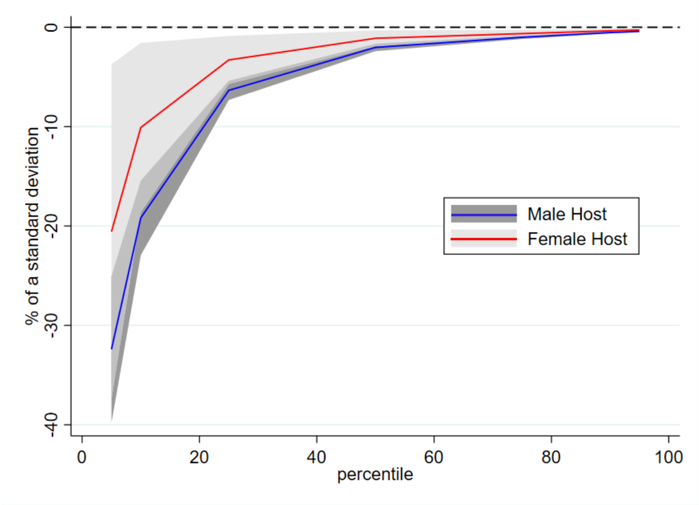


Figure 2.6: **Male hosts experience a drop in review sentiment that is twice as large at every quantile of the distribution, with lower quantiles more affected**

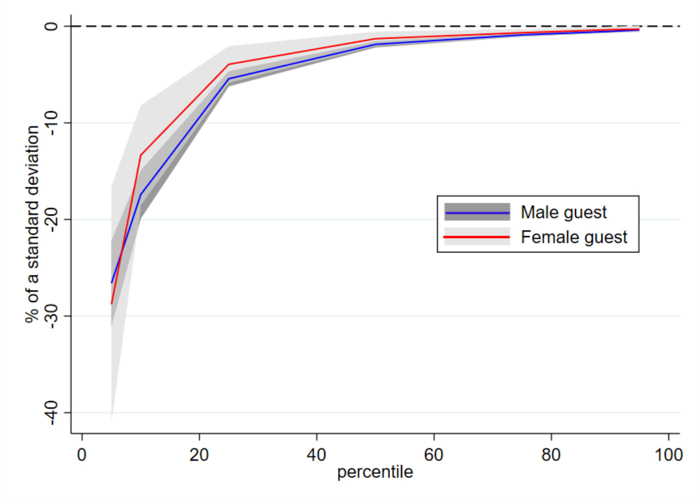
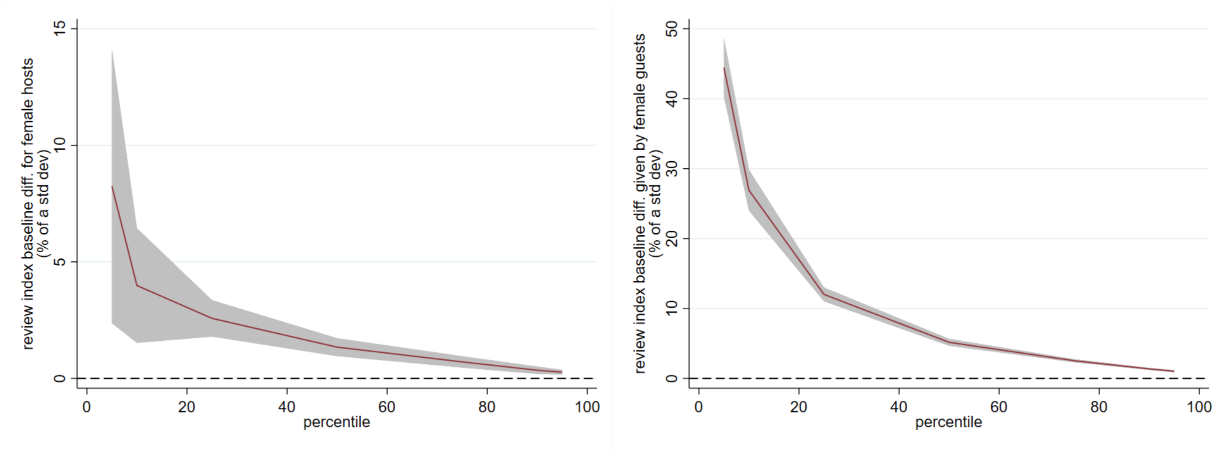


Figure 2.7: **Male and female guests react similarly at every quantile, with reviews written about the worst stays most affected**

sentiment index 0.99 for a high quality stay, such as one at the 95th percentile. This means that while male hosts had an advantage due to retaliation/reciprocity prior to July 2014, women hosts were still receiving more positive reviews, especially for reviews at the low end of the distribution.

While the quantile regressions provide heterogeneity, they do not shed light on compound index scores between -1 and 0.66. The range between the 5th and 95th quantiles represent 90% of the distribution but a smaller range of index scores from 0.66 to .992. Intuitively, the quantile regression results show the effect for very positive, positive, and neutral-positive reviews. Even



(a) **Female hosts receive more positive reviews across quantiles prior to July 2014**

(b) **Female guests give more positive reviews across quantiles prior to July 2014**

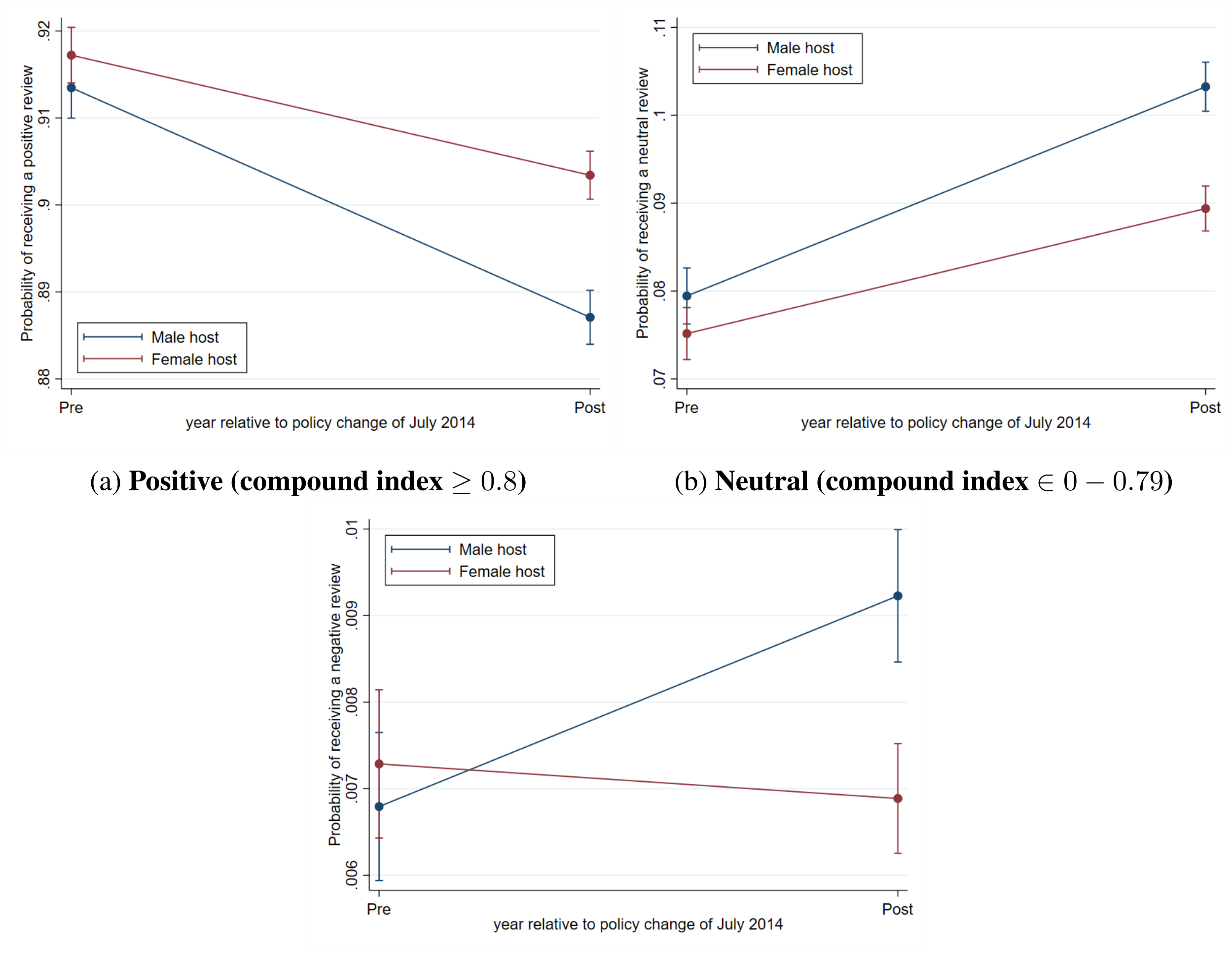
Figure 2.8: **Women write and receive more positive reviews at baseline at every quantile, especially at the low end of the review sentiment distribution**

though the 5th percentile is at low end of the distribution in percentile terms, it is quite neutral (neutral-positive), and does not accurately reflect the ‘terrible’ reviews.

To better understand retaliation’s impact on intuitively ’neutral-negative’ and ’negative’ reviews, I turn to logistic regressions. Logistic regressions allow me to define negative and neutral reviews as binary variables and assess the probability of them occurring.

Logit results continue to display that the leftward shift is stronger for male hosts, as evidenced by the larger drop in intuitively positive reviews (compound index score above 0.8), along with larger increases in neutral ($0 \leq$ index < 0.8) and negative (index score < 0) reviews (Figure 2.9). Using these definitions, reviews are overwhelmingly positive (90.1%). Less than one percent are negative (0.8%) and 9.0% are neutral. The probability of receiving a negative review does not change for female hosts, but increases from 0.7 to 0.9% for male hosts (Panel 2.9c). Given the low baseline, this is almost a 30% increase in the chance of receiving an intuitively negative review for male hosts. The probability of receiving a neutral review increases for both men and women, with the change for male hosts increasing by 2.3 percentage points and 1.5 p.p for female hosts (Panel 2.9b). Relative to the low baselines around 8%, this is a 29% increase for men and a 20% increase for women. Panel 2.9a shows that positive reviews become less likely by 2.5 p.p. for male hosts (a 2.7% decline) and by 1.5 p.p for female hosts (a 1.6% decline). Although positive reviews are still the most likely, the chance of receiving non-positive reviews increases substantially with the implementation of simultaneous reveal.

Overall the policy change induces a fraction of positive reviews to become neutral or even negative (for male hosts). These represent the fraction of stays that were not positive but were reviewed positively prior to July 2014 due to the inability to be honest. This is further evidence that retaliation and reciprocity constricted behavior of review-givers to a larger degree when reviewing men.



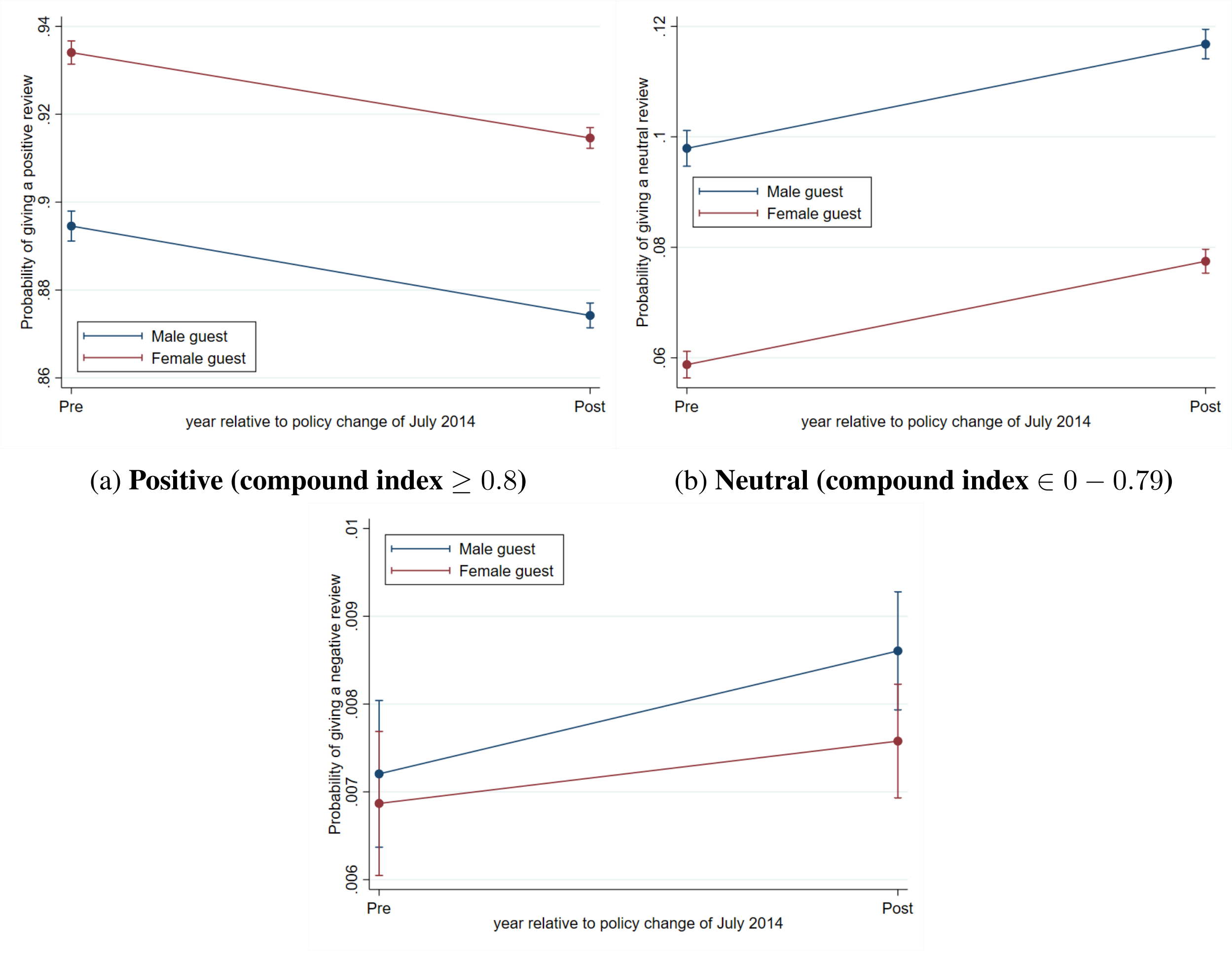
(a) **Positive (compound index** ≥ 0.8**)**

(b) **Neutral (compound index** $\in 0 - 0.79$**)**

(c) **Negative (compound index** < 0**)**
Graphs show 95% confidence intervals.

Figure 2.9: **Logit regressions show that all hosts receive less positive and more neutral reviews, and male hosts are also more likely to receive negative reviews**

Logistic regressions by guest gender show that male and female guests change their review-giving behavior in the same way. Figure 2.10 exhibits the parallel sentiment adjustment across the policy change. The female guest positivity premium can be seen by observing that female guests write more positive reviews (level shift up in panel c) and fewer neutral and negative reviews (level shift down in panels a and b).



(a) **Positive (compound index** ≥ 0.8**)**

(b) **Neutral (compound index** $\in 0 - 0.79$**)**

(c) **Negative (compound index** < 0**)**
Graphs show 95% confidence intervals.

Figure 2.10: **Logit regressions show that guests change behavior in same way regardless of gender, as evidenced by the parallel changes across July 2014**

All previously discussed analysis uses the compound index, which is a holistic measure of intensity and polarity or valence of sentiment. I now turn to the analysis on length and component indices of the fraction of review text that is $\{$negative, neutral, and positive$\}$ as categorized by the text analysis method.

Reviews become 12.5 characters (or 1 to 3 words) shorter on average after the policy change (Appendix Figure 2.A.2). The bulk of reviews are between 200 and 500 characters (roughly 30-60 to 70-130 words). The distribution of review lengths from mid-2014 to mid-2015 overlaps significantly with the distribution of the prior year. At the same time, the shorter length could be evidence of people adding fewer emphasis words to their reviews.

There are two other ways to measure the negativity of reviews using the negative component index. The extensive margin of the number of reviews with some amount of negative words (i.e. a non-zero negativity index) decreases slightly across the policy change. Before, 34.9% of reviews had some negative text, which ticks down to 33.7%. The intensive margin of the average fraction of a review that is negative increases. Conditional on having some negative text, the average fraction of text that is negative increases from 4.1 to 4.3%. These conflict, suggesting that usage of strictly negative words is roughly unchanged on average.

Composition shift analysis shows that reviews towards male hosts become more negative, while reviews towards female hosts simply become less positive. Table 2.10 reiterates that all reviews become less positive and more neutral, but only some reviews add negative text. These results are obtained by regressing individual component indices on gender, policy structure and review length. Reviews written for male hosts are revised to be less positive by -1.1%, more neutral by +1.1%, and more negative by +.05%. This increase in negative text is small but statistically significant and equates to a 7.3% increase in the fraction of the review that is negative. Note that the index picks up words that are unfavorable, so sentences such as “the bathroom was not what you want after a long flight” may not be classified as negative while “the bathroom was a filthy dump” would be. As such, the small change is still informative of the existence of negative feedback. This negativity change is not apparent for female hosts, whose reviews become 0.77 of a percent less positive, 0.78 percent more neutral, and no change to the negative component index (insignificant -0.016%).

Supplemental analysis shows that the increase in negativity is especially pronounced for

Table 2.10: **Review composition confirms that reviews become statistically significantly more negative for male hosts**

	(1) NEGATIVE	(2) NEUTRAL	(3) POSITIVE
post	0.0529***	1.07***	-1.13***
	(0.0187)	(0.0736)	(0.0757)
female host * post	-0.0691***	-0.288***	0.361***
	(0.0262)	(0.101)	(0.105)
review length (x1000)	0.0173***	0.179***	-0.196***
	(0.000212)	(0.00122)	(0.00127)
N	243,123	243,123	243,123
adj. R^2	0.035	0.206	0.227
N clusters	14372	14372	14372
Location controls	City	City	City
Time controls	Month FE	Month FE	Month FE
Host FE	Yes	Yes	Yes

Standard errors in parentheses

* $p < .10$, ** $p < 0.05$, *** $p < 0.01$

Outcome indices measure the fraction of review that fits within that category.

the worst reviews written for male hosts. Quantile regression of the component indices identifies the magnitude of the effect by the degree of negativity. Appendix Figure 2.A.3 highlights the increasing negative shift for reviews above the 75th percentile of proportion of negative words (Panel a).[51]

Performing the same component analysis by the gender of guest reiterates that gender of host is more important for review characteristics. Both male and female guests give less positive and more neutral reviews with no change in negativity. Further breaking down interactions by the 2x2 host-guest gender pair, we see that both male and female guests give reviews with more negative text to male hosts (Appendix Table 2.A.7). Neither change the proportion of negative comments to female hosts.

In addition to sentiment index outcomes, I test whether the quantity of reviews increase. As mentioned above, we expect that the possibility of retaliation induces people to write more positive

[51] Reviews with different fractions of neutral and positive words have less varied policy responses. Appendix Figure 2.A.3b shows that reviews that were a little neutral and very neutral change in a similar way. Very positive reviews are less affected than reviews that a little positive, with a more pronounced move away from positive sentiment for male hosts (Appendix Figure 2.A.3c).

reviews and maybe even refrain from writing a review at all. On the other hand, reciprocity induces people to write more positive reviews, but wouldn't inhibit review writing.

Figures 2.2 and 2.11 show the volume of reviews between July 2013 and 2015. The fitted trend line for the post period has a steeper slope than the fitted line for the pre period, but does not appear to jump on July 10, 2014 (Figure 2.11). A regression based approach and associated F test confirms that the jump in the number of reviews for the month following July 10, 2014 relative to the month prior is not significant but that the trend pivot in number of reviews per day is significantly higher after the policy change (F = 10.76, p = 0.0011).[52] This means that the policy change did not result in a significant increase in the reviewing rate, at least not for our sample of hosts who were active for the entire two year period. Previous research has found that around 70% of stays result in a review (Fradkin et al 2021; Mousavi & Zhao 2022). Fradkin and co-authors (2021) find that review rates written by guests for hosts increased by 1.7%. This is a small amount and my results are consistent with their findings.[53] An important caveat is that I only observe reviews that are written. In other words, I cannot rule out that bookings were simply increasing across the entire period.

