THE UNIVERSITY OF CHICAGO

ESSAYS ON THE ECONOMICS OF CONTROVERSIAL POLICIES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES
KENNETH C. GRIFFIN DEPARTMENT OF ECONOMICS
AND
THE FACULTY OF THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY

RAFAEL JOSYMAR JIMÉNEZ DURÁN

CHICAGO, ILLINOIS
JUNE 2022

To my parents, for their endless love.

To Marta, for coloring my days.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I'm deeply grateful to my dissertation committee for their advice and support. Leonardo Bursztyn's behavioral economics class inspired me to switch fields, and his clean experiments expanded my research possibility frontier. Pietro Tebaldi's enthusiasm for my research encouraged me when I needed it. Ali Hortaçsu was the first one to see the potential of my job market paper. Kevin Murphy taught me the powerful toolkit of price theory.

I encountered great support at the Economics Department, Booth, and many other places. To mention a few names: Ufuk Akcigit, Fernando Álvarez, David Argente, Manasi Deshpande, Michael Dinerstein, Julio Elías, Lars Hansen, Justin Holz, Diego Jiménez, Eduardo Laguna Müggenburg, Francesco Lippi, John List, Casey Mulligan, and Javier Pérez.

The Ph.D. offices at Booth and Economics helped me navigate through the administrative procedures.

My extended first year study group, "The Weyazos," provided invaluable support. I share many unforgettable memories of these past six years with many other friends.

Finally, Marta Prato recommended me to take that first behavioral economics class that changed everything.

ABSTRACT

This dissertation contains essays of the underlying economics of three controversial policies.

The first paper addresses the practice of social media platforms of banning users and removing posts to moderate their content. This "speech policing" remains controversial because little is known about its consequences and the costs and benefits for different individuals. I conduct two field experiments on Twitter to examine the effect of moderating hate speech on user behavior and welfare. Randomly reporting posts for violating the rules against hateful conduct increases the likelihood that Twitter removes them. Reporting does not affect the activity on the platform of the posts' authors or their likelihood of reposting hate, but it does increase the activity of those attacked by the posts. These results are consistent with a model in which content moderation is a quality decision for platforms that increases user engagement and hence advertising revenue. The second experiment shows that changing users' perceived content removal does not change their willingness to pause using social media, a measure of consumer surplus. My results imply that content moderation does not necessarily moderate users, but it marginally increases advertising revenue. It can be consistent with both profit- and welfare-maximization if out-of-platform externalities are small.

In the second paper, coauthored with Justin Holz and Eduardo Laguna, we study anti-price gouging laws in the US. Thirty-four states prohibit price increases during emergencies and many individuals take costly actions to report violators. We use an experiment to measure the willingness to pay to report sellers who increase prices of personal protective equipment. Over 75% of subjects pay to report even if others are willing to purchase at those prices. The willingness to pay is polarized and increases with price. We argue that reports contain information about a desire to prevent third-party transactions at illegal prices. The mechanism driving reports varies by good: we find a distaste for profits for hand sanitizers but not for face masks.

In the third paper, coauthored with Fernando Álvarez, David Argente, and Francesco Lippi (now published in Alvarez et al. (2022)), we study the possibility of banning cash payments in Mexico. We use two quasi-natural experiments that encouraged the use of debit cards and facilitated the use of ATMs in Mexico to estimate the elasticity of crime and informality to the availability of cash as a means of payment. We then construct a simple model to quantify the private costs of restricting cash usage in the economy. Our model captures the degree of substitution between cash and other payment methods at the intensive and extensive margins. We estimate the welfare effects of restricting cash by means of three key inputs: i) the elasticity of substitution between cash and credit, ii) the share of expenditures in cash by type of good obtained from detailed micro data, and iii) the elasticity of crimes to the availability of cash as means of payment. The social benefits of restricting cash usage are driven by the reduction of some criminal activities. The costs arise from the distortions that the anti-cash regulation imposes on the individual choices regarding the means of payment. We find that the private costs of heavily taxing the use of cash in Mexico outweigh the social benefits that we identify.