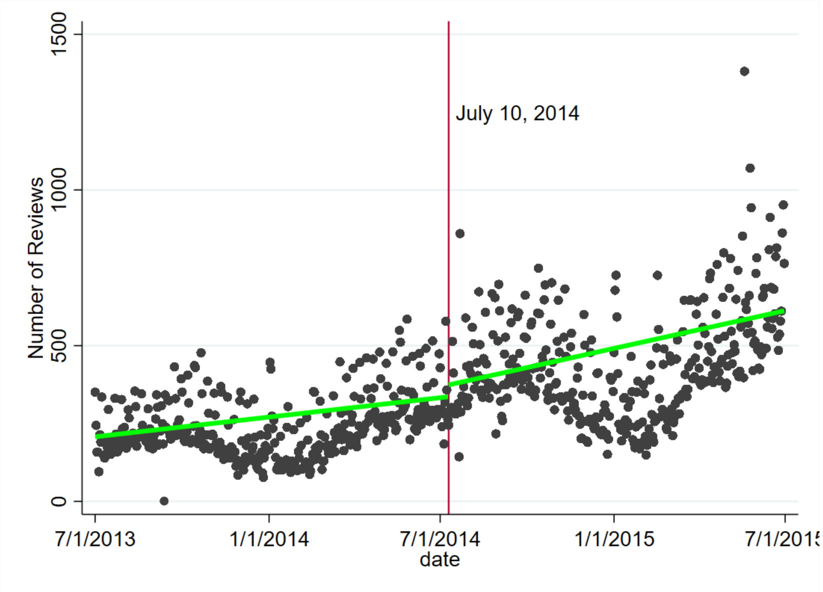


Figure 2.11: **The number of reviews written experiences a break in trend but no jump**

[52] In a regression of the number of reviews per day on post, month FE, a "pre" time trend and a "post" time trend.
[53] Fradkin et al (2021) find that review rates increased by 9.8% of reviews written by hosts for guests.

It is worth noting that the trend break may not be a result of the policy change only. Airbnb was increasing in popularity dramatically between 2014 and 2015 and this could simply reflect the increasing number of Airbnb bookings/stays. As Figure 2.11 shows, there are a few outliers in Summer of 2015, but these are not driving the results. Excluding the outliers does not eliminate the significance of trend difference.[54]

Overall, we see a consistent story whereby retaliation and reciprocity create disparate impacts by providing an artificial advantage for male hosts. Simultaneously revealing reviews lowers the positivity of comments written for hosts, and men are twice as affected. Not only do reviews become less positive, but they become more negative towards male hosts, suggesting that guests react to fear of retaliation from men. Reviews for female hosts become less positive but do not increase in negativity. This implies that guests were not so much constrained by retaliation, but reacted to anticipation of reciprocity when staying with a female host. Both male and female guests modify their review-giving behavior in the same way, which suggests that men and women are not differentially deterred by retaliation or motivated by reciprocity. Altogether, the Airbnb review policy with asynchronous reveal induced gender bias in host reviews due to differential beliefs and/or actions of retaliation and reciprocity by male hosts.

2.7.1 Robustness

Results are robust to alternate gender imputations, definitions of post, and whether the guest likely met the host.[55]

Since gender is critical to the analysis, I perform a series of robustness checks on gender imputation methods. Results hold whether I use a subset of higher-probability imputations or a different algorithm altogether.

The first gender robustness check uses names from gender-guesser that are classified as 'male' or 'female', excluding names that are 'mostly male' and 'mostly female'. This is a stricter

[54] Excluding the top and bottom 1% of reviews continues to identify a significant trend break (F = 6.21; p = 0.0129).

[55] In the future, I will conduct additional analysis to test for differential attrition of Airbnb hosts by gender following July 2014.

classification that keeps 78.4% of the observations from the main specification. These are people with names that are more commonly associated with one gender, and so who have a higher probability of being correctly gendered by the algorithm.

Following this line of thinking, I create two more stringent subsets. One consists of only individuals with the 40 most common, strongly gendered names. Hosts with the top 20 female names and top 20 male names are included, where these 20 names are the most frequent that are not associated with both genders.[56] There are 58,300 review level observations received by hosts named one of these top names (45,400 written by guests with these names). Honing in further, I perform analysis for individuals named John and Sara(h) only. These two names are the most frequent in the dataset. There are fewer observations that meet this classification (6200 for hosts and 4600 by guests), leading to large confidence intervals.

I also test the gender imputation by comparing results using a different python package that relies on separate source data. GenderComputer includes name-to-gender data for a different set of countries, including Brazil.[57] GenderComputer has been used by other economists including Bohren et al (2019).

The main and alternate gender imputation tools often agree on gender. If gender-guesser, the main indicator, predicts an individual's gender is female or male, genderComputer agrees 98.2% of the time.

Gender-guesser is the preferred gender imputation method since it is more conservative than genderComputer. Names that are unpredicted from gender-guesser are often assigned male or female by genderComputer. For example, of those classified as androgynous by gender-guesser, only 5.5% remain unclassified by genderComputer; 54% are assigned female and 41% are male. Conversely, only 30% of names that are not assigned a gender by genderComputer are assigned

[56] Alex, Chris, and Sam are common but excluded since they are nicknames for both men and women. The top 20 male names are Adam, Andrew, Brian, Daniel, David, Eric, James, Jason, Jay, Joe, John, Kevin, Mark, Matt, Matthew, Michael, Paul, Robert, Scott, and Sean. The top 20 female names are Amy, Anna, Christina, Elizabeth, Emily, Heather, Jennifer, Jessica, Karen, Kate, Laura, Lisa, Melissa, Michelle, Rachel, Sara/Sarah, Stephanie, and Susan (in alphabetical order).

[57] GenderComputer was created by Vasilescu, Capiluppi, and Serebrenik (2014). GenderComputer covers fewer countries than gender-guesser, but includes Canada, Brazil, and Australia. For a full list of sources by country, see `https://github.com/tue-mdse/genderComputer/blob/master/nameLists/nameLists.md`.

male or female by gender-guesser. These observations are in line with Santamaría & Mihaljević (2018). At the same time, very few observations are classified as male under one algorithm and female under the other; 1.07% are male in gender-guesser and female in genderComputer, and 0.99% vice versa.

The alternative algorithm genderComputer includes Brazil as one of the countries, while gender-guesser does not include any Central or South American countries. As such, it will be a little better at identifying gender for some names.

Therefore, imputed gender is largely similar but not exact between the two methods. Since the regressions identify off a different group of men and women, finding similar results is suggestive that the particular gender imputation tool is not driving results.

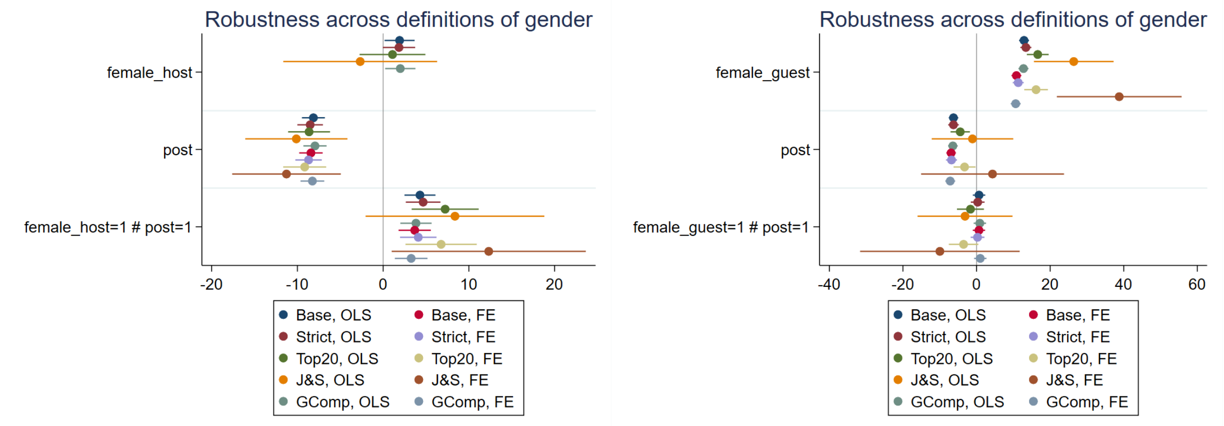
Finally, I manually assign gender using gendered language of review comments for 500 randomly-selected hosts.[58] This manual gender robustness test is currently in progress. When I finish imputing host gender using gendered pronouns used in comments of the reviews written about them, I can compare the gender imputed by gender-guesser to the true gender for a sense of accuracy in my particular sample. I will also run regressions on only this subset with nearly 100% gender accuracy and compare results to the main specification.[59]

Differential effects of the policy by host and guest gender are consistently estimated across various gender imputation methods (Figure 2.12). Male hosts are persistently twice as affected by the policy change as female hosts, while guests change their review giving behavior in the same way regardless of own gender. For full regression results see Appendix Tables 2.A.10 and 2.A.11.

Results are robust to different cutoffs for before and after the policy change. Appendix Tables 2.A.8 and 2.A.9 show that regardless of defining post as beginning on July 1 2014, July 10 2014, or August 1 2014, results are nearly identical. This is also true if I exclude July 2014 or the

[58] For each host with at least two reviews within ± 1 year of the policy change (one before and one after), I used a random number generator to assign a number between 0 and 100. I selected the top 4% which equates to 498 hosts and 21,215 host-review observations.

[59] There are multiple added benefits of this task. One is being able to identify which listings are hosted by couples, regardless of if the host name lists one or both names. This will also give me a sense of the amount of couples who are in my main dataset and assigned the gender according to the self-entered name. Another is that I can identify gender for uncommon names and for individuals who choose a nondescript host name such as “M”.



(a) **Policy's differential impact on hosts is robust to different gender imputation methods**

(b) **Lack of differential impact on guests is robust to different gender imputation methods**

Figure 2.12: **Results are robust across gender imputation methods**
Base refers to the main gender imputation of gender-guesser. Strict is a subset of Base that excludes 'mostly male' and 'mostly female'. Top 20 refers to the 20 most common male and female names. J&S stands for John and Sara(h). GComp refers to GenderComputer, the second gender imputation algorithm.

April through July 2014, the time period when some hosts may have been treated according to the pilot-like experiment discussed in Fradkin et al (2021). This robustness is crucial because there is some timing ambiguity in reviews that are treated. Which review structure the host and guest use is determined by the checkout date which I do not observe. A review that is posted on July 15, 2014 could be subject to either review policy.

Some Airbnb stays involve staying with the host while others are entire home rentals where you may or may not meet the host.[60] It is possible that the policy change interacts with gender depending on whether the guest met the host. To test this, I use room type as a proxy for likelihood of meeting the host, where room type is entire home, private room, or shared room. Appendix Tables 2.A.12 and 2.A.13 show that there does not seem to be an interaction with whether the guest met the host. Guests give more honest and negative reviews to a magnitude that is twice as large for male hosts regardless of room type.

I am in the process of analyzing differential attrition after the policy change. Intuitively, if hosts with the lowest reviews stop hosting on Airbnb, then attrition after the policy change is

[60] Sometimes you will meet them to receive the keys or if you need anything during the stay. Other times you will not meet the host (e.g. if they are out of town or have a system where you can get the key from the concierge, etc.).

expected to be from those who were most affected by the honesty-inducing policy change, i.e. male hosts. This idea is present in Cabral & Hortaçsu (2010), where they find that negative reviews on Ebay decrease sales and increase the likelihood of more negative reviews, and the likelihood of exiting. Since men are twice as affected, there may be differential drop out whereby male hosts are more likely to stop hosting on Airbnb post July 2014. To check the attrition issue, there are a few cities scraped just before and multiple times within a year after the policy change. If men differentially drop out after the policy change, then the effect I find is an underestimate of the differential impact by gender, since it analyzes hosts who continue to stay active for 1 year.[61]

2.7.2 Further Discussion

Context is key to understanding results. Honesty reduces bias in this system at the review stage, but this would not be the case under the taste-based model of discrimination. With taste-based discrimination, honesty would result in people reviewing according to their biases. The reason Airbnb reviews differ is because people have already selected into interacting with each other by the time reviews are written. To be reviewing each other, a guest has chosen the host's listing and the host approved the stay.[62] Recall from Section 2.3 that there is significant discrimination at the acceptance and booking stage, so honesty at this earlier stage may not reduce bias. This emphasizes that context is extremely important for understanding effects for various groups of people. Airbnb's review system benefits from removing retaliatory power and reciprocity incentives to move along the retaliation-honesty tradeoff because they are a group of individuals who have selected into interacting.

The fact that men have an advantage due to greater fear of retaliation from them is unsurprising given previous literature. Dehdari et al (2019) finds that men are more likely to retaliate and Liyanarachchi & Adler (2011) find that accountants over age 45 who are women are more deterred

[61] While it is interesting to know if there is differential joining of Airbnb by gender after policy change, I use a fixed set of hosts who have hosted since before the policy change so this would not bias my results.

[62] Hosts can turn on automatic booking (Instant Book) which began in 2010 (The Airbnb Story n.b.). For this reason, some hosts may not be choosing the guest (beyond certain requirements). Guests are still choosing the host, so reviews that I observe (written by the guest for the host) can still be interpreted as being selected into the interaction.

by the threat of retaliation. Additionally, women are perceived and treated differently than men in many situations. There is no reason that behavior around retaliation would not differ by gender. If men are believed to be more likely to retaliate in Airbnb reviews, then avoidance behavior is rational. Men may or may not be more likely to retaliate but only beliefs are needed to see this behavior. People assume men will be more likely to retaliate and react accordingly.

Women's reviews are higher on average both before and after the policy change. This may be because hosts who are women offer higher quality stays on average. It may also result from the female guest positivity premium since female hosts in my sample receive slightly more reviews from female guests than male guests (Table 2.4).

My findings may be underestimates if honesty of reviews induces better quality stays. Hosts may realize that reviews will become more honest and choose to improve aspects of the stay. For example, a host may decide to clean the apartment more often, knowing that they were 'getting away with' not cleaning as often since guests feared retaliation. If this phenomenon took place more often for male hosts, I am estimating a smaller gender differential than the true direct effect.

Although I do not observe reviews that guests receive, I can extrapolate my results under some assumptions. If we assume that the gender pattern for reviews that hosts receive is maintained for the reviews that guests receive,[63] then male guests also had artificially elevated reviews when retaliation was possible. This means that retaliation could have led to unfair advantages for male guests in terms of probability of being accepted by hosts. I also look forward to future research that analyzes retaliation and reciprocity dynamics by opposite-sex and same-sex couples.

2.8 Conclusion

In this paper I exploit the natural experiment of the unexpected review process change in Airbnb reviews to study gender differences in retaliation and reciprocity.

Overall, allowing retaliation distorts reviews upwards. Not only is this an issue because trust relies on reviews signaling quality, but the misrepresentation is exaggerated for men. Re-

[63] Recall that I do not have reviews written for guests.

taliation and reciprocity induced a gender bias in reviews due to differential fear of retaliation from male hosts. Removing the ability to retaliate creates incentives for review writers to be honest. When the reviews become more honest, this offsets the male advantage resulting from fear of retaliation and anticipation of reciprocity. Both male and female guests modify review giving behavior to limit the potential of retaliation and/or maximize positive reciprocity in the same way.

Retaliation and reciprocity falsely exaggerated the quality of male hosts' properties relative to women's, and hence, compounded income inequality. Reviews create one's reputation and lead to future bookings on the platform. This issue matters because the majority of hosts use their earnings to supplement their income.

These findings are relevant to inform review policy for Airbnb, but also other two-way or peer-to-peer review systems. Review systems that allow reciprocity and retaliation are inherently not gender-neutral. More broadly, gender differences in evaluations extend to peer-to-peer systems.

This paper is one building block of many towards reducing disparities in a myriad of evaluation systems. Future work in this area will unveil bias for other groups of people (by race/ethnicity, disability, religion, sexual orientation, etc) induced by retaliation and other review system characteristics. A fruitful area of research is studying review policy features that reduce discrimination at the earlier stage of host-guest interactions that manifests as acceptance and booking rate discrepancies.

Chapter 2 is currently being prepared for submission for publication of the material. The dissertation author, Amanda Bonheur, was the primary investigator and author of this material.

2.9 References

Ahuja, Rishi, and Ronan C. Lyons (2017). The Silent Treatment: LGBT Discrimination in the Sharing Economy. *Trinity Economics Papers, Trinity College Dublin, Department of Economics*, tep1917. https://ideas.repec.org/p/tcd/tcduee/tep1917.html

Airbnb's Growing Community of 60+ Women Hosts (2016). *Airbnb Citizen*. Available at https://www.airbnbcitizen.com/wp-content/uploads/2016/03/Airbnb_60_Plus_Women_Report.pdf

Airhosts Forum (n.b.) https://airhostsforum.com/

Alsudais, Abdulkareem (2021). Incorrect Data in the Widely Used Inside Airbnb Dataset. *Decision Support Systems, 141*. DOI: 10.1016/j.dss.2020.113453.

Ameri, Mason, Sean Edmund Rogers, Lisa Schur, Dandouglas Kruse (2020). No Room at the Inn? Disability Access in the New Sharing Economy. *Academy of Management Discoveries, Vol. 6*, No. 2. https://doi.org/10.5465/amd.2018.0054

Anders, William (2017). Answer by William Anders to "On Airbnb, should the host or the guest review first? Why?" *Quora*. Accessed August 16 2021. Available at https://www.quora.com/On-Airbnb-should-the-host-or-the-guest-review-first-Why

Astaire (2017). Astaire's answer to "To review or not to review (when you know the experience was mutually disagreeable and the guest will surely leave a bad review in return) - POLL". *Airhosts Forum*. Available at https://airhostsforum.com/t/to-review-or-not-to-review-when-you-know-the-experience-was-mutually-disagreeable-and-the-guest-will-surely-leave-a-bad-review-in-return-poll/16627

Average Airbnb Occupancy Rates By City (n.b.). *AllTheRooms*. Accessed August 2022, https://www.alltherooms.com/analytics/average-airbnb-occupancy-rates-by-city/

Bad reviews: Why people write them, and what they expect (2019, Jan 31). *Trustpilot Business blog*. Accessed 2020, https://uk.business.trustpilot.com/guides-reports/learn-from-customers/bad-reviews-why-people-write-them-and-what-they-expect

Baldwin, Roberto (2014). Airbnb updates design and introduces controversial new Bélo logo. *The Next Web*. https://thenextweb.com/news/airbnb-updates-design-introduces-new-belo-logo

Barron, Kyle, Edward Kung, and Davide Proserpio (2021). The Effect of Home-Sharing on House Prices and Rents: Evidence from Airbnb. *Marketing Science, 40*(1). DOI: 10.1287/mksc.2020.1227

Bird, Steven, Ewan Klein, and Edward Loper (2009). Natural Language Processing with Python. *O'Reilly Media, Inc.* 1st edition. DOI: 10.5555/1717171. ISBN 0596516495.

Bohren, J. Aislinn, Alex Imas, and Michael Rosenberg (2019). The Dynamics of Discrimination: Theory and Evidence. *American Economic Review, 109* (10): 3395-3436. DOI: 10.1257/aer.20171829

Bolton, Gary, Ben Greiner, and Axel Ockenfels (2013). Engineering Trust: Reciprocity in the Production of Reputation Information. *Management Science 59*(2):256-285. https://doi.org/10.1287/mnsc.1120.1609

Boring, Anne (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics, 145*, 27-41. https://doi.org/10.1016/j.jpubeco.2016.11.006.

Buchan, Nancy, Rachel Croson, and Sara Solnick (2008). Trust and Gender: An Examination of Behavior and Beliefs in the Investment Game. *Journal of Economic Behavior & Organization, 68*, 466-476. DOI: 10.1016/j.jebo.2007.10.006.

Building Trust With a New Review System (2014, July 10). *Airbnb*. Accessed August 2017, available at https://web.archive.org/web/20170825123153/http://blog.atairbnb.com/building-trust-new-review-system

Calabral, Luís, and Ali Hortaçsu (2010). The Dynamics of Seller Reputation: Evidence from eBay. *Journal of Industrial Economics, vol. 58*, issue 1, 54-78. https://doi.org/10.1111/j.1467-6451.2010.00405.x

Cameron, Steffani (2017). AirBNB Reviews: Why Honesty and Expectations Matter. *Full Nomad blog*. Available at https://www.fullnomad.com/2017/08/06/airbnb-honest-reviews-expectations-matter/

Chaudhuri, Ananish, and Lata Gangadharan (2003). Gender Differences in Trust and Reciprocity. *University of Auckland, Economics Department Working Papers*, Number 136. https://www.researchgate.net/publication/246292193_Gender_Differences_in_Trust_and_Reciprocity_1

Correll, Shelley, and Caroline Simand (2016). Vague Feedback is Holding Women Back. *Harvard Business Review*. https://hbr.org/2016/04/research-vague-feedback-is-holding-women-back

Croson, Rachel, and Nancy Buchan (1999). Gender and Culture: International Experimental Evidence from Trust Games. *American Economic Review, 89* (2): 386-391. DOI: 10.1257/aer.89.2.386

Dann, David, Timm Teubner, and Christof Weinhardt (2019). Poster child and guinea pig – insights from a structured literature review on Airbnb. *International Journal of Contemporary Hospitality Management, Vol. 31*, No. 1, pp. 427-473. https://doi.org/10.1108/IJCHM-03-2018-0186

Dehdari, Sirus Håfström, Emma Heikensten, and Siri Isaksson (2019). What Goes Around (Sometimes) Comes Around: Gender Differences in Retaliation. *Working Paper*. Available at SSRN, http://dx.doi.org/10.2139/ssrn.3378279

Dellarocas, Chrysanthos, and Charles A. Wood (2008). The Sound of Silence in Online Feedback:

Estimating Trading Risks in the Presence of Reporting Bias. *Management Science 54*(3):460-476. https://doi.org/10.1287/mnsc.1070.0747

Dittrich, Marcus (2015). Gender differences in trust and reciprocity: evidence from a large-scale experiment with heterogeneous subjects. *Applied Economics, 47*:36, 3825-3838, DOI: 10.1080/00036846.2015.1019036

Edelman, Benjamin, Michael Luca, and Dan Svirsky (2017). Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. *American Economic Journal: Applied Economics, 9* (2): 1-22. DOI: 10.1257/app.20160213

Fatoki, Olawale (2013). Internal Whistleblowing Intentions of Accounting Students in South Africa: The Impact of Fear of Retaliation, Materiality and Gender. *Journal of Social Sciences*. https://www.tandfonline.com/doi/abs/10.1080/09718923.2013.11893202

Fradkin, Andrey, Elena Grewal, and David Holtz (2021). Reciprocity and Unveiling in Two-Sided Reputation Systems: Evidence from an Experiment on Airbnb. *Marketing Science, 40*(6):1013-1029. https://doi.org/10.1287/mksc.2021.1311

Galvao, Antonio F., and Kengo Kato (2017). Chapter 19: Quantile Regression Methods for Longitudinal Data. In “Handbook of Quantile Regression”, edited by Roger Koenker, Victor Chernozhukov, Xuming He, and Limin Peng. 1st Edition. *Chapman and Hall/CRC*. ISBN 9781315120256

Garbarino, Ellen, and Robert Slonim (2009). The robustness of trust and reciprocity across a heterogeneous U.S. population. *Journal of Economic Behavior & Organization, 69*, issue 3, p. 226-240. https://econpapers.repec.org/RePEc:eee:jeborg:v:69:y:2009:i:3:p:226-240

Gecko, Zed (2007). Overview of the program “gender” by Jörg Michael. *AutoHotKey Forum*. Available at https://www.autohotkey.com/board/topic/20260-gender-verification-by-forename-cmd-line-tool-db/

Hardonk, Loïs (2020). Gender Discrimination on Airbnb: the Effect of Host Ethnicity. *Utrecht University Student Theses Repository*. https://studenttheses.uu.nl/handle/20.500.12932/36118

Heinz, Matthias, Steffen Juranek, and Holger A. Rau (2012). Do women behave more reciprocally than men? Gender differences in real effort dictator games. *Journal of Economic Behavior & Organization, Elsevier, vol. 83*(1), pages 105-110. DOI: 10.1016/j.jebo.2011.06.015

Hengel, Erin (2022). Publishing While Female: are Women Held to Higher Standards? Evidence from Peer Review. *The Economic Journal, Royal Economic Society, vol. 132*(648), pages 2951-2991. https://ideas.repec.org/a/oup/econjl/v132y2022i648p2951-2991..html

HNN Newswire (2015). Airbnb Tops $1.15b in NYC Economic Activity. *CoStar*. https://www.costar.com/article/1375536831/airbnb-tops-115b-in-nyc-economic-activity

How to Get 5-Star Reviews on Airbnb (2024). *AirDNA blog*. https://www.airdna.co/blog/how-to-get-five-star-reviews-on-airbnb

How Search Results Work (n.b.). *Airbnb Help Center*. Accessed May 2021, https://www.airbnb.com/help/article/39

Hutto, C.J., and Eric Gilbert (2015). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM*. https://www.researchgate.net/publication/275828927_VADER_A_Parsimonious_Rule-based_Model_for_Sentiment_Analysis_of_Social_Media_Text

Iakubovskyi, Dmytro (2018, Oct 1). Digging into Airbnb data: reviews, sentiments, superhosts, and prices prediction (part 1). *Medium.com: Towards Data Science*. Accessed March 2020, https://towardsdatascience.com/digging-into-airbnb-data-reviews-sentiments-superhosts-and-prices-prediction-part1-6c80ccb26c6a

InsideAirbnb (n.b.) https://insideairbnb.com/. Data at https://insideairbnb.com/get-the-data

I.V., Shravan (2016). Sentiment Analysis in Python using NLTK. *Open Source For U*. https://www.opensourceforu.com/2016/12/analysing-sentiments-nltk/

Kakar, Venoo, Joel Voelz, Julia Wu, and Julisa Franco (2018). The Visible Host: Does race guide Airbnb rental rates in San Francisco? *Journal of Housing Economics, Volume 40*, Pages 25-40, ISSN 1051-1377, https://doi.org/10.1016/j.jhe.2017.08.001

Koenker, Roger, Victor Chernozhukov, Xuming He, and Limin Peng (Editors) (2017). Handbook of Quantile Regression (1st ed.). *Chapman and Hall/CRC*. https://doi.org/10.1201/9781315120256

Kovachevska, Marija (2020). 28 Amazing Airbnb Statistics You Should Know Before Booking. *Capital Counselor*. Accessed May 20, 2021. https://capitalcounselor.com/airbnb-statistics/

Kundro, Timothy G., and Nancy P. Rothbard (2022). Does Power Protect Female Moral Objectors? How and When Moral Objectors' Gender, Power, and Use of Organizational Frames Influence Perceived Self-Control and Experienced Retaliation. *Academy of Management Journal*. https://doi.org/10.5465/amj.2019.1383

Ledesma, Josue (2020). Why do people write reviews? What our research revealed. *Trustpilot Business Blog*. Available at https://business.trustpilot.com/reviews/learn-from-customers/why-do-people-write-reviews-what-our-research-revealed

Li, Xiaodi (2018). Which Neighborhoods Join the Sharing Economy and Why? The Case of the Short-term Rental Market in New York City. *AEA Conference Paper*. https://www.aeaweb.org/conference/2019/preliminary/paper/r3z6nbi2

Liyanarachchi, Gregory, and Ralph Adler (2011). Accountants' Whistle-Blowing Intentions: The Impact of Retaliation, Age, and Gender. *Australian Accounting Review*. https://doi.org/10.1111/j.1835-2561.2011.00134.x

Machado, José A.F., and J.M.C. Santos Silva (2019). Quantiles via moments. *Journal of Econometrics, Volume 213*, Issue 1, Pages 145-173. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2019.04.009

Macnell, Lillian, Adam Driscoll, and Andrea Hunt (2014). What's in a Name: Exposing Gender Bias in Student Ratings of Teaching. *Innovative Higher Education*. doi:10.1007/s10755-014-9313-4.

Marchenko, Anya (2019). The impact of host race and gender on prices on Airbnb. *Journal of Housing Economics, Volume 46*, 101635, ISSN 1051-1377, https://doi.org/10.1016/j.jhe.2019.101635.

Mitchell, Kristina, and Jonathan Martin (2018). Gender Bias in Student Evaluations. *Political Science & Politics (PS), 51*(3), 648-652. doi:10.1017/S104909651800001X

Mousavi, Reza, and Kexin Zhao (2022). Examining the Impacts of Airbnb Review Policy Change on Listing Reviews. *Journal of the Association for Information Systems, 23*(1), 303-328. DOI: 10.17705/1jais.00720

Morris, Carolyn (2021). How Much Are People Making From the Sharing Economy? *Earnest, Data Blog*, August 13, 2021. Available at https://www.earnest.com/blog/sharing-economy-income-data/

Nosko, Chris, and Steven Tadelis (2015). The Limits of Reputation in Platform Markets: An Empirical Analysis and Field Experiment. *National Bureau of Economic Research (NBER) Working Paper Series*, number 20830. http://www.nber.org/papers/w20830

Our Community of Teacher Hosts (2020, Dec 4). *Airbnb News*. https://news.airbnb.com/our-community-of-teacher-hosts/

Penaflorida II, Rexly (2020, Dec 2). A Guide to How Airbnb Reviews Work. *Review Trackers Blog*. https://www.reviewtrackers.com/blog/airbnb-reviews/

Pérez, Israel Saeta (2016). Gender-guesser. *Python Package Index*. https://pypi.python.org/pypi/gender-guesser

Porges, Seth (2014, Oct 17). The Strange Game Theory of AirBNB Reviews. *Forbes*. https://www.forbes.com/sites/sethporges/2014/10/17/the-strange-game-theory-of-airbnb-reviews/

Protalinski, Emil (2014, July 10). Airbnb revamps its review system: Hosts and guests see feedback simultaneously, review period cut to 14 days. *The Next Web blog*. https://thenextweb.com/news/airbnb-revamps-review-system-hosts-guests-see-feedback-simultaneously-review-period-cut-14-days

Rehg, Michael T., Marcia P. Miceli, Janet P. Near, and James R. Van Scotter (2008). Antecedents and Outcomes of Retaliation against Whistleblowers: Gender Differences and Power Relationships. *Organization Science*, 19(2), 221–240. http://www.jstor.org/stable/25146176

Rosenberg, Eli (2018). Another sign of hard times for teachers? They make up nearly 10 percent of Airbnb hosts. *The Mercury News*. https://www.mercurynews.com/2018/08/20/another-sign-of-hard-times-for-teachers-they-make-up-nearly-10-percent-of-airbnb-hosts/

Santamaría, Lucía, and Helena Mihaljević (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science 4:e156*. https://doi.org/10.7717/peerj-cs.156

Shapiro, Nick (2017, Dec 1). Perfect Strangers: How Airbnb is building trust between hosts and guests. *Airbnb Newsroom*. https://news.airbnb.com/perfect-strangers-how-airbnb-is-building-trust-between-hosts-and-guests/

Shared Opportunity: How Airbnb Benefits Communities (2015). *Federal Trade Commission & Airbnb*. Accessed July 2021, https://www.ftc.gov/system/files/documents/public_comments/2015/05/01740-96152.pdf

Shatford, Scott (2018). What is Airbnb's Superhost Status Really Worth? *AirDNA*. https://www.airdna.co/blog/airbnb_superhost_status#:~:text=Airbnb%20launched%20their%20Superhost%20program,with%20a%20special%20VIP%20status

Smith, Aaron (2016). Gig Work, Online Selling and Home Sharing. *Pew Research Center*. https://www.pewresearch.org/internet/2016/11/17/gig-work-online-selling-and-home-sharing/

Tadelis, Steven (2016). Reputation and Feedback Systems in Online Platform Markets. *Annual Review of Economics, 8*:321-340. https://doi.org/10.1146/annurev-economics-080315-015325

Teacher Hosts Earned $230 Million in 2019 (Aug 2020). *Airbnb Newsroom*. https://news.airbnb.com/teacher-hosts-earned-230-million-in-2019/

The Airbnb Story (n.b.). *Airbnb Newsroom*. Accessed August 2021. Available at https://news.airbnb.com/about-us/

Vasilescu, Bogdan N., Andrea Capiluppi, and Alexander Serebrenik (2014). Gender, representation and online participation: a quantitative study. *Interacting with Computers, 26*(5), 488-511. https://doi.org/10.1093/iwc/iwt047

What are some tips that new hosts on AirBnB can use to be successful with their first few guests? (n.b.) *Quora thread*. https://www.quora.com/What-are-some-tips-that-new-hosts-on-AirBnB-can-use-to-be-successful-with-their-first-few-guests

Why would they write that?! The psychology of customer reviews (2018). *National Strategic Group*. Available at https://www.nationalstrategic.com/why-would-they-write-that-the-psychology-of-customer-reviews/

Yates, Tyler (2020). How Much Are People Making From the Sharing Economy? *Earnest, Data Blog*. Accessed March 2020. Available at https://web.archive.org/web/20201225192234/https://www.earnest.com/blog/sharing-economy-income-data/#webpage

Zervas, Georgios, Davide Proserpio, and John Byers (2021). A First Look at Online Reputation on Airbnb, Where Every Stay is Above Average. *Marketing Letters, 32* (1), No 1, 16 pages. DOI: 10.1007/s11002-020-09546-4

Appendix 2.A

Miscellaneous Tables & Figures

Table 2.A.1: **Scrape dates for the cities in the sample**

Area	Scrape date
Boston	3 Oct 2015
Chicago	3 Oct 2015
Los Angeles	25 Jul 2015
Nashville	3 Oct 2015
New Orleans	2 Sep 2015
New York City	1 Aug 2015
Portland	2 Sep 2015
San Francisco	2 Sep 2015
Santa Cruz County	15 Oct 2015
Washington DC	3 Oct 2015

Table 2.A.2: **Sample is representative, resembles excluded observations on observables**

%		**In Sample**	**Excluded**
City	Boston	3	4.5
	Chicago	5.6	7.7
	Los Angeles	21.1	19.7
	Nashville	2.8	3.8
	New Orleans	7.2	4.5
	New York City	32.9	33.9
	Portland	7.1	6.6
	San Francisco	14.1	11.4
	Santa Cruz County	1.7	2.2
	Washington DC	4.6	5.7
Listing Type	Entire Home	60.6	57.7
	Private Room	37.8	40.1
	Shared Room	1.7	2.2
Review Date	Min	7/1/2013	10/6/2008
	Max	6/30/2015	10/17/2015
N	Total	243,123	803,474

Table 2.A.3: **Balance between early and late-joining hosts**

	Female host (%)	Male host (%)	No. of reviews per month (avg)	Price* ($)	N (listings)
Joined before July 2013 (in sample)	33.1	32.3	0.53	171.49	31,701
Joined between July 2013 and July 2014	32.6	34.0	0.69	177.92	17,465
Joined between July 2014 and July 2015	32.9	33.1	0.88	186.31	21,618

* Price is as of the scrape date in Fall of 2015, and does not reflect any variations over time.

Table 2.A.4: **Balance in Who Reviews around Policy Change**

	Gender	Property Type		Rental Type	
%	Female Guest	House	Apartment	Entire home	Private Room
Pre (July 2013 to 2014)	52.8	30.2	63.3	59.7	38.7
Post (July 2014 to 2015)	51.9	29.6	64.2	61.0	37.3

OLS results are nearly identical to the panel with listing FE results. The OLS regressions include time-invariant controls that are not included in the FE specifications since they are subsumed in the listing fixed effects. Controls include city or zipcode fixed effects, month or month x year fixed effects, and other characteristics of the listing such as property type (house, apartment building, etc) and room type (entire apartment, private room, or shared room).

$$\begin{aligned} ReviewSentimentIndex_{i,t} = \beta_1 FemaleHost_i + \beta_2 Post_t + \beta_3 (FemaleHost_i \times Post_t) + \\ \gamma(City_i + Time_t + PropertyType_i + RoomType_i) + \epsilon_{i,t} \end{aligned} \tag{2.A.1}$$

Table 2.A.5: **Policy's impact is half as large for female hosts (OLS), in line with main FE results**

z_review_compound	(1)	(2)	(3)	(4)
female_host	1.984**	1.914**	1.476*	1.522*
	(0.892)	(0.893)	(0.868)	(0.868)
post	-8.142***	-8.149***	-8.201***	
	(0.682)	(0.683)	(0.662)	
female_host * post	4.262***	4.309***	4.458***	4.407***
	(0.921)	(0.922)	(0.908)	(0.907)
N	242947	242947	241872	241872
adj. R^2	0.004	0.005	0.014	0.014
Location_controls	City	City	Zipcode	Zipcode
Time_controls	No	Month FE	Month FE	Month x year

Standard errors in parentheses

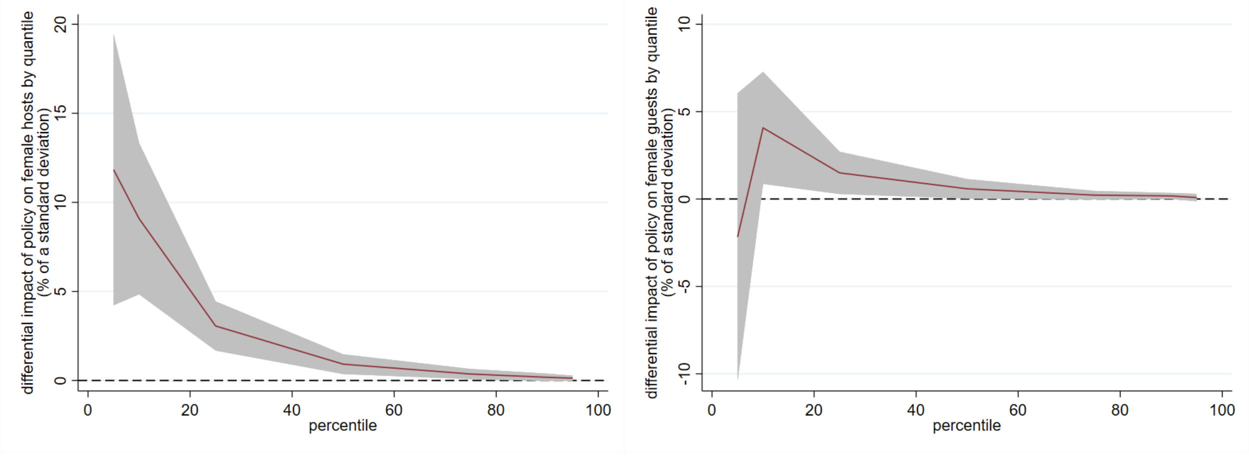
* $p < .10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.6: **Male and female guests react in the same way (OLS), in line with main FE results**

z_review_compound	(1)	(2)	(3)	(4)
female_guest	12.92^{***}	12.89^{***}	12.45^{***}	12.46^{***}
	(0.691)	(0.691)	(0.687)	(0.687)
post	-6.292^{***}	-6.281^{***}	-6.258^{***}	
	(0.657)	(0.657)	(0.651)	
female_guest * post	0.693	0.705	0.689	0.689
	(0.860)	(0.860)	(0.857)	(0.857)
N	242947	242947	241872	241872
adj. R^2	0.008	0.008	0.017	0.018
Location_controls	City	City	Zipcode	Zipcode
Time_controls	No	Month FE	Month FE	Month x year

Standard errors in parentheses

* $p < .10$, ** $p < 0.05$, *** $p < 0.01$



(a) **Female hosts less negatively affected by policy change, especially at bottom of distribution**

(b) **Female guests give slightly better reviews for all but best and worst stays**

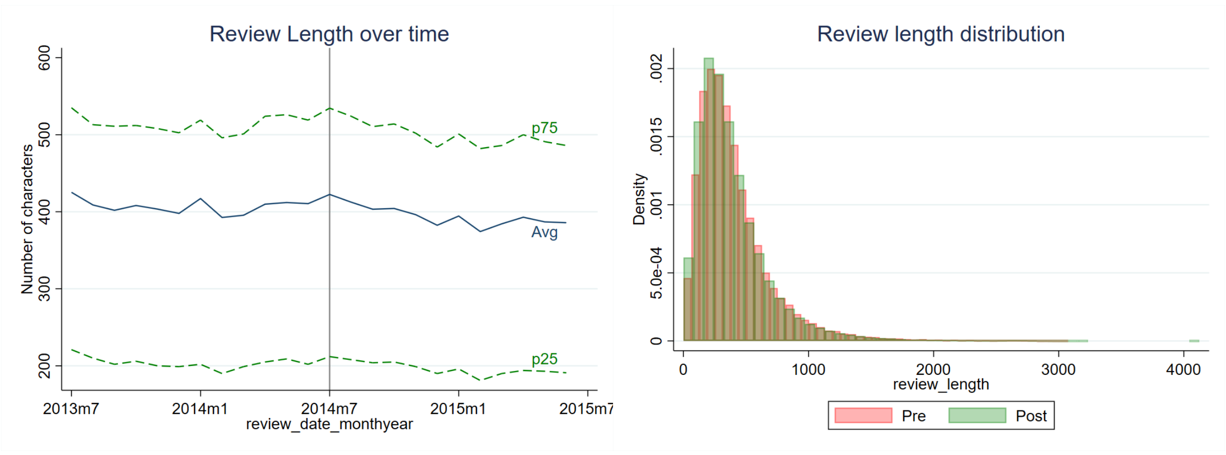
Figure 2.A.1: **Differential impact of policy by gender by quantile**

Table 2.A.7: **Gender of host is more important than gender of guest (2x2)**

	(1) compound	(2) negative	(3) neutral	(4) positive	(5) compound	(6) negative	(7) neutral	(8) positive
female guest	8.063***	-0.00116***	-0.00301***	0.00406***	9.984***	-0.00164***	-0.00488***	0.00660***
	(0.974)	(0.000264)	(0.00107)	(0.00110)	(0.949)	(0.000265)	(0.00105)	(0.00110)
post	-8.999***	0.000768***	0.0101***	-0.0109***	-3.755***	-0.000280	0.00589***	-0.00555***
	(0.933)	(0.000258)	(0.00106)	(0.00109)	(0.935)	(0.000281)	(0.00105)	(0.00109)
female guest * post	1.925	-0.000431	0.00135	-0.000855	-1.215	0.000199	0.00351***	-0.00374***
	(1.238)	(0.000331)	(0.00133)	(0.00137)	(1.165)	(0.000335)	(0.00131)	(0.00136)
N	121523	121523	121523	121523	121600	121600	121600	121600
adj. R^2	0.017	0.035	0.206	0.227	0.020	0.036	0.207	0.227
Time_controls	Month	Month	Month	Month	Month	Month	Month	Month
Host_FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Host_Gender	Male	Male	Male	Male	Female	Female	Female	Female

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$



(a) **Reviews become 12.5 characters shorter on average**

(b) **Review length distribution marginally shifts left**

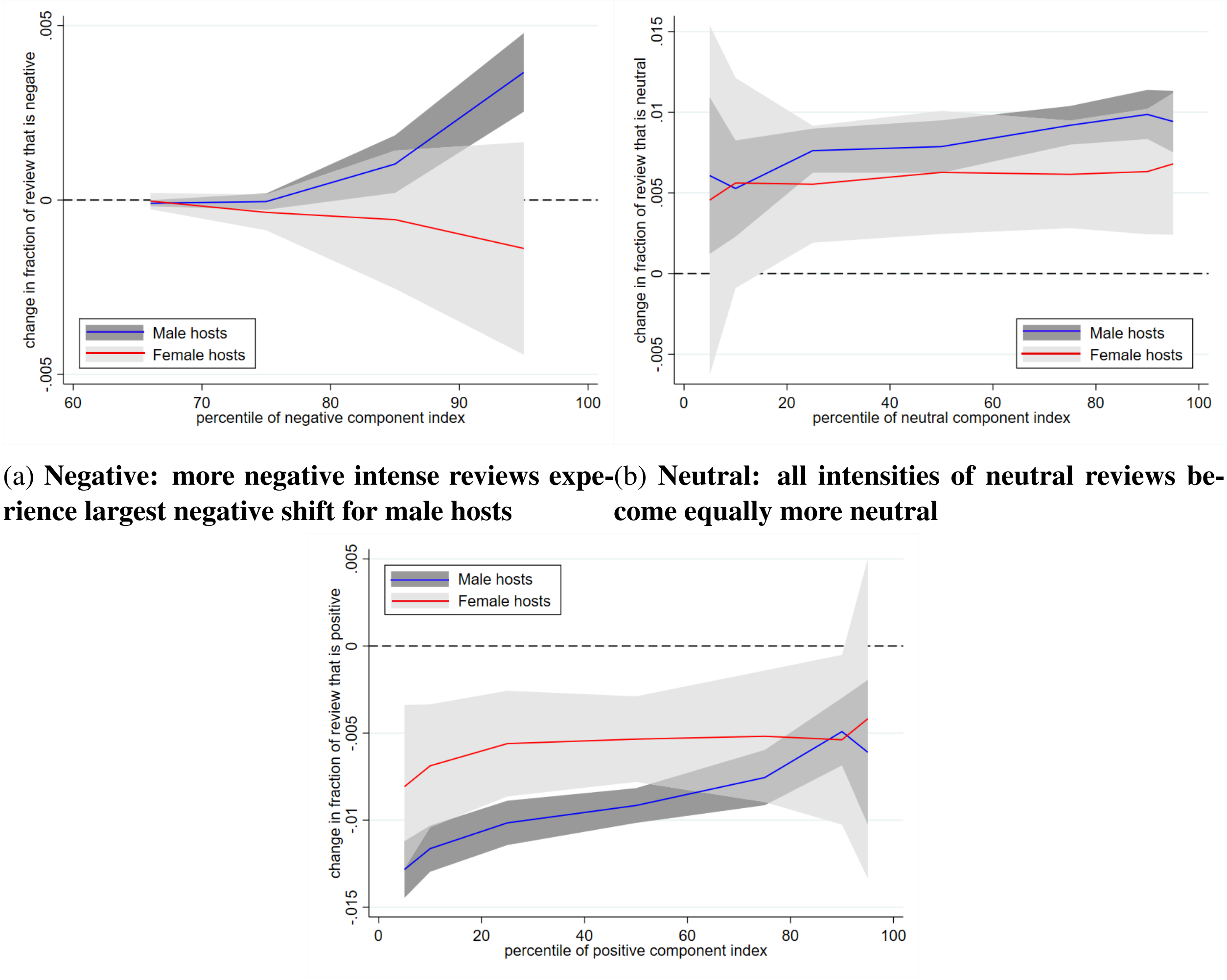
Figure 2.A.2: **Reviews become slightly shorter**

Table 2.A.8: **Main results of differential effect for male hosts robust across alternative specifications of post**

review compound (z)	(1)	(2)	(3)	(4)	(5)
post	-8.427***	-8.073***	-8.174***	-8.484***	-7.708***
	(0.696)	(0.698)	(0.703)	(0.722)	(0.846)
female host x post	3.684***	3.603***	2.926***	3.495***	4.171***
	(0.958)	(0.960)	(0.946)	(0.987)	(1.145)
N	243123	243123	243123	232335	203849
adj. R^2	0.001	0.001	0.001	0.001	0.001
N clusters	14372	14372	14372	14332	14201
Time controls	Month FE	Month FE	Month FE	Month FE	Month FE
Listing FE	Yes	Yes	Yes	Yes	Yes
Post definition	Jul 1	Jul 10	Aug 1	Aug 1	Aug 1
Excluded dates				Jul	Apr-Jul

Standard errors in parentheses. Column (1) matches main specification of Column (2) in Table 2.8.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$



(a) **Negative: more negative intense reviews experience largest negative shift for male hosts**

(b) **Neutral: all intensities of neutral reviews become equally more neutral**

(c) **Positive: reviews with a smaller share of positive words experience a larger drop**

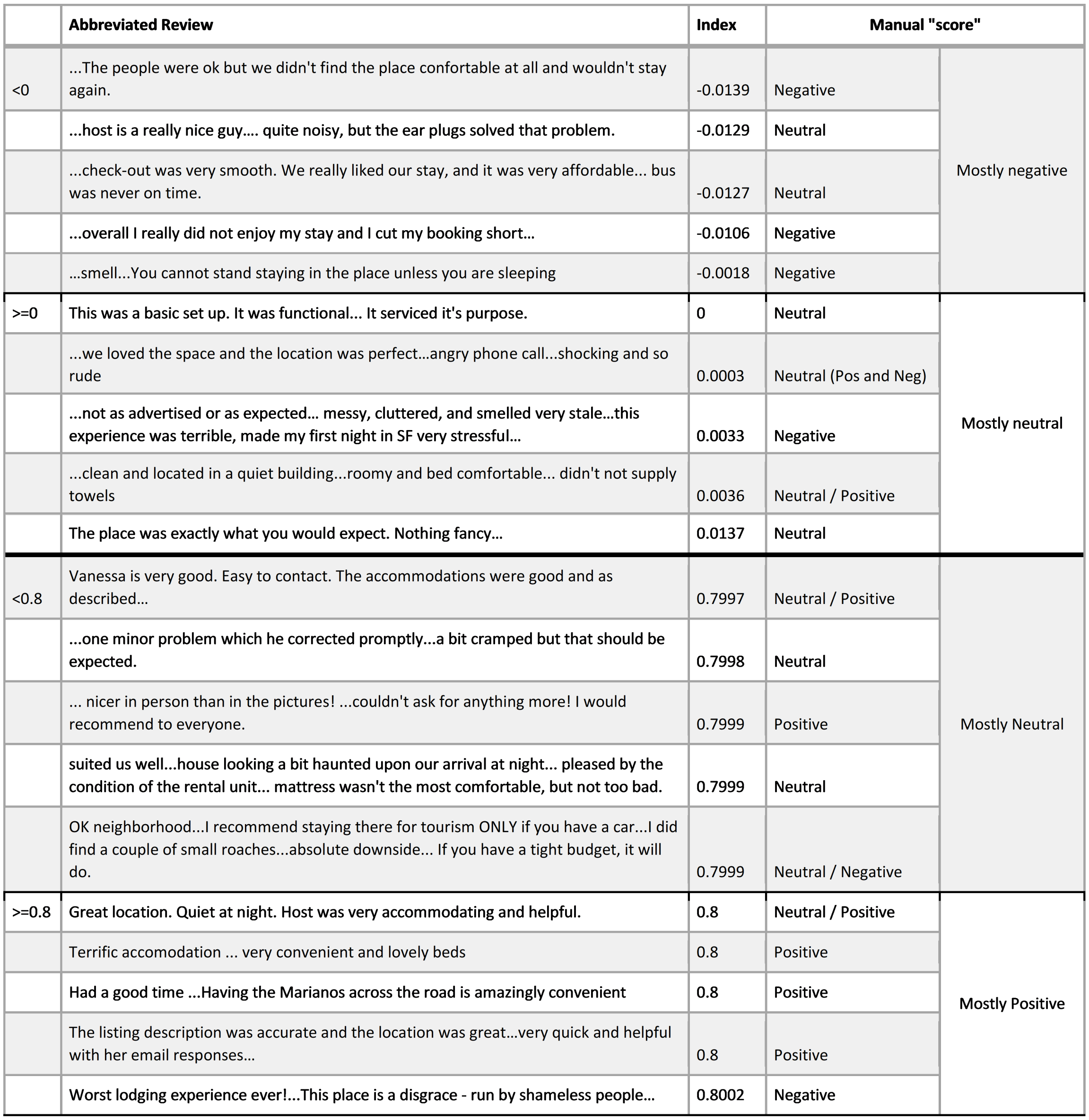
Figure 2.A.3: **By fraction of review that is {negative, neutral, positive} words, the most negative reviews become even more negative towards male hosts**

Table 2.A.9: **Main results of no differential effect by guest gender is robust to alternative specifications of post**

review compound (z)	(1)	(2)	(3)	(4)	(5)
female guest	10.83***	10.73***	10.81***	10.81***	11.48***
	(0.681)	(0.675)	(0.644)	(0.682)	(0.807)
post	-6.923***	-6.703***	-7.106***	-7.133***	-5.579***
	(0.669)	(0.667)	(0.681)	(0.693)	(0.825)
female guest x post	0.673	0.836	0.752	0.769	0.153
	(0.852)	(0.857)	(0.851)	(0.873)	(0.980)
N	243123	243123	243123	232335	203849
adj. R^2	0.004	0.004	0.004	0.004	0.005
N clusters	14372	14372	14372	14332	14201
Time controls	Month FE	Month FE	Month FE	Month FE	Month FE
Listing FE	Yes	Yes	Yes	Yes	Yes
Post definition	Jul 1	Jul 10	Aug 1	Aug 1	Aug 1
Excluded dates				Jul	Apr-Jul

Standard errors in parentheses. Column (1) matches the main specification of Column (2) in Table 2.9.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$



	Abbreviated Review	Index	Manual "score"	
<0	...The people were ok but we didn't find the place confortable at all and wouldn't stay again.	-0.0139	Negative	Mostly negative
	...host is a really nice guy…. quite noisy, but the ear plugs solved that problem.	-0.0129	Neutral	
	...check-out was very smooth. We really liked our stay, and it was very affordable... bus was never on time.	-0.0127	Neutral	
	...overall I really did not enjoy my stay and I cut my booking short…	-0.0106	Negative	
	…smell...You cannot stand staying in the place unless you are sleeping	-0.0018	Negative	
>=0	This was a basic set up. It was functional... It serviced it's purpose.	0	Neutral	Mostly neutral
	...we loved the space and the location was perfect…angry phone call...shocking and so rude	0.0003	Neutral (Pos and Neg)	
	...not as advertised or as expected… messy, cluttered, and smelled very stale…this experience was terrible, made my first night in SF very stressful…	0.0033	Negative	
	...clean and located in a quiet building...roomy and bed comfortable... didn't not supply towels	0.0036	Neutral / Positive	
	The place was exactly what you would expect. Nothing fancy…	0.0137	Neutral	
<0.8	Vanessa is very good. Easy to contact. The accommodations were good and as described…	0.7997	Neutral / Positive	Mostly Neutral
	...one minor problem which he corrected promptly...a bit cramped but that should be expected.	0.7998	Neutral	
	... nicer in person than in the pictures! ...couldn't ask for anything more! I would recommend to everyone.	0.7999	Positive	
	suited us well...house looking a bit haunted upon our arrival at night... pleased by the condition of the rental unit... mattress wasn't the most comfortable, but not too bad.	0.7999	Neutral	
	OK neighborhood...I recommend staying there for tourism ONLY if you have a car...I did find a couple of small roaches...absolute downside... If you have a tight budget, it will do.	0.7999	Neutral / Negative	
>=0.8	Great location. Quiet at night. Host was very accommodating and helpful.	0.8	Neutral / Positive	Mostly Positive
	Terrific accomodation ... very convenient and lovely beds	0.8	Positive	
	Had a good time ...Having the Marianos across the road is amazingly convenient	0.8	Positive	
	The listing description was accurate and the location was great…very quick and helpful with her email responses…	0.8	Positive	
	Worst lodging experience ever!...This place is a disgrace - run by shameless people…	0.8002	Negative	

Figure 2.A.4: **Reviews near Logit Cutoffs**

Table 2.A.10: **Host results by gender consistent across gender imputation methods**

	(1) Base	(2) Strict	(3) Top20	(4) J&S	(5) GComp	(6) Base	(7) Strict	(8) Top20	(9) J&S	(10) GComp
female_host	1.919**	1.847*	1.092	-2.675	2.002**	0	0	0	0	0
	(0.893)	(0.963)	(1.957)	(4.562)	(0.900)	(.)	(.)	(.)	(.)	(.)
post	-8.133***	-8.513***	-8.640***	-10.12***	-7.948***	-8.419***	-8.698***	-9.151***	-11.27***	-8.252***
	(0.683)	(0.761)	(1.247)	(3.030)	(0.689)	(0.698)	(0.787)	(1.279)	(3.222)	(0.706)
1.female_host#1.post	4.295***	4.664***	7.233***	8.377	3.830***	3.682***	4.098***	6.761***	12.32**	3.270***
	(0.923)	(1.032)	(1.994)	(5.300)	(0.931)	(0.960)	(1.086)	(2.119)	(5.757)	(0.972)
N	242692	190437	58319	6244	238733	242692	190437	58319	6244	238733
adj. R^2	0.005	0.005	0.007	0.008	0.004	0.001	0.001	0.001	0.003	0.001
N_clust	14366	12143	3277	347	14231	14366	12143	3277	347	14231
Location_controls	City	City	City	City	City	City	City	City	City	City
Time_controls	Month FE	Month FE	Month FE	Month FE	Month FE	Month FE	Month FE	Month FE	Month FE	Month FE
Host_FE	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses

* $p < .10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.11: **Guest results by gender consistent across gender imputation methods**

	(1) Base	(2) Strict	(3) Top20	(4) J&S	(5) GComp	(6) Base	(7) Strict	(8) Top20	(9) J&S	(10) GComp
female_guest	12.91***	13.40***	16.63***	26.42***	12.76***	10.84***	11.33***	16.17***	38.77***	10.64***
	(0.692)	(0.763)	(1.512)	(5.522)	(0.694)	(0.682)	(0.760)	(1.641)	(8.645)	(0.685)
post	-6.261***	-6.268***	-4.430***	-1.099	-6.430***	-6.906***	-6.799***	-3.258**	4.348	-7.148***
	(0.658)	(0.736)	(1.344)	(5.651)	(0.671)	(0.670)	(0.761)	(1.523)	(9.904)	(0.684)
1.female_guest#1.post	0.681	0.315	-1.651	-3.127	0.899	0.655	0.280	-3.522*	-9.964	1.014
	(0.861)	(0.953)	(1.868)	(6.580)	(0.863)	(0.854)	(0.956)	(2.046)	(11.06)	(0.855)
N	242692	190437	45402	4628	238733	242692	190437	45402	4628	238733
adj. R^2	0.008	0.009	0.010	0.017	0.008	0.004	0.005	0.006	0.039	0.004
N_clust	14366	12143	10115	3186	14231	14366	12143	10115	3186	14231
Location_controls	City	City	City	City	City	City	City	City	City	City
Time_controls	Month FE	Month FE	Month FE	Month FE	Month FE	Month FE	Month FE	Month FE	Month FE	Month FE
Host_FE	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses

* $p < .10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.12: **Gender differential for Hosts persist across room types**

z_review_compound	(1)	(2)	(3)	(4)
female_host	1.914**	3.268***	0.197	-10.32
	(0.893)	(1.121)	(1.483)	(7.185)
post	-8.149***	-9.012***	-6.154***	-16.39***
	(0.683)	(0.873)	(1.113)	(4.692)
female host x post	4.309***	3.753***	4.365***	9.949
	(0.922)	(1.178)	(1.489)	(7.582)
N	242947	147092	91765	4090
adj. R^2	0.005	0.006	0.003	0.006
N_clust	14367	9048	4993	326
Location_controls	City	City	City	City
Time_controls	Month FE	Month FE	Month FE	Month FE
Room_type	All	Entire Home	Private Room	Shared Room

Standard errors in parentheses

* $p < .10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.13: **Lack of gender differential by Guests persists across room types**

z_review_compound	(1)	(2)	(3)	(4)
female_guest	12.89***	12.41***	13.53***	12.72**
	(0.691)	(0.900)	(1.092)	(5.426)
post	-6.281***	-7.572***	-3.986***	-14.78***
	(0.657)	(0.857)	(1.044)	(4.586)
female guest x post	0.705	0.880	0.374	6.305
	(0.860)	(1.127)	(1.352)	(6.330)
N	242947	147092	91765	4090
adj. R^2	0.008	0.009	0.007	0.012
N_clust	14367	9048	4993	326
Location_controls	City	City	City	City
Time_controls	Month FE	Month FE	Month FE	Month FE
Room_type	All	Entire Home	Private Room	Shared Room

Standard errors in parentheses

* $p < .10$, ** $p < 0.05$, *** $p < 0.01$

Appendix 2.B

Model Details and Extensions

This Appendix 2.B will detail the model and break down how each component depends on x, y, θ, ϕ, and gender. Then we will walk through retaliation's role by analyzing each decision game by review rules. Lastly, I discuss loosening the assumption that guests review first.

Private net benefits $b(x, \theta)$ consist of expressive joy e, other regarding preferences s, and a time cost c. Expressive joy $e(x, \theta)$ is the utility derived from expressing ones experience. Leaving a bad Amazon rating when the product arrived damaged to get it off your chest and leaving a great review for a pillow that has stopped your neck pain are both examples of expressive joy.[1] These also highlight the interaction of review sentiment x and quality of stay θ in determining the value of e. Expressive joy is most beneficial when it accurately reflects the quality. In the first example, both x and θ are negative, while in the last example both are positive. In other words, $e(\cdot)$ is maximized when $x = \theta$. By the same token, expressive joy for individual j is also maximized when $y = \phi$.

Other regarding preferences represents utility from knowing your review will benefit the society of users. A person's reviews create their reputation on the platform and shape future interactions. Uber riders with a higher rating are more likely to be accepted by drivers. Unsafe Uber drivers can be kicked off the platform. These other regarding preferences apply to both positive and negative reviews. As a guest, posting a positive review benefits the host in their future endeavors. If you had a great stay, you want others to know so the host continues to get bookings. A negative

[1] Note that I am assuming the bad review was not written in order to obtain a refund, and that the positive review was not written for payment. These concerns are present on Amazon but very unlikely on Airbnb.

review benefits potential future guests. Posting a negative review warns them of your experience. By the same logic as before, these social preferences $s(x, \theta)$ are maximized when review sentiment matches the quality of the stay. That is, when $x = \theta$ for the guest and when $y = \phi$ for the host.

Reviews are not costless due to the opportunity cost of time. I abstract away from this by defining $b(\cdot)$ as the net benefits. However, one could model this by $c(\cdot)$ taking on the form of a small constant if a review is written, and zero otherwise. That is, $c(x) = c\ \mathbb{1}\{x \neq 0\}$ and $c(y) = c\ \mathbb{1}\{y \neq 0\}$. The opportunity cost term is assumed to be small. The average review is a couple of sentences and can be written from a mobile device on the go.

In addition to benefiting directly from writing reviews, utility depends largely on the type of review that is received. Reviews written about me stay with me for all future interactions on the platform. Reviews influence who accepts me as a guest. As a host, this would affect future bookings and so income from the app. This concept is denoted by the inclusion of the other player's choice in an individual's utility function. Concretely, U_i is a function of j's choice y and U_j depends on i's decision over x. Everyone wants to receive positive reviews, so U_i is increasing in y and U_j is increasing in x. Since y, x are discrete let me be more specific. Receiving a positive review yields the most utility, followed by a neutral review, and then a negative review. Note that unless the review is negative, a review is beneficial to increase the volume of reviews for reputation. By this logic, the utility of a neutral review is greater than zero. In shorthand for the utility of the type of review received, $U_i(y = +) > U_i(y =\sim) > 0 > U_i(y = -)$.

One's expected probability distribution over the type of review they will receive depends on many factors. Namely, the quality of the stay as experienced by both parties θ, ϕ, the review they write for the other party, and gender. We will ignore gender for now and add it back in shortly. Consider a situation where I am a host. A guest, who was wonderful, just checked out. I think they also had a pleasant stay. I plan on leaving a positive review for the guest. However, before I do so, they write a negative review for me. Since I can read their review before writing mine, I decide to write a negative review for them saying they were disrespectful. This is an example of retaliation and how the guest's behavior impacted the host's. In fact, the host copied the guest's behavior in a tit-for-tat mentality. Before July 2014 when reviews were posted immediately, the probability of the other party taking an action is conditional on one's own choices. So in addition to knowing how you experienced the stay and having an expectation of how they experienced the stay, your review behavior affects your utility through the review you receive.

More formally, one's own action taken is positively correlated with the expected probability of the other party taking that same action. Holding θ and ϕ constant, $E_i(p(y = +))$ is increasing in x.[2] Following this logic, the probability of receiving a negative review $E_i(p(y = -))$ is decreasing in x (i.e. there is a lower chance of receiving a negative review the better review you write, and a higher chance of receiving a negative review when you write a negative review). Neutral reviews can be thought of as a mild positive or a mild negative. Consider the case from the previous paragraph. If instead of receiving a negative review as the host I received a neutral review, I would revise my behavior to also give a neutral review in return. This way I have mildly punished the other player for not writing a positive review for me. This is a 'mild positive'. Alternatively, I could have been planning to write a negative review for the guest. I see the guest writes me a great review. I still want to be a little honest about their behavior, so I leave them a neutral review. In this case the neutral review is a 'mild negative'. Also notice that in both of these scenarios, the host revised their decision towards the sentiment of the guest. Therefore the probability the host writes a neutral review is highest when the guest writes a neutral review.

The paragraphs above provide examples where the sentiment of the review written by the second mover is correlated with the sentiment of the first person's review. Fradkin et al (2021) document this phenomenon. The authors have data on reviews written and received by guests and hosts. They find that reviews are less correlated with each other when reviews are simultaneous reveal. Specifically, they measure the correlation in reviews across two different measures: the labeled review text, and the lowest rating. They find large and statistically significant decreases in the correlation of ratings in both measures. The correlation of positive text fell by 50% and the correlation of ratings fell by 48%. Conditional on a host leaving a negative review first, guests

[2] Increasing in x refers to moving from $-$, to $\sim$, to $+$, which is the discrete analog to the continuous idea of review sentiment.

responding with a negative review decreased from 7% to 2.2% of the time. This is evidence that sentiment of reviews under sequential reveal are correlated to a non-zero degree.

This phenomenon and intuition of review correlation and tit-for-tat is also confirmed by the active presence of Airbnb hosts online. Forums such as Quora and Airhosts Forum discuss that when the guest was bad ($\phi = (-)$), there are two schools of thought held by hosts on how to react.[3] One school of thought says that hosts have a responsibility towards reviewing honestly to help other hosts. This is similar to my other regarding preferences term. The other school of thought is that since their own reviews are very important, wait to review until after the guest reviews, to be able to counter strike if needed.

Before analyzing what the model predicts about behavior, let's discuss why I consider the guest the first mover. There are two reasons for this. First, because the stays where the guest reviews first are likely most affected by the policy change, and second, to match the empirical specification.

Bad and neutral stays are likely most affected by the policy change. Great stays will be reviewed positively before and after, but so-so and negative stays can now be reviewed honestly. There are no stipulations on who reviews first, although there are some unwritten 'rules'. Hosts care more about reviews since they have more skin in the game. Building a large set of positive reviews is essential for procuring future bookings and hence, future income. Hosts also use the platform more often and so have a better understanding of the review system. If a host thinks they guest had a good stay, they will often review first and write nice things to prod a good review in return. On the other hand, if a host thinks the guest had a bad stay, they will wait for the guest to review first. In this way they can retaliate if the guest writes a negative review. The threat of retaliation deters guests from reviewing honestly. Since I want to focus on the neutral and bad stays that are affected by the honesty-inducing policy change, the stays where guests review first are those of interest. A number of blogs discuss how hosts follow this intuition (Porges 2014 &

[3] For example, Astaire (2017) discuss host reviewing responsibilities and Anders (2017) discusses waiting to review to "counter strike" if needed. Also see What are some tips that new hosts on AirBnB can use (n.b.) and AirHosts Forum (n.b.)

Anders 2017) and sellers on Ebay act similarly (Bolton et al 2013).

Due to data limitations I only observe reviews written by guests for hosts. First movers' behavior is most affected by the policy change since their choice no longer needs to consider knock-on effects. Since I can identify behavior changes for guests, defining the guest as the first reviewer allows the model to reflect the empirical analysis.

However, if we were to loosen this assumption and let the host review first sometimes, the main takeaways stay the same but my results are attenuated. In the sequential review system, if the host reviews first, then the subsequent guest's review decision for the host has no uncertainty. The guest sees the review written for them and has no fear of retaliation or anticipation of reciprocity, but instead, can retaliate or reciprocate themselves. This means that these guests who were reviewed first by the hosts, are less affected by the policy change. Therefore the set of reviews written by guests for hosts is attenuated when you allow for some hosts to review first. Astaire (2017) and Anders (2017) suggest that this is a subset of guests, but the model implies that my results are attenuated due to the fact that anyone can be the first mover prior to July 2014.

Chapter 3

The Hidden Cost of Strict Job Qualification Requirements: Application Gaps, Diversity, and Perceptions about Hiring

Amanda Bonheur and Tanner Eastmond [0]

Abstract

Despite years of policy and revised corporate practice intended to correct inequality in the hiring process, application gaps persist for women and individuals from underrepresented racial minority groups. This study explores whether it is possible to narrow this application gap and promote diversity in the applicant pool by including encouraging and informative language around qualification requirements in job ads. We do so using a large-scale, "reverse audit study" field experiment where we randomize the content of job ads and observe job seeker behavior. Specifically, we established a non-profit firm to act as an intermediary in the job search process. This firm reposts real job ads and collects information from job seekers interested in applying. We randomize whether we encourage people to apply even if they don't meet all of the listed quali-

[0] Department of Economics, University of California San Diego. Authors can be reached at abonheur@ucsd.edu and teastmond@ucsd.edu. This research is conducted under UCSD IRB #806291. Pre-registration of the pilot can be found on AsPredicted #148892. This work has been supported (in part) by Grant #2301-41761 from the Russell Sage Foundation and by the W.E. Upjohn Institute for Employment Research. Any opinions expressed are those of the principal investigator alone and should not be construed as representing the opinions of either funder. We also thank the Yankelovich Center for Social Science Research and the UC San Diego Department of Economics Diversity and Research fund for generous financial support with this research.

fications and whether we inform them that companies routinely hire individuals who do not have all qualifications. Preliminary results show that this light touch intervention nudges more people into applying. Further analysis will study how the intervention changes perceptions of the hiring process and whether it has larger impacts on women, individuals from underrepresented racial minority groups, and people with non-traditional employment backgrounds.

3.1 Introduction

Despite years of policy and revised corporate practice intended to correct inequality in the hiring process, gaps persist for women and individuals from underrepresented racial minority groups at every level of the hiring process. Past news articles (Clark, 2014; Mohr, 2014; Sakowitz, 2018; Zucker, 2020) and academic work have shown that qualified workers in underrepresented groups are less likely to apply to jobs in the first place (Abraham, Hallermeier, & Stein, 2023; Castilla & Rho, 2023; Chaturvedi, Mahajan, & Siddique, 2021; K. B. Coffman, Collis, & Kulkarni, 2019; Flory, Leibbrandt, & List, 2015; Flory, Leibbrandt, Rott, & Stoddard, 2021; Fluchtmann, Glenny, Harmon, & Maibom, 2024; Gaucher, Friesen, & Kay, 2011; Leibbrandt & List, 2018; Linos, 2017; Llinares-Insa, González-Navarro, Córdoba-Iñesta, & Zacarés-González, 2018; Pager & Pedulla, 2015; Wille & Derous, 2017). This paper focuses on this first step of the hiring process, seeking to understand who applies to jobs, and whether we can encourage job seekers from historically disadvantaged groups to apply for more and better jobs.

In particular, we explore the true extent of the application gap for underrepresented workers across a variety of industries and whether changing the language surrounding the listed qualifications in the job ad can induce more of these workers to apply. This is motivated in part by a finding from a Hewlett Packard internal report that says “men apply for a job promotion when they meet 60% of the qualifications, but women apply only if they meet 100% of them” (Clark, 2014; Mohr, 2014; Sakowitz, 2018). Furthermore, women are 16% less likely to apply to a job after viewing it, apply to 20% fewer jobs overall, and are less likely to apply for ‘stretch roles’ (Tockey & Ignatova, 2019). Taken together, these findings would not be concerning if companies only ever hired workers that met every qualification from the job ad. Though companies are very thoughtful about the qualifications they put in their listings, there is no way to fully describe a perfect candidate with a short list of possible experiences and traits. This is not only true in theory, but in practice as well: Half (2019) finds that 84% of companies are willing to hire and train employees up where needed and 62% of employees have been offered a position even when they did not satisfy all listed required qualifications. By not applying in the first place, workers who would ultimately be an excellent match for the prospective employer remove themselves from the applicant pool, even though companies frequently extend offers to similar people.

Our study evaluates whether high quality job seekers can be induced to apply for jobs by randomizing the language around the list of qualifications in job ads. We do this using a non-profit corporation that we set up, the Job Connections Project (JCP).[1] The Job Connections Project re-posts open job ads from other companies for full time positions and advertises them on various online job boards, thus resembling a recruiting firm. Job seekers who click through our job ads from online job boards are randomly assigned to treatment or control versions of the job ad, then shown a brief survey, before being routed to the actual application. If assigned to the control version, nothing is changed about the job ad.[2] The various treatment arms add language encouraging the candidate to apply even if they do not meet all qualifications, informing them that companies routinely hire people without all listed qualifications, and informing them that women are less likely to apply without all qualifications. This step is interposed in the standard job application process and is minimally disruptive to their job search, but to compensate them for their time we offer several free services to help job seekers.[3] This design is similar in spirit to resume audit studies, where realism is preserved by randomizing the content of resumes sent to real hiring managers. Our 'reverse' audit study instead randomizes content in real job listings, which allows us to observes job seeker behavior towards real jobs, measure real-time perceptions of the job, and understand job seeker attitudes towards the job market more broadly.

Crucially, our randomization does not change any of the listed qualifications in each listing. Keeping the listed qualifications fixed is important and relevant since companies thoughtfully choose the qualifications in job postings and altering qualifications is not feasible for many roles. Additionally, previous research has found that changing the qualifications can change the perceived

[1] Our company serves two major goals: help job seekers better match with jobs and study the labor market to improve the search and match process broadly. The Job Connections Project is incorporated as a non-profit in the State of California.

[2] Except for de-identifying the company name, like recruiting firms, to reduce noise from company reputation.

[3] This includes resume feedback and an application autofill tool.

rigor of the role and is not guaranteed to increase applicant diversity (Abraham et al., 2023). We focus on encouragement and information about the hiring process because, while some have suggested that this difference in applying behavior is a professional confidence problem (Rojas, 2021; Sandberg & Scovell, 2013), differing perceptions about the hiring process are a more likely culprit. Mohr (2014) surveys people about their application behavior and found that common reasons for not applying to a job include not wanting to waste time if they do not have a chance, fear of failure, and simply following the rules, with women being more likely to report these feelings. In other words, women "thought that the required qualifications were ... well, required qualifications. What held them back from applying was not a mistaken perception about themselves, but a mistaken perception about the hiring process." Not only does this impact women, but there is some evidence that other historically disadvantaged groups such as racial/ethnic minorities may be impacted as well (Avery & McKay, 2006; Wille & Derous, 2017).

Initial pilot results suggest that job seekers exposed to treatment are more likely to continue toward applying for the job, and furthermore that this is driven primarily by the simplest treatment arm that only encourages them to still apply if they do not meet all listed qualifications with no other information. We are still collecting data for the pilot, but will soon also be able to speak to the demographic composition of prospective applicants,[4] their quality, their perceptions of the job ad itself, their broader perceptions of the job market, and occupational differences. These data are collected from engagement with JCP's website, survey answers, and job seeker resumes.

We hypothesize that women, BIPOC individuals (Black, Indigenous, and People of Color),[5] and people who are Skilled Through Alternative Routes rather than a bachelor's degree (STARs)[6] will be more affected by encouragement and information about 'required' qualifications. We further hypothesize that the treatment will nudge qualified individuals (e.g. those who meet 7 to 9 out

[4] Throughout this paper we call job seekers 'prospective applicants' if they click through our site indicating that they intend to apply, though we do not actually observe whether or not they do ultimately apply to the job.

[5] Pager and Pedulla (2015) find that Black people cast a wider net in their search relative to similarly situated White people, as an adaptation to deal with labor market discrimination. Due to this fact, that Black workers may already apply to a large breadth of roles, we may not see a large change for Black job seekers.

[6] I use this language since Opportunity @ Work, a non-profit organization that works to advance economic mobility, refers to these individuals as Skilled Through Alternative Routes or STARs.

of 10 qualifications) into applying. Overall, this means that we expect a change in the demographic composition of the applicant pool with an equal or higher number of applicants who could perform the job well. This project aims to improve outcomes for those traditionally disadvantaged in the labor market.

Understanding who applies for jobs under different ways of presenting qualifications will both (i) help people who have been historically disadvantaged in the labor market and (ii) help companies with their hiring initiatives. In the two-sided job match process, employers choose how they present openings in job ads, and potential employees make application decisions. Just as employees navigate the hiring process, many employers struggle to recruit diverse applicants (Kessler, Low, & Sullivan, 2019). Job ad language is one of the common themes in articles about how-to-attract-a-diverse-workforce,[7] but no rigorous test has been performed yet. This project will provide valuable insight into how job seekers behave in the current job market. It has the potential to provide employers with a tool to attract a more diverse applicant pool and mitigate application gaps.

There are three main contributions of this work. The first is that we observe real job seeker behavior towards real, full-time positions. Our ability to preserve realism expands upon previous literature that has used fake job ads (Burn, Firoozi, Ladd, & Neumark, 2022) or short-term positions (Castilla & Rho, 2023; Del Carpio & Guadalupe, 2021). Additionally, we are able to study effects across companies and occupations while abstracting away from company-specific reputation (Abraham et al., 2023).[8] Second, our survey design enables us to study mechanisms behind application decisions. Specifically, we measure perceptions of the hiring process, confidence, feelings of wasting time, interest in the role, self-assessments versus how they think hiring managers will view their ability, and more. The third contribution of this work is that the intervention is a partial solution that is easily implementable across a variety of contexts. We keep listed qualifications the same without removing or changing requirements, which is less effort from companies

[7] See, for example https://bit.ly/3AjOhVn, https://bit.ly/3dxmtUG, and https://bit.ly/3JXLPHj.

[8] Abraham et al. (2023) randomizes information for corporate jobs for Uber. Since Uber is a well known company, people likely view their job ads through the lens of their perception of them.

and is less likely to affect the perceived rigor of the role (Abraham et al., 2023). If treatment is effective in attracting different types of workers who could perform the job well, then this is a simple, readily adoptable policy tool for firms. No test scores or additional certifications are required to encourage women to apply (K. B. Coffman et al., 2019). Given that changing a sentence or two in a job ad is low-cost from a firm's perspective, the resulting policy recommendation may have a high likelihood of being adopted.

In addition to the strong policy relevance, this project is at the frontier of methodologies to study inequality. We have created a new infrastructure to run reverse audit study experiments and gather original data.

Anticipated Contribution to the Literature

We are not the first to study application decisions of women and other traditionally disadvantaged groups in the labor market, but we innovate in realism, breadth, and depth.

Disparities arise throughout the hiring pipeline; there are gaps in who applies, who gets interviews, how interviews are rated, promotions, and more (Abraham et al., 2023; Avery & McKay, 2006; Benson, Li, & Shue, in press; Bertrand & Mullainathan, 2004; Burn et al., 2022; Chaturvedi et al., 2021; Clark, 2014; Goldin & Rouse, 2000; Kline, Rose, & Walters, 2022; Mohr, 2014; Sakowitz, 2018; Tockey & Ignatova, 2019; Wille & Derous, 2017). This paper focuses on application gaps and its connection to perceptions about the hiring process.

We contribute to the social sciences literature about the existence, causes, and remedies for application gaps by gender and race (Barbulescu & Bidwell, 2013; Ekstrom, 1981; Llinares-Insa et al., 2018; Pager & Pedulla, 2015; Reskin & Bielby, 2005). Most notable for this project, women are less likely to apply for a job unless they have all of the qualifications listed, whereas men apply with a fraction of the qualifications (Mohr, 2014; Sakowitz, 2018). A Gender Insights Report from LinkedIn (Tockey & Ignatova, 2019) similarly found that women are 16% less likely to apply to a job after viewing it, apply to 20% fewer jobs overall, and are less likely to apply for 'stretch' roles. At first, people thought this was likely a professional confidence problem (Rojas, 2021; Sandberg & Scovell, 2013), but more recent research has shown it is more about their beliefs about the hiring "rules" (Mohr, 2014; Zucker, 2020). Further, there is little research into application behavior of non-binary and other gender minority individuals.[9] Aksoy, Exley, and Kessler (2024) shows that gender minority individuals exhibit less confidence and less favorable self-evaluations, and we hope to be the first to document application gaps, depending on sample size. Relatedly, African American job seekers cast a wider net in their job search, as a response to labor market discrimination (Pager & Pedulla, 2015). This evidence demonstrates that perceptions about the hiring manager and hiring process are crucial components of application decisions.

These application gaps are likely inefficient because people are sometimes hired into stretch roles. Half (2019) found that 84% of companies are willing to hire and train up and 62% of employees have been offered a position when they were 'underqualified'.

Previous studies have shown that altering aspects of job advertisement language can inadvertently turn away or attract certain types of job seekers. For example, job applicants are affected by stereotyped language or gendered skills (Burn, Button, Menguia Corella, & Neumark, 2019; Burn et al., 2022; Chaturvedi et al., 2021; Kuhn, Shen, & Zhang, 2020), the inclusion of diversity and EEO statements (Dover, Major, & Kaiser, 2016; Flory et al., 2021; Gaucher et al., 2011; Hurst, 2022; Leibbrandt & List, 2018), information about the success of marginalized groups (Choi, Pacelli, Rennekamp, & Tomar, 2022; Del Carpio & Guadalupe, 2021), highlighting personal benefits (Linos, 2017), and the framing of factual information about the job (L. C. Coffman, Featherstone, & Kessler, 2017; Dal Bó, Finan, & Rossi, 2013; Flory et al., 2015; Gee, 2019; Marinescu & Wolthoff, 2020; Samek, 2015). These papers answer different questions than ours, but suggest that less strict language around required qualifications could be effective in modifying behavior. On the other hand, Castilla and Rho (2023) find negligible effects of the gendering of job postings or of the recruiter, implying that small language changes may not be very important for some online job postings.

The closest work is Abraham et al. (2023), which finds that making listed qualifications less demanding encourages people to apply and reduces the skill gap between male and female

[9] For example, transgender individuals and people who identify as genderqueer.

applicants, but also changed perceptions of the rigor of the role. We complement this work by keeping the listed qualifications fixed while changing the wording around the qualifications. This distinction is important since companies choose which qualifications are needed for job postings, so removing or altering the qualifications themselves is not feasible for many roles. Similarly, K. B. Coffman et al. (2019) find that adding strict test score cutoffs reduces ambiguity around qualifications and reduces the gender gap in willingness to apply, but test scores are not readily available for most roles.[10]

This paper innovates on previous audit study designs by proposing an infrastructure for a 'reverse' audit study which focuses on job seeker response instead of firm behavior. Audit studies have been used to research discrimination from firms by sending fake resumes to companies (Gaddis, 2018). These have found gaps in callback rates by race (Bertrand & Mullainathan, 2004; Kline et al., 2022; Quillian, Pager, Hexel, & Midtbøen, 2017); age (Farber, Silverman, & von Wachter, 2016); gender (Bohren, Imas, & Rosenberg, 2019); religion, attractiveness, sexual orientation, and more (Bertrand & Duflo, 2017).

Most other research in this space have either used fake job ads, studied short-term work such as internships, or partnered with one company.[11] Burn et al. (2022) uses a similar concept and method to our reverse audit study design, and we innovate on their approach to ensure realism and reproducibility. They answer a different question and encountered a number of implementation difficulties due to posting fake job ads.[12] We learn from their work and create an experimental design that not only dodges many of their implementation hurdles, but also posts ads for full-time jobs across firms, meaning that we can establish realistic effects independent from a company's reputation.

[10] For instance, there is no test or score cutoff to define someone as a 'strong communicator'. Additionally, even there are tests for some skills, most jobs are not on a standardized platform such as UpWork, the one used in K. B. Coffman et al. (2019).

[11] For examples, see Abraham et al. (2023); Burn et al. (2022); Del Carpio and Guadalupe (2021); Flory et al. (2021). One paper that does use real job seekers, real jobs, and across firms is Kuhn et al. (2020). They study explicit requests for applicants of a particular gender in a Chinese context. Alternatively, we study language surrounding job qualifications that imply varying levels of strictness in the US context.

[12] These include technical difficulties such as having to have phone numbers and research assistants for every fake job ad they posted, which limited their sample size.

Overall, our paper expands upon the current literature by studying job seekers within their job search process, across types of roles, and with an in-depth study of mechanisms.

3.2 Methodology

We design a large-scale 'reverse' audit study field experiment where we randomize the content in real job ads to explore how job seekers respond to information in jobs ads about the listed qualifications. This methodology is in the spirit of the large set of resume audit studies in the literature (e.g. Bertrand & Mullainathan, 2004; Kline et al., 2022). To accomplish this we established a non-profit company, the Job Connections Project (JCP), that acts as an intermediary within the regular job application process to preserve realism. The JCP serves two primary purposes simultaneously. First, it serves as a platform to post jobs to learn about real job seeker responses to the content of job ads and the labor market more broadly. Second, it provides services to those job seekers to compensate them for the small incursion into their time and to help them in their job search.

Using data from the Current Population Survey in 2023, we first identify 10 occupational categories with differing levels of baseline representation of women, Black workers, and Hispanic workers for inclusion in our study. These occupations are shown in Table 3.1. With these occupations in hand, we conduct our experiment using our platform as follows.

We post job ads in occupation/job title and geographical clusters. We find and post 3-5 open job listings at a time for the chosen category in a randomly chosen city in the US. All jobs are for full-time positions and have a clear list of qualifications.

For job listings used in the study, we remove identifying information for the hiring company,[13] repost the job ad on the JCP website, and advertise the positions on online job boards. When advertising on multiple large job listing websites, we do not include the qualifications in the preview of the job listing included on the external job board to prevent contamination of the

[13] We remove identifying information for the hiring company to limit reputation-related confounding factors. This is common practice among recruiting firms and so will not be out of the ordinary.

Table 3.1: **Occupations used in the study and their fraction of representation**

	Fraction of employed persons who are:		
		Black/African	**Hispanic/**
Occupation	**Women**	**American**	**Latino**
Engineers, various	Low	Low	Low
Credit counselors and loan officers**	Median	Median	Median
Human resources manager	High	High	High
Preschool and kindergarten teachers	High	High	Median
Elementary and middle school teachers	High	Median	Median
Secondary teachers	Median	Low	Low
Postsecondary teachers	Median	Median	Low
Public relations specialists	High	Median	Low
Wholesale and retail buyers	Median	Low	High
Training and Development specialists	Median	High	Low
Computer/data/software occupations**	Low	Median	Low
Mental health and guidance counselors	High	High	Median

Low means the fraction of that identity is less than the 36th percentile of that group's participation across all occupations. Median is between the 36th and 63rd percentile, while High is above the 64th percentile.

**Occupations used in pilot.

treatment. Job seekers then come across our ads on these sites and click through if they are interested in applying. Once they click through, they are redirected to our company's website, https://jobconnectionsproject.org/, and are randomized by IP address to see either the control or one of the treatment versions of the job ad and told that we are a non-profit trying to learn about the job market (Figure 3.1).

More specifically, one can think of the Job Connections Project as a helpful intermediary in the job search process. A typical job search involves job seekers finding a job ad on an online job board, clicking "Read More/Continue", being redirected to the hiring company's careers page where they read the full job ad, clicking "Apply Now", and finally, applying for the job. Jobs that are posted by the JCP have additional steps in the middle of the process. When job seekers click "Read More/Continue" on job ads posted by the JCP, they are redirected to the Job Connections Project website where they read the full job ad with language randomization. They then can click "Continue", are offered JCP's services, asked to complete an optional 2-minute survey, then click

The Job Connections Project is a non-profit company that advertises open positions for other companies. Please read the hiring company's job ad below, then click 'Continue'.

Figure 3.1: **Example of what job seekers see when read job ad on JCP's website**

"Apply now on company website". At this point they are debriefed about being part of a study, re-directed to the hiring company's website, and advised to re-read the full job ad. At this point they can continue and apply as normal.

Our control and treatment settings describe the listed qualifications using whatever heading the original job post had.[14] In the control arm, the job ad is otherwise completely unaltered.[15] The first treatment (which we call hereafter the 'Encourage' treatment) includes a blurb saying: "Don't meet every single requirement? If you're excited about this role but your past experience doesn't align perfectly with the job description, we encourage you to apply anyways. You may be just the right candidate for this role." This blurb is embedded in the job ad at the end of the qualifications section. The second treatment ('Encourage + Hiring Info' treatment) is the same, except we add the statement "Most companies routinely hire individuals who lack some of the stated required skills" to the blurb. The third treatment arm ('Encourage + Hiring Info + Women Info' treatment) includes the "Most companies routinely hire. . . " statement, along with "Studies have shown that women are less likely to apply to jobs unless they meet every single qualification." The fourth treatment arm matches the third ('Encourage + Hiring Info + Women Info'), but the blurb comes from the JCP. In this fourth treatment ('Encourage + Hiring Info + Women Info from JCP'), instead of being

[14] Examples include: Qualifications, Required Qualifications, Knowledge and Skills, Required Skills/Competencies, Experience, and similar.

[15] Except for removing the identifying information of the hiring company, as noted earlier.

embedded in the job advertisement itself, the blurb appears in a box under the text "Tip from the Job Connections Project:".[16] Overall, treatment interventions are a variation of the following language:

> "Don't meet every single requirement? Studies have shown that women are less likely to apply to jobs unless they meet every single qualification, but most companies routinely hire individuals who lack some of the stated required skills. So if you're excited about this role but your past experience doesn't perfectly align with the job description, we encourage you to apply anyways. You may be just the right candidate for this role."

After reading the full job ad with randomized intervention language, interested job seekers click 'Continue' and are presented with the free services offered by the JCP (resume feedback and access to a Chrome extension that automatically fills out their information on other job applications)[17] and a short survey. The services are offered first, directly succeeded by a survey about their views toward this role to help us learn about the job match process. This 2-minute survey asks for their likelihood of applying to the position, likelihood of accepting the position if offered, current employment status, basic demographics (race, gender identity), and a set of questions about how they fit with the role. These last questions give us self-assessments of quality and fit for the specific role. Specifically, we elicit how well they believe they could perform the job, their own perceptions of the proportion of qualifications they meet, whether they have other relevant skills that weren't specifically asked for, their guess of whether the hiring manager will recognize their potential, and their guess as to their chances of being invited for an interview. The last part of the survey elicits perceptions of the job ad itself. We ask agree-disagree Likert scale questions of how they view the following: the hiring company's leniency when reviewing candidates relative to other companies, whether they will be wasting their time if they apply to this job, whether they want to apply to more 'stretch' roles, how people should apply when they meet most but not all of listed qualifications, and whether companies in general stick to required qualifications.[18]

Survey answers combined with how they interact with our website allow us to determine

[16] For examples of what this looks like in a job ad, see Appendix Figures 3.A.1 and 3.A.2.

[17] The Chrome Extension is available in the Chrome Web Store at the following link https://chromewebstore.google.com/detail/job-connections-project-j/apjpojgndgefpkgpmkifhhgmepkgieki.

[18] To see the full survey instrument, see Appendix 3.B.

whether the presentation of required qualifications can affect perceptions and the likelihood that people apply for jobs they could perform well.

The field experiment design allows us to measure a few different outcomes of how application behavior is affected by treatment using interaction with our website, survey answers, and job seeker resumes. The outcomes are the number, composition, and quality of job seekers. The quantity of likely applicants is tracked using both self-reported likelihood of applying and click-through rate to the hiring company's webpage. Specifically, we measure the proportion of those who view a job ad on our website and click 'Continue', and the proportion that clicks 'Apply now on Company website' versus not continuing or selecting 'No longer interested, show me other jobs'. We can measure click-through rates for all individuals who encounter our website and self-reported likelihood of applying for survey respondents. While we do not observe actual application rates,[19] these two proxies will be proportional and able to detect treatment effects. Demographic and employment status characteristics are collected in the survey to identify composition effects. Self assessments from the Madlib style question of how they fit with the job being advertised, including how well they believe they could perform the job and whether they think the hiring manager will see their potential give us some measures of quality. We also measure quality using listed employment history, education, and skills from individuals who share their resume with us. We believe that our treatment will lead to more people applying across the skill distribution. While people who are less likely to perform the job well may also apply more, we predict that the number of excellent candidates that would perform the job well will also increase, giving employers a larger desired applicant pool.

Importantly, we care about who is most affected by job ad language, and whether the intervention can encourage more women, Black individuals, Hispanic people, and or people with non-traditional education/employment backgrounds to apply. Therefore we will explore treatment effects by race, gender, education, and employment status. We know that women tend to take

[19] Those who select 'Apply on Company Website' will be redirected to the original job ad on the hiring company's website where they will be able to apply for the open job. At this point we will not interact with or receive information from the job seeker any longer. Those who select 'Not interested, show me other jobs' will be redirected to a page on our website that lists multiple related jobs.

themselves out of the running, and we hope to find evidence that more women have intent to apply when presented with less strict language surrounding required qualifications.[20]

The strength and believability of the encouragement language may vary depending on the industry and/or current amount of representation of women and BIPOC individuals. For this reason, we select a handful of occupations with varying levels of current diversity, such that we can identify heterogeneous effects along this margin. This is important for understanding when effects translate to other contexts. The occupations are chosen using 2023 data from the Current Population Survey (CPS).[21] Table 3.1 shows the 10 chosen occupations and whether their gender and race/ethnicity fractions are average, low, or high. Due to the smaller size of our pilot, our pilot includes 2 of these occupations: Data Science/Computer Occupations and Loan Officer/Credit Counselor.[22]

Further, our setup allows us to unpack mechanisms behind behavior. We will use the survey to examine whether people's perceptions move in conjunction with their application behavior. As stated above, we gather data on perceptions of own fit, how others view them as a candidate, if this job is worth applying to, and more. If people who are nudged into applying are also less likely to think they are wasting their time when they view treated job ads, then fear of wasting time is a factor driving application gaps that can be influenced by job ad language.

We are at the frontier of studying real job seekers in relation to actual full-time jobs across multiple employers. The contributions of this project include: (1) We quantify the impact of language surrounding job qualifications, without changing the qualifications themselves, on applicant pool diversity in a real job setting. (2) We explore mechanisms behind the results using survey evidence, including feelings of wasting time, perceived rules, etc. (3) We provide concrete policy recommendations to attract a wider group of applicants and reduce gaps. (4) We innovate on previous methods by developing a reverse audit study methodology and founding the Job Connections

[20] We anticipate that we may also find larger treatment effects for individuals who are currently employed, since they can be more discretionary with respect to job ad language than currently unemployed individuals.

[21] The CPS detailed occupation data by race/ethnicity and sex can be found at https://www.bls.gov/cps/cpsaat11.htm.

[22] The data science/tech was chosen over engineers, who also have low representation among women and non-White individuals, since there have been layoffs around the time of our pilot, such that our ability to reach job seekers is higher.

Project non-profit. This setup creates a novel dataset, allows analysis at-scale across industries and job types, and will enable future work to deepen understanding of job seeker behavior.

3.3 Results

This section describes preliminary results from the pilot between April 4, 2024 and May 17, 2024, and are subject to change. In these 6 weeks, the JCP posted 59 jobs and advertised 26 of them on various online job boards. These consisted of 13 data scientist and 13 credit counselor/loan officer positions. These job ads garnered 518 observations, where each observation is a unique person and job ad pairing. This is from 484 unique job seekers who have come across our website.[23] Of these 518 people looking at a job ad on our website, 186 (36%) clicked continue at the bottom of the job description, 99 (19.1% of the full sample, or just over half of those who clicked continue) filled out some part of the survey or uploaded their resume, and 119 (23%) clicked 'Apply now on company website'.

Roughly half of job seekers were randomized into seeing control ads (245, 47.3%), while the other half saw a treatment ad (273, 52.7%). Within the treated individuals, they are equally distributed among each version of Treatment Encourage, Encourage + Hiring Info, Encourage + Hiring Info + Women Info, and Encourage + Hiring Info + Women Info from the JCP. Note that this randomization is occurring at the IP and job ad level, meaning that we can make comparisons across treatment or within each open position.

Preliminary results show that the treatment language embedded in the job ad may encourage applications, while the language coming as a 'tip from the Job Connections Project' may decrease applications. There are 3 ways we measure application likelihood; clicking continue, self-reporting likelihood of applying; and clicking 'Apply now on company website'. The first outcome is whether treatment language (encouragement and information around required qualifications) induces more people to click 'Continue'. This would be the first step to show more intent

[23] This means that 93.4% of job seekers interact with the JCP once, looking at only one JCP job ad, while a small fraction also look at other jobs posted by the JCP.

to apply to the position.

Individuals who are shown job ads with the 'Encouragement' treatment variation are about 5 to 9 percentage points (13-25%) more likely to click continue after reading the job ad (Table 3.2).[24] Column 3 shows results without job ad fixed effects, while column 4 adds job ad fixed effects, giving us a within-job-ad effect estimate. The job ad fixed effects capture variation in job characteristics that influence application behavior.[25] While the effect of the encouragement treatment is not statistically significant, we expect more power as the sample size grows.

Interestingly, job seekers who see treatment 4, the encouragement and information intervention listed as a tip from the JCP, are *less* likely than control individuals to click continue (Table 3.2).[26] Specifically they are 9 to 13 p.p. (26-37%) less likely to click continue. This makes the pooled treatment effect negligible in the current full sample (Columns 1 and 2).

Table 3.2: **Job seekers in the 'Encourage' treatment are more likely to click continue, while those in the 'Encourage + Info from the JCP' are less likely to click continue**

Clicked Continue	(1)	(2)	(3)	(4)
Treated	0.015	-0.009		
	(0.034)	(0.035)		
Treatment Encourage			0.088	0.047
			(0.058)	(0.054)
Encourage+Hiring Info			-0.009	-0.018
			(0.043)	(0.040)
Encourage+Hiring+Women Info			0.057	0.036
			(0.064)	(0.060)
Encourage+Hiring+Women Info (JCP)			-0.092*	-0.129***
			(0.047)	(0.043)
N	518	518	518	518
N clusters	32	32	32	32
Job Ad FE		Yes		Yes
R^2	0.000	0.175	0.010	0.185

Standard errors in parentheses, clustered by job ad. Results are similar if use robust standard errors.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

[24] Relative to the baseline of 35.1% of control individuals who click continue. Related logistic estimates can be found in Appendix Table 3.A.1.

[25] Future iterations will also add a version without job ad FE that controls for various characteristics about the job ads, to better understand application decisions.

[26] Investigation is ongoing to determine if this is because it means that job seekers are less likely to think the hiring company is more lenient (i.e. perceptions about the open position), or if it is a salience effect about the JCP not being the hiring company.

There appears to be a larger increase for credit counselor/loan officer relative to data science positions, suggesting that occupation and/or baseline representation matters (Table 3.3). The credit counselor occupation has medium representation of women, Black, and Latinx individuals and a larger increase in the likelihood of clicking continue among treated individuals, while data scientists are currently less diverse. Both occupations continue to have a decrease in the likelihood of clicking continue for those in treatment 4 with the tip from the Job Connections Project.

Table 3.3: **Job seekers looking at credit counselor/loan officer positions are more encouraged by treatment to click continue, relative to data scientist roles**

	Data Scientist				Credit/Loan Officer			
Clicked Continue	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treated	-0.006	-0.019			0.108	0.021		
	(0.035)	(0.033)			(0.118)	(0.136)		
Treatment Encourage			0.097	0.049			-0.017	0.009
			(0.059)	(0.049)			(0.225)	(0.297)
Encourage+Hiring Info			-0.038	-0.028			0.118	0.000
			(0.044)	(0.044)			(0.113)	(0.098)
Encourage+Hiring+Women Info			-0.016	-0.008			0.361*	0.222
			(0.054)	(0.051)			(0.203)	(0.220)
Encourage+Hiring+Women Info (JCP)			-0.079	-0.108**			-0.148	-0.225
			(0.055)	(0.040)			(0.124)	(0.171)
N	423	423	423	423	95	95	95	95
N clusters	16	16	16	16	17	17	17	17
Job Ad FE		Yes		Yes		Yes		Yes
R^2	0.000	0.167	0.009	0.174	0.014	0.202	0.106	0.262

Standard errors in parentheses, clustered by job ad. Results are similar if use robust standard errors.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Another measure of intent to apply is the number who click ‘Apply now on company website’ on the next page. Conditional on clicking continue, treated individuals are equally likely to click ‘Apply now’ as control individuals. As our sample grows, we could also look at self-reported answers to “How likely are you to apply to this job?” and whether they vary by treatment.

Our main interest is whether we can disproportionaly encourage more women and other traditionally disadvantaged groups into applying more. To study this, we turn to the sample of job seekers who filled out some portion of the survey or uploaded their resume, providing us more information about their demographic and other characteristics.

Ninety-nine job seekers filled out some portion of the survey or uploaded their resume, with most of them filling out every question of the 2-minute survey. Descriptive statistics of those who chose to share their information is shown in Table 3.4. Overall, we see substantial variation in our pool of job seekers.

Table 3.4: **Descriptive statistics of job seekers who shared additional information show substantial heterogeneity**

	Summary
N	99
Gender	
Genderqueer or gender fluid	1 (1.2%)
Man	67 (81.7%)
Woman	14 (17.1%)
Race	
A race/ethnicity not listed here	2 (2.4%)
Asian or Pacific Islander	38 (45.8%)
Black or African American	12 (14.5%)
Hispanic or Latino	14 (16.9%)
Multiracial or Biracial	2 (2.4%)
White or Caucasian	15 (18.1%)

	Summary
Highest Education	
High School	5 (6.0%)
Associate's or some college	5 (6.0%)
Bachelor's degree	39 (47.0%)
Graduate degree	34 (41.0%)
Relevant Experience (years)	7.281 (5.796)
Employment Status	
Employed (full-time)	41 (48.8%)
Employed (part-time)	12 (14.3%)
On temporary leave	1 (1.2%)
Unemployed	30 (35.7%)

In our sample of survey respondents we also observe a few patterns by gender and race (Table 3.4). We observe more men than women who are intending to apply to these positions, which makes sense given the current gender make-up of these occupations. We have a variety of racial groups, a range of years of experience, and both people who are currently employed and unemployed. We have mostly people with college degrees or higher, which fits with the types of job ads used in the pilot.

Since treatment assignment is random, we can use the composition of survey respondents to determine who was encouraged by treatment. Identification by this method relies on the assumption that the likelihood of filling out the survey conditional on clicking continue is not differentially affected by treatment. This is plausible, especially for treatments 1, 2, and 3 where intervention language is embedded in the job ad and appears to job seekers similar to the control.[27]

The current gender sample is too small to effectively answer the pertinent question of whether women are differentially encouraged by the treatment language. Table 3.5 shows that our current survey sample has only 7 women in the control and 7 in the treatment.

[27] Further work is ongoing to more concretely answer this question.

Table 3.5: **Current survey sample has low representation of women**

	Control		Treated	
	N	(%)	N	(%)
Genderqueer	1	2.6	0	0.0
Man	30	79.0	37	84.1
Women	7	18.4	7	15.9
Total	38	100	44	100

Underrepresented racial groups may be more affected by the encouragement and information treatment language. Using data from those who fill out the survey, Table 3.6 shows how the racial composition differs in the treatment relative to the control. While not yet significant (Appendix Table 3.A.2), Black, Latinx, and Multiracial people appear to be more likely to click continue in the treatment group. Asian individuals are similarly likely to click continue, and White people may be slightly discouraged (although the sample is too small to definitively say).

Table 3.6: **Black, Latinx, and Multiracial individuals may be more affected by treatment encouragement language**

	Control		Treated	
	N	(%)	N	(%)
Asian	19	48.7	19	43.2
Black/Latinx/Multi	11	**28.2**	19	**43.2**
White	9	23.1	6	13.6
Total	39	100	44	100

Tables 3.7 and 3.8 show our potential mechanism questions about own confidence, self-assessment of own skills, perceptions of the hiring manager, perceptions of this company, and more general assessments of the hiring process. There is substantial variation in most categories, meaning these variables have the potential to provide insights into behavior. There is strong alignment between showing intent to apply by clicking continue and self-reporting a high likelihood of both applying to the job and accepting the job if offered. There is also variation in perceptions of how lenient the company is, whether applying would be a waste of time, whether they want to apply for more stretch roles, and more.

Table 3.7: **Variation in self-reported likelihood of applying and how well they align with the position along a number of dimensions**

	Summary
N	99
Likelihood apply	
Very Likely	74 (83.1%)
Likely	10 (11.2%)
Equally Likely and Unlikely	4 (4.5%)
Very Unlikely	1 (1.1%)
Likelihood accept	
Very Likely	75 (88.2%)
Likely	9 (10.6%)
Equally Likely and Unlikely	1 (1.2%)
Perform on job	
I could do the job well	63 (85.1%)
the job would be a stretch, but I could learn quickly	9 (12.2%)
this job would be a real challenge	2 (2.7%)
Fraction of qualifications	
meet all	44 (57.1%)
meet most	29 (37.7%)
meet some	4 (5.2%)
Other relevant skills	
many	63 (82.9%)
some	13 (17.1%)
Hiring manager see my potential	
would	74 (97.4%)
may or may not	1 (1.3%)
would not	1 (1.3%)
Likelihood get interview	
very likely	56 (71.8%)
likely	18 (23.1%)
equally likely and unlikely	2 (2.6%)
unlikely	2 (2.6%)

Table 3.8: **Variation in perceptions of job seekers about the company, role, and application process more generally**

	Summary
N	99
This company lenient	
Strongly disagree	8 (11.6%)
Disagree	11 (15.9%)
Neither agree nor disagree	27 (39.1%)
Agree	17 (24.6%)
Strongly agree	6 (8.7%)
Applying would be waste of time	
Strongly disagree	36 (52.2%)
Disagree	17 (24.6%)
Neither agree nor disagree	11 (15.9%)
Agree	3 (4.3%)
Strongly agree	2 (2.9%)
I want to apply to more stretch roles	
Strongly disagree	6 (8.8%)
Disagree	7 (10.3%)
Neither agree nor disagree	15 (22.1%)
Agree	28 (41.2%)
Strongly agree	12 (17.6%)

	Summary
People should apply with most qualif.	
Strongly disagree	5 (7.2%)
Disagree	1 (1.4%)
Neither agree nor disagree	10 (14.5%)
Agree	27 (39.1%)
Strongly agree	26 (37.7%)
Companies sticking req'd qualif. less	
Strongly disagree	8 (11.4%)
Disagree	3 (4.3%)
Neither agree nor disagree	25 (35.7%)
Agree	22 (31.4%)
Strongly agree	12 (17.1%)

3.4 Discussion & Policy Implications

Findings from this study have clear and useful policy implications for all parties involved in the job search-and-match process. This paper will either provide an evidence-backed solution to mitigate application gaps or insights into why it doesn't work and proposals for other solutions.

If using accessible language around qualifications has the hypothesized effect, then this intervention provides more opportunities for job seekers traditionally disadvantaged in the labor market. The slight interventions can be used to improve equality in overall placement outcomes for women, BIPOC workers, and STARs. The more that companies adopt these tested job ad modifications, the more job seekers who are influenced by the changes can benefit by being encouraged to give themselves a chance at 'stretch roles'.[28] Job seekers could also infer which firms care about diversity through the use of language encouraging a broader range of people to apply.

Results will inform best practices for companies to attract a diverse set of applicants that can be implemented without financial, time, or capacity barriers. These minor changes to job ads are a low-cost intervention from the firm's perspective.[29] At the same time, making these changes will benefit companies through more diverse desired applicant pools. Further, the experimental language around job qualifications can be applied to all types of roles for any industry. We will analyze heterogeneous effects to tailor recommendations.

Policymakers at various levels can integrate insights into policies. Online job boards could create guides for companies that want to advertise open positions on their platform. Hiring consultants and Diversity & Inclusion (D&I) officers can advocate for use of this empirically-supported tool to mitigate application gaps. This benefits companies since they will see a larger range of applicants, which could lead to better matches since fewer people would be left out of the applicant pool.

[28] Applying to 'stretch roles' is a good thing because sometimes people who do not meet all qualifications are hired. In fact, 62% of survey respondents in Half (2019) were offered jobs when they didn't have all qualifications.

[29] One potential cost is the company needing to sort through a larger number of applications. We anticipate that the interventions will induce applications from capable candidates and also less qualified individuals. We will analyze the size of this trade-off to inform our policy recommendation.

3.5 Conclusion

Overall, this research has important ramifications for the labor market. We hope to identify one lever than can be used strategically to alleviate application gaps. Current perceptions about job ads may be discouraging women and others from applying to some jobs, particularly stretch roles. This research tests a simple and hopefully effective way to reduce this problem, namely, using encouragement around job qualifications and information about the hiring process.

Results are subject to change as data collection is currently ongoing.

Chapter 3, in part, is currently being prepared for submission for publication of the material and is coauthored with Eastmond, Tanner S. The dissertation author was the primary researcher and author of this material.

3.6 References

Abraham, L., Hallermeier, J., & Stein, A. (2023). Words matter: Experimental evidence from job applications. *Forthcoming (Revise & Resubmit), Journal of Economic Behavior & Organization.* Retrieved from https://drive.google.com/file/d/1mOl_P9ezjGY0Dfo1Bzus47yfCUY8LyYf/view

Aksoy, B., Exley, C. L., & Kessler, J. B. (2024, January). The gender minority gaps in confidence and self-evaluation. *National Bureau of Economic Research Working Paper Series*(32061). Retrieved from http://www.nber.org/papers/w32061 doi: 10.3386/w32061

Avery, D. R., & McKay, P. F. (2006). Target practice: An organizational impression management approach to attracting minority and female job applicants. *Personnel Psychology*, *59*(1), 157-187. doi: https://doi.org/10.1111/j.1744-6570.2006.00807.x

Barbulescu, R., & Bidwell, M. J. (2013). Do women choose different jobs from men? mechanisms of application segregation in the market for managerial workers. *Organization Science*, *24*(3), 737-756.

Benson, A., Li, D., & Shue, K. (in press). 'potential' and the gender promotion gap. *R&R, American Economic Review*. Retrieved from https://danielle-li.github.io/assets/docs/PotentialAndTheGenderPromotionGap.pdf

Bertrand, M., & Duflo, E. (2017). Chapter 8 - field experiments on discrimination. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of field experiments* (Vol. 1, p. 309-393). North-Holland. doi: https://doi.org/10.1016/bs.hefe.2016.08.004

Bertrand, M., & Mullainathan, S. (2004, September). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, *94*(4), 991-1013. Retrieved from https://www.aeaweb.org/articles?id=10.1257/0002828042002561

Bohren, J. A., Imas, A., & Rosenberg, M. (2019, October). The dynamics of discrimination: Theory and evidence. *American Economic Review*, *109*(10), 3395-3436. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.20171829 doi: 10.1257/aer.20171829

Burn, I., Button, P., Menguia Corella, L. F., & Neumark, D. (2019, December). *Older workers need not apply? ageist language in job ads and age discrimination in hiring* (Working Paper No. 26552). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w26552 doi: 10.3386/w26552

Burn, I., Firoozi, D., Ladd, D., & Neumark, D. (2022, July). *Help really wanted? the impact of age stereotypes in job ads on applications from older workers* (Working Paper No. 30287). National Bureau of Economic Research. doi: 10.3386/w30287

Castilla, E. J., & Rho, H. J. (2023). The gendering of job postings in the online recruitment process. *Management Science*, *0*(0). Retrieved from https://doi.org/10.1287/mnsc.2023.4674

Chaturvedi, S., Mahajan, K., & Siddique, Z. (2021). Words matter: Gender, jobs and applicant behavior. *IZA Institute of Labor Economics Discussion Paper*(14497).

Choi, J. H., Pacelli, J., Rennekamp, K. M., & Tomar, S. (2022). Do jobseekers value diversity information? evidence from a field experiment. *Journal of Accounting Research (Accepted)*. Retrieved from https://www.utah-wac.org/2022/Papers/choi_UWAC.pdf

Clark, N. F. (2014, April 28). Act now to shrink the confidence gap. *Forbes*. Retrieved from https://www.forbes.com/sites/womensmedia/2014/04/28/act-now-to-shrink-the-confidence-gap/?sh=547c2ed05c41

Coffman, K. B., Collis, M., & Kulkarni, L. (2019, November). Whether to apply. *Harvard Business School Working Paper*(20-062). (Revised June 2021)

Coffman, L. C., Featherstone, C. R., & Kessler, J. B. (2017). Can social information affect what job you choose and keep? *American Economic Journal: Applied Economics*, *9*(1), 96–117.

Dal Bó, E., Finan, F., & Rossi, M. A. (2013, 04). Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service. *The Quarterly Journal of Economics*, *128*(3), 1169-1218. Retrieved from https://doi.org/10.1093/qje/qjt008

Del Carpio, L., & Guadalupe, M. (2021). More women in tech? evidence from a field experiment addressing social identity. *Management Science*.

Dover, T. L., Major, B., & Kaiser, C. R. (2016). Members of high-status groups are threatened by pro-diversity organizational messages. *Journal of Experimental Social Psychology*, *62*, 58–67.

Ekstrom, R. B. (1981). Psychological and sociological perspectives on women's paid and unpaid work choices. *Advances in Consumer Research*, *08*, 580-584. Retrieved from https://www.acrwebsite.org/volumes/5863/volumes/v08/NA-08

Farber, H. S., Silverman, D., & von Wachter, T. (2016, May). Determinants of callbacks to job applications: An audit study. *American Economic Review*, *106*(5), 314-18. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.p20161010 doi: 10.1257/aer.p20161010

Flory, J. A., Leibbrandt, A., & List, J. A. (2015). Do competitive workplaces deter female workers? a large-scale natural field experiment on job entry decisions. *The Review of Economic Studies*, *82*(1), 122–155.

Flory, J. A., Leibbrandt, A., Rott, C., & Stoddard, O. (2021). Increasing workplace diversity evidence from a recruiting experiment at a fortune 500 company. *Journal of Human Resources*, *56*(1), 73–92.

Fluchtmann, J., Glenny, A. M., Harmon, N. A., & Maibom, J. (2024, May). The gender application gap: Do men and women apply for the same jobs? *American Economic Journal: Economic Policy*, *16*(2), 182-219. Retrieved from https://www.aeaweb.org/articles?id=10.1257/pol.20210607 doi: 10.1257/pol.20210607

Gaddis, S. M. (2018). An introduction to audit studies in the social sciences. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance* (p. 3-44). Cham, Switzerland: Springer International Publishing.

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology*, *101*(1), 109–128. Retrieved from https://doi.org/10.1037/a0022530

Gee, L. K. (2019). The more you know: Information effects on job application rates in a large field experiment. *Management Science*, *65*(5), 2077–2094.

Goldin, C., & Rouse, C. (2000, September). Orchestrating impartiality: The impact of ”blind” auditions on female musicians. *American Economic Review*, *90*(4), 715-741. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.90.4.715 doi: 10.1257/aer.90.4.715

Half, R. (2019, March 19). Survey: 42 percent of job applicants don’t meet skills requirements, but companies are willing to train up. *Cision PR Newswire*. Retrieved from https://www.prnewswire.com/news-releases/survey-42-percent-of-job-applicants-dont-meet-skills-requirements-but-companies-are-willing-to-train-up-300813540.html

Hurst, R. (2022, January 31). *Workplace Backlash? Workforce Diversity, Status Threat, and the Contractionary Effects of Pro-Diversity Claims* (Working Paper). Retrieved from https://ssrn.com/abstract=3789682

Kessler, J. B., Low, C., & Sullivan, C. D. (2019, November). Incentivized resume rating: Eliciting employer preferences without deception. *American Economic Review*, *109*(11), 3713-44. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.20181714 doi: 10.1257/aer.20181714

Kline, P., Rose, E. K., & Walters, C. R. (2022). Systemic discrimination among large us employers. *The Quarterly Journal of Economics*, *137*(4), 1963–2036.

Kuhn, P., Shen, K., & Zhang, S. (2020). Gender-targeted job ads in the recruitment process: Facts from a chinese job board. *Journal of Development Economics*, *147*, 102531.

Leibbrandt, A., & List, J. A. (2018, September). *Do equal employment opportunity statements backfire? evidence from a natural field experiment on job-entry decisions* (Working Paper No. 25035). National Bureau of Economic Research. doi: 10.3386/w25035

Linos, E. (2017). More Than Public Service: A Field Experiment on Job Advertisements and Diversity in the Police. *Journal of Public Administration Research and Theory*, *28*(1), 67-85. Retrieved from https://doi.org/10.1093/jopart/mux032 doi: 10.1093/jopart/mux032

Llinares-Insa, L. I., González-Navarro, P., Córdoba-Iñesta, A. I., & Zacarés-González, J. J. (2018). Women's job search competence: A question of motivation, behavior, or gender. *Frontiers in Psychology*, *9*. Retrieved from https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00137

Marinescu, I., & Wolthoff, R. (2020). Opening the black box of the matching function: The power of words. *Journal of Labor Economics*, *38*(2), 535-568. Retrieved from https://doi.org/10.1086/705903 doi: 10.1086/705903

Mohr, T. S. (2014). Why women don't apply for jobs unless they're 100% qualified. *Harvard Business Review*, *25*. Retrieved from https://hbr.org/2014/08/why-women-dont-apply-for-jobs-unless-theyre-100-qualified

Pager, D., & Pedulla, D. S. (2015). Race, self-selection, and the job search process. *American Journal of Sociology*, *120*(4), 1005–1054. Retrieved from https://doi.org/10.1086/681072

Quillian, L., Pager, D., Hexel, O., & Midtbøen, A. H. (2017). Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences*, *114*(41), 10870-10875. doi: 10.1073/pnas.1706255114

Reskin, B. F., & Bielby, D. D. (2005). A sociological perspective on gender and career outcomes. *The Journal of Economic Perspectives*, *19*(1), 71–86. Retrieved 2023-01-16, from http://www.jstor.org/stable/4134993

Rojas, M. (2021). Dear female job seeker: Apply for the job, ignore the 'qualifications'. *Fast Company*. Retrieved from https://www.fastcompany.com/90661349/dear-female-jobseeker-apply-for-the-job-ignore-the-qualifications

Sakowitz, J. (2018). Uncovering the gendered dimensions of job hunting. *Stanford News & The Clayman Insitute for Gender Research*. Retrieved from https://gender.stanford.edu/news/uncovering-gendered-dimensions-job-hunting

Samek, A. (2015). A university-wide field experiment on gender differences in job entry decisions. *Manuscript, UWM*.

Sandberg, S., & Scovell, N. (2013). *Lean In*. Alfred A. Knopf.

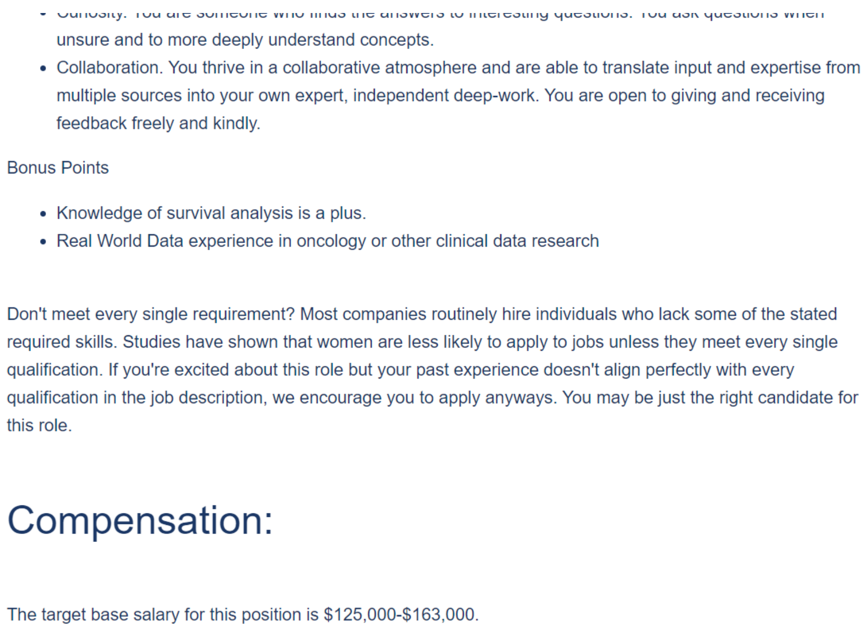
Tockey, D., & Ignatova, M. (2019). *Gender insights report: How women find jobs differently* (Tech. Rep.). LinkedIn Talent Solutions. Retrieved from https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions-lodestone/body/pdf/Gender-Insights-Report.pdf

Wille, L., & Derous, E. (2017). Getting the words right: When wording of job ads affects ethnic minorities' application decisions. *Management Communication Quarterly*, *31*(4), 533-558. doi: 10.1177/0893318917699885

Zucker, R. (2020, January 27). Is that stretch job right for you? *Harvard Business Review*. Retrieved from https://hbr.org/2020/01/is-that-stretch-job-right-for-you

Chapter 3 Appendices

3.A Additional Figures & Tables



- Curiosity. You are someone who finds the answers to interesting questions. You ask questions when unsure and to more deeply understand concepts.
- Collaboration. You thrive in a collaborative atmosphere and are able to translate input and expertise from multiple sources into your own expert, independent deep-work. You are open to giving and receiving feedback freely and kindly.

Bonus Points

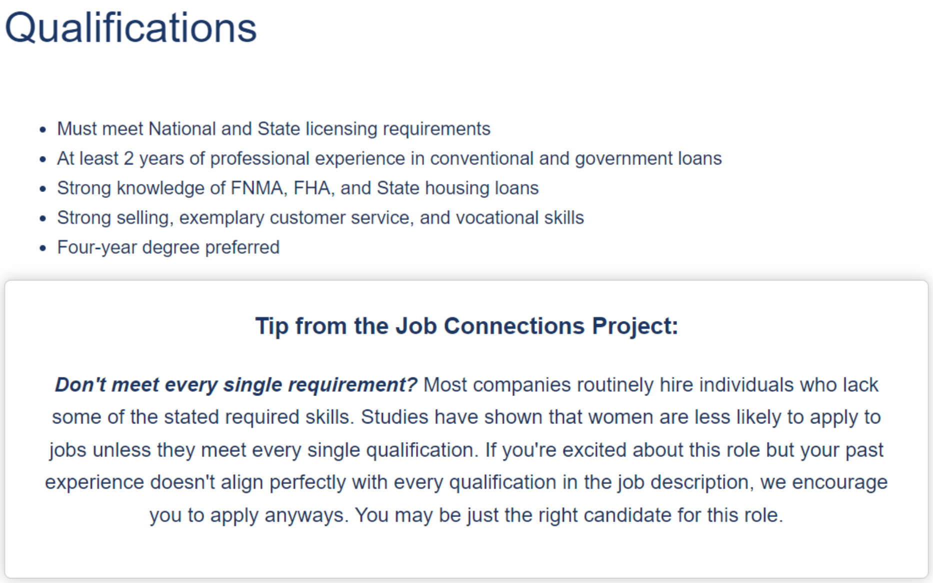
- Knowledge of survival analysis is a plus.
- Real World Data experience in oncology or other clinical data research

Don't meet every single requirement? Most companies routinely hire individuals who lack some of the stated required skills. Studies have shown that women are less likely to apply to jobs unless they meet every single qualification. If you're excited about this role but your past experience doesn't align perfectly with every qualification in the job description, we encourage you to apply anyways. You may be just the right candidate for this role.

Compensation:

The target base salary for this position is $125,000-$163,000.

Figure 3.A.1: **Example of Treatment 3 language embedded in job ad**



Qualifications

- Must meet National and State licensing requirements
- At least 2 years of professional experience in conventional and government loans
- Strong knowledge of FNMA, FHA, and State housing loans
- Strong selling, exemplary customer service, and vocational skills
- Four-year degree preferred

Tip from the Job Connections Project:

Don't meet every single requirement? Most companies routinely hire individuals who lack some of the stated required skills. Studies have shown that women are less likely to apply to jobs unless they meet every single qualification. If you're excited about this role but your past experience doesn't align perfectly with every qualification in the job description, we encourage you to apply anyways. You may be just the right candidate for this role.

Figure 3.A.2: **Example of Treatment 4, language in a box as a tip from the JCP in job ad**

Table 3.A.1: **Job seekers in the 'Encourage' treatment are more likely to click continue, while those in the 'Encourage + Info from the JCP' are less likely to click continue (Logit)**

Clicked Continue (Odds Ratios)	(1)	(2)	(3)	(4)
Treated	1.069	0.954		
	(0.160)	(0.169)		
Treatment Encourage			1.449	1.248
			(0.346)	(0.321)
Encourage+Hiring Info			0.963	0.910
			(0.182)	(0.195)
Encourage+Hiring+Women Info			1.274	1.193
			(0.347)	(0.348)
Encourage+Hiring+Women Info (JCP)			0.645*	0.506***
			(0.151)	(0.122)
N	518	506	518	506
N clusters	32	22	32	22
Job Ad FE		Yes		Yes
Pseudo R^2	0.000	0.124	0.008	0.133

Exponentiated coefficients; Standard errors in parentheses

Standard errors are clustered by job ad. Results are similar if use robust standard errors.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

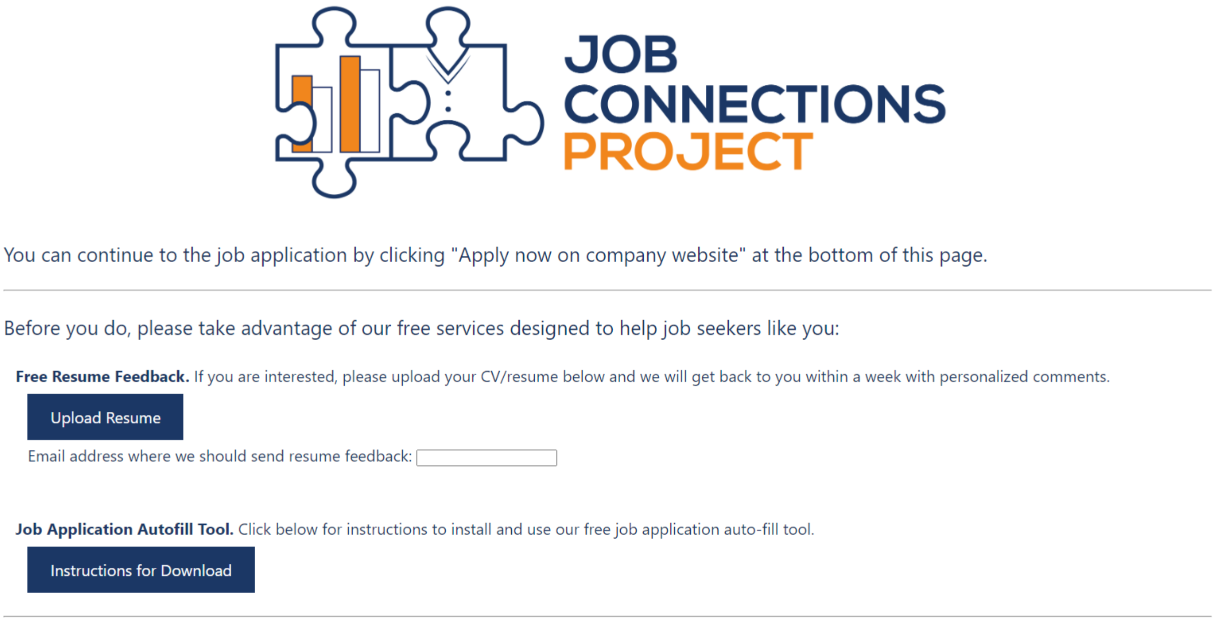
Table 3.A.2: **Black, Latinx, and Multiracial individuals may be more affected by treatment encouragement language, although results are not statistically significant**

	(1)	(2)	(3)
(Odds Ratios)	Asian	White	Black/Latinx/ Multi/Other
Treated	0.800	0.526	1.935
	(0.354)	(0.306)	(0.906)
N	83	83	83
r2_p	0.002	0.016	0.019

Exponentiated coefficients; Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

3.B Survey Instrument



You can continue to the job application by clicking "Apply now on company website" at the bottom of this page.

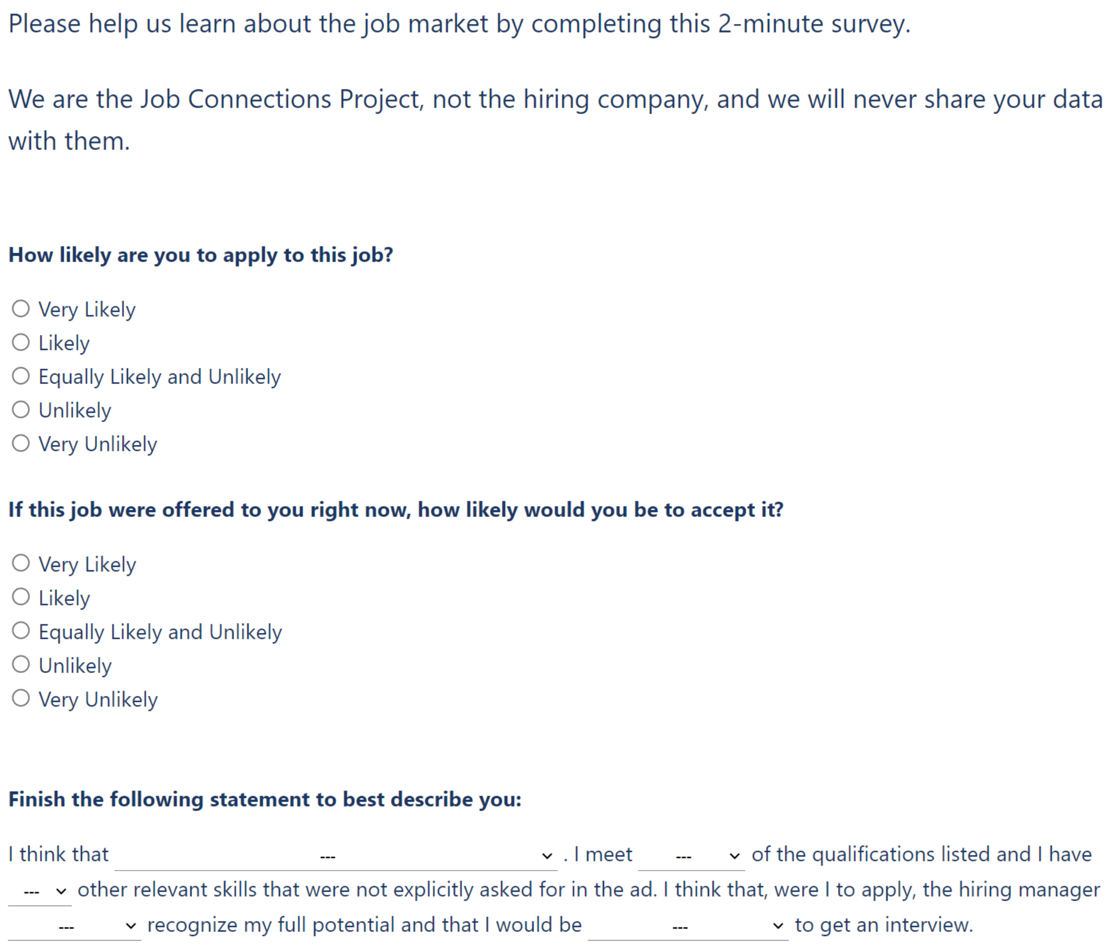
Before you do, please take advantage of our free services designed to help job seekers like you:

Free Resume Feedback. If you are interested, please upload your CV/resume below and we will get back to you within a week with personalized comments.

Email address where we should send resume feedback: ______

Job Application Autofill Tool. Click below for instructions to install and use our free job application auto-fill tool.

Figure 3.B.1: **Top of survey page with offering of job seeker services**



Please help us learn about the job market by completing this 2-minute survey.

We are the Job Connections Project, not the hiring company, and we will never share your data with them.

How likely are you to apply to this job?

○ Very Likely
○ Likely
○ Equally Likely and Unlikely
○ Unlikely
○ Very Unlikely

If this job were offered to you right now, how likely would you be to accept it?

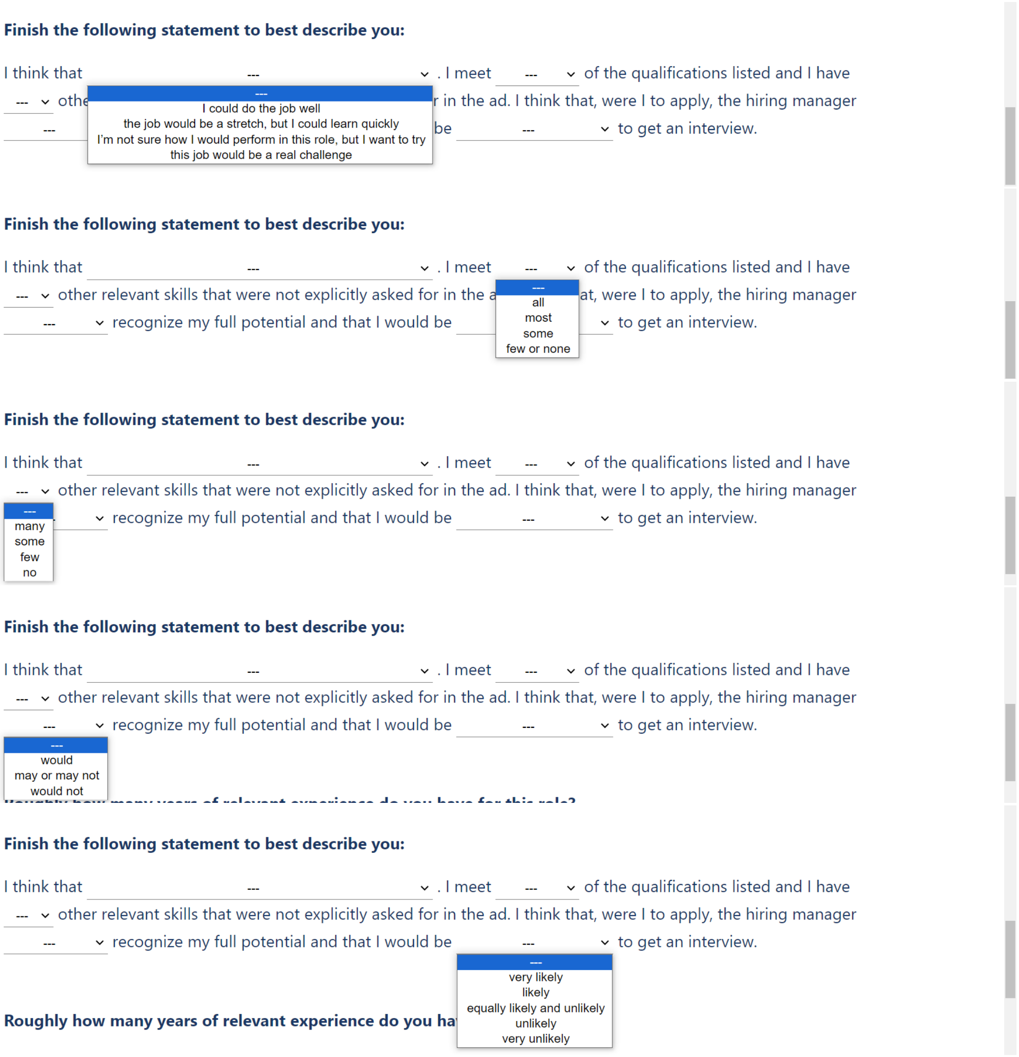
○ Very Likely
○ Likely
○ Equally Likely and Unlikely
○ Unlikely
○ Very Unlikely

Finish the following statement to best describe you:

I think that --- ⌄ . I meet --- ⌄ of the qualifications listed and I have --- ⌄ other relevant skills that were not explicitly asked for in the ad. I think that, were I to apply, the hiring manager --- ⌄ recognize my full potential and that I would be --- ⌄ to get an interview.

Figure 3.B.2: **First set of questions on survey page**

where the dropdown menus contain:



Finish the following statement to best describe you:

I think that --- ⌄ . I meet --- ⌄ of the qualifications listed and I have
--- ⌄ othe r in the ad. I think that, were I to apply, the hiring manager
--- be --- ⌄ to get an interview.

I could do the job well
the job would be a stretch, but I could learn quickly
I'm not sure how I would perform in this role, but I want to try
this job would be a real challenge

Finish the following statement to best describe you:

I think that --- ⌄ . I meet --- ⌄ of the qualifications listed and I have
--- ⌄ other relevant skills that were not explicitly asked for in the a at, were I to apply, the hiring manager
--- ⌄ recognize my full potential and that I would be --- ⌄ to get an interview.

---
all
most
some
few or none

Finish the following statement to best describe you:

I think that --- ⌄ . I meet --- ⌄ of the qualifications listed and I have
--- ⌄ other relevant skills that were not explicitly asked for in the ad. I think that, were I to apply, the hiring manager
⌄ recognize my full potential and that I would be --- ⌄ to get an interview.

---
many
some
few
no

Finish the following statement to best describe you:

I think that --- ⌄ . I meet --- ⌄ of the qualifications listed and I have
--- ⌄ other relevant skills that were not explicitly asked for in the ad. I think that, were I to apply, the hiring manager
--- ⌄ recognize my full potential and that I would be --- ⌄ to get an interview.

---
would
may or may not
would not

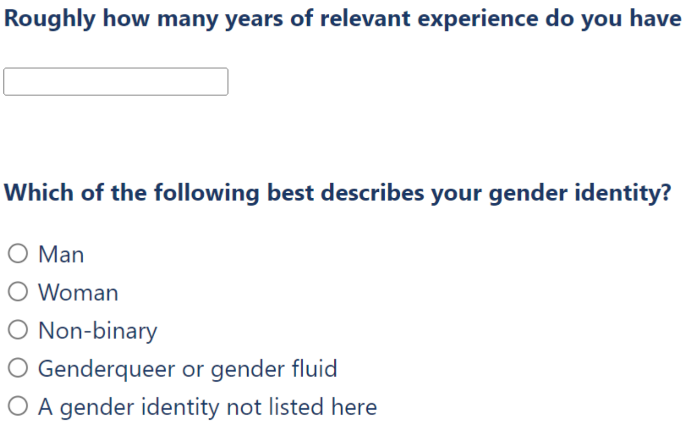
Finish the following statement to best describe you:

I think that --- ⌄ . I meet --- ⌄ of the qualifications listed and I have
--- ⌄ other relevant skills that were not explicitly asked for in the ad. I think that, were I to apply, the hiring manager
--- ⌄ recognize my full potential and that I would be --- ⌄ to get an interview.

---
very likely
likely
equally likely and unlikely
unlikely
very unlikely

Roughly how many years of relevant experience do you ha

Figure 3.B.3: **Dropdown menu options of Madlib style survey question**

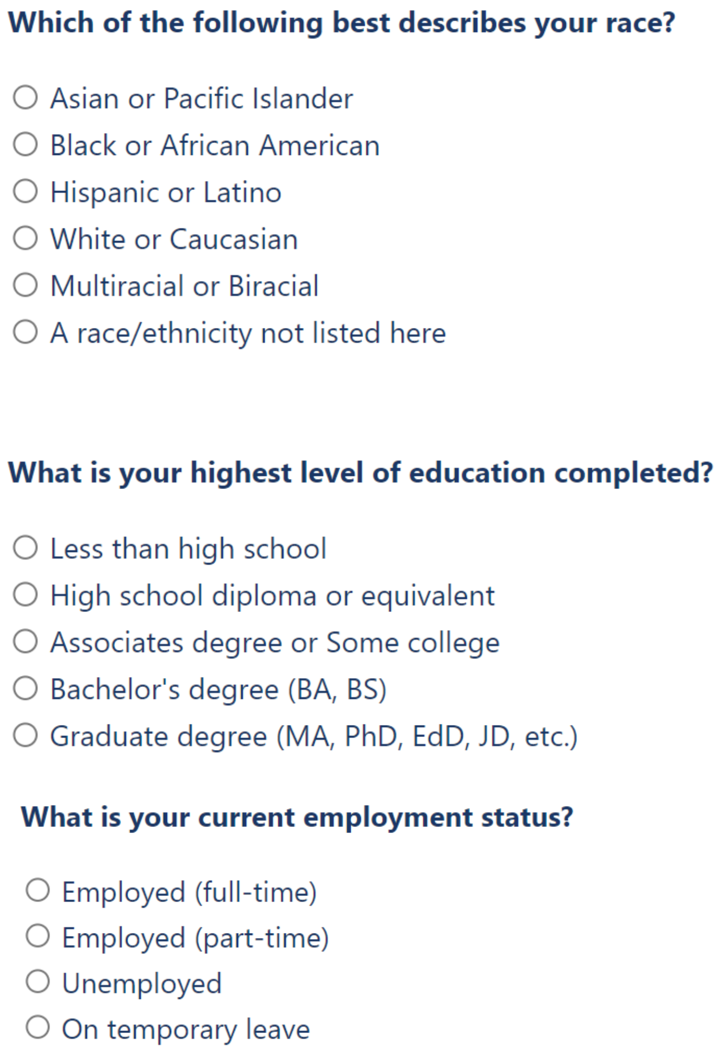


Roughly how many years of relevant experience do you have for this role?

Which of the following best describes your gender identity?

○ Man
○ Woman
○ Non-binary
○ Genderqueer or gender fluid
○ A gender identity not listed here

Figure 3.B.4: **Next set of questions on the survey page**



Which of the following best describes your race?

○ Asian or Pacific Islander
○ Black or African American
○ Hispanic or Latino
○ White or Caucasian
○ Multiracial or Biracial
○ A race/ethnicity not listed here

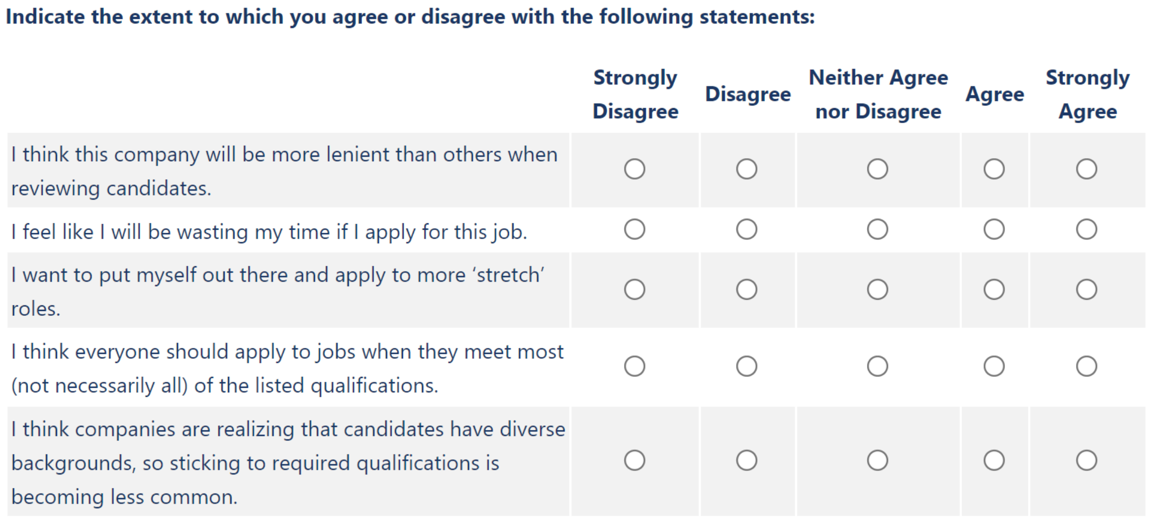
What is your highest level of education completed?

○ Less than high school
○ High school diploma or equivalent
○ Associates degree or Some college
○ Bachelor's degree (BA, BS)
○ Graduate degree (MA, PhD, EdD, JD, etc.)

What is your current employment status?

○ Employed (full-time)
○ Employed (part-time)
○ Unemployed
○ On temporary leave

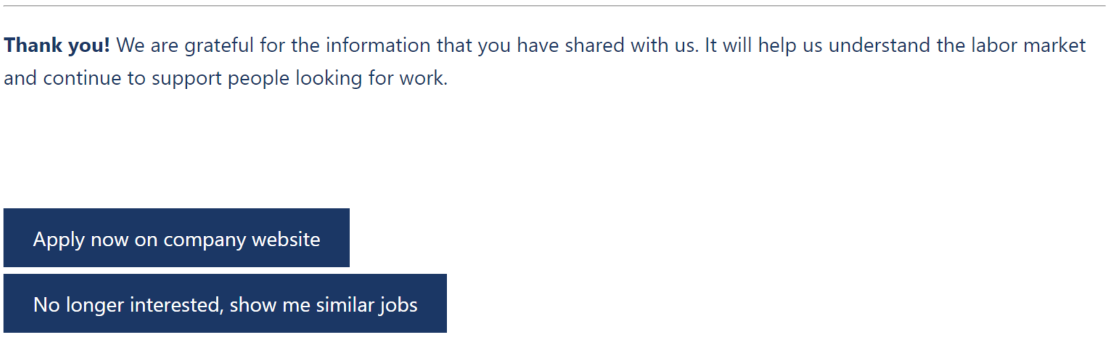
Figure 3.B.5: **Third set of questions on the survey page**



Indicate the extent to which you agree or disagree with the following statements:

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
I think this company will be more lenient than others when reviewing candidates.	○	○	○	○	○
I feel like I will be wasting my time if I apply for this job.	○	○	○	○	○
I want to put myself out there and apply to more 'stretch' roles.	○	○	○	○	○
I think everyone should apply to jobs when they meet most (not necessarily all) of the listed qualifications.	○	○	○	○	○
I think companies are realizing that candidates have diverse backgrounds, so sticking to required qualifications is becoming less common.	○	○	○	○	○

Figure 3.B.6: **Likert scale questions at end of survey page**



Thank you! We are grateful for the information that you have shared with us. It will help us understand the labor market and continue to support people looking for work.

Apply now on company website
No longer interested, show me similar jobs

Figure 3.B.7: **End of survey page, with button to "Apply now on company website"**