CHAPTER 1

THE ECONOMICS OF CONTENT MODERATION: THEORY AND EXPERIMENTAL EVIDENCE FROM HATE SPEECH ON TWITTER

1.1 Introduction

Social media is the "modern public square," according to the U.S. Supreme Court[1]—a place where speech happens among individuals with different backgrounds and ideologies. Yet, the biggest strengths of platforms—their size and diversity—also represent their greatest challenges. Forty percent of people have experienced online harassment (Anti-Defamation League, 2021), and studies document real-world consequences of online speech on hate crimes (Bursztyn et al., 2019; Müller and Schwarz, 2020a), and election outcomes (Fujiwara et al., 2021). Despite these consequences, few governments have crafted laws or regulation of online content (Carlson, 2021).

As a result, social media companies self-regulate and issue community guidelines that forbid not only illegal content but also some combination of hate speech, misinformation, harassment, spam, sexual content, and graphical content (Gillespie, 2018). Platforms moderate content by enforcing these guidelines with sanctions such as removing posts or accounts. Still, even if it is widespread, "policing speech" remains controversial (Kaye, 2019), in part due to scarce data and studies about this practice. The debate oscillates between arguments about freedom of expression (Strossen, 2018) and the harms that can be caused by online content (Waldron, 2012).

This paper contributes to the discussion by providing theory and experimental evidence of how moderation works, how it changes online behavior, and how to weigh its welfare gains and losses to different users. Guided by a model, I run two large-scale field experiments to

1. Packingham v. North Carolina, 137 S. Ct. 1730, 1737 (2017).

document the consequences of content moderation on user behavior and welfare. The focus is on hate speech on Twitter as a prominent example. One quarter of U.S. adults use this platform, and hate speech is its most sanctioned violation (Pew Research Center, 2021; Twitter, 2020b).

I begin by modeling a platform on which users spend time and interact. The platform maximizes profits by choosing its prices—the frequency at which it displays ads—and its content removal rate, which reduces spillovers between users. As in Weyl (2010), the pricing policy is what allows the platform to effectively choose the amount of time that users spend on it. The moderation policy is a quality decision that maximizes the users' willingness to engage with ads. When setting its moderation rate, the platform trades off the change in censored and non-censored users' engagement with ads. Because moderation is costly, it is only profitable if it increases the activity of at least some users. In other words, it makes sense for profit-maximizing platforms to restrict the content of some of their users if this increases the overall engagement with ads. Thus, a key parameter of the platform's decisions is how users change their time on the platform in response to moderation.

The first experiment provides information about this parameter by leveraging the reporting tool of the platform that allows flagging content that violates the rules. Twitter combines these reports with algorithms to detect violations and enforce its guidelines. It then chooses from a wide range of sanctions at the Tweet or user level, such as reducing Tweet or user visibility (also called shadowbanning), temporarily locking accounts, removing Tweets, or suspending (banning) users. Because reports increase the likelihood that Twitter detects content and, plausibly, do not affect users directly, they instrument moderation as long as sanctions are perfectly observed. Thus, reporting overcomes the challenge of moderation not being randomly assigned.

Over the course of two months, I sampled 6,000 Tweets containing slurs about disability, which constitute 98% of the sample, or that deny the Holocaust. These slurs are covered

by Twitter's hateful-conduct policy.[2] The sample included different spellings of the slurs to capture attempts to evade detection algorithms, excluded bots, inactive accounts, and other quality filters. Users enter the sample once, so there is one Tweet per user.

The day after they were posted, I randomly reported half of the Tweets for violating the rules against hateful conduct. I then collected daily server-level data of users' sanctions, their behavior and their followers' behavior, and the behavior of the users that the Tweets replied to, if any. The data comes from Twitter's Application Programming Interface (API) and other sources such as Google's Perspective API (Wulczyn et al., 2017), Botometer's API (Yang et al., 2020), and shadowban.eu's API (Merrer et al., 2020).

The first set of results show that reporting has a first-stage impact on sanctions. Reported Tweets are 66% (1.4 percentage points or 0.08 standard deviations) more likely to be removed within three weeks by Twitter than non-reported Tweets, with an F-statistic of 11. The treatment does not significantly change user suspensions and shadowbans, the other observable sanctions.[3] However, I also find evidence of "unobservable" sanctions, such as temporarily locking users' accounts, which I obtained from the updates that Twitter sent me after I reported users.[4] Hence, reports remain a valid instrument for all sanctions even if they violate the exclusion restriction for observed sanctions.

The second set of results concern the reduced-form impact of reports on user behavior on the platform. This estimation is possible because accounts do not disappear after reporting. I find that reports do not reduce the users' Twitter activity or their likelihood of reposting

2. The policy mentions the Holocaust and slurs that reinforce negative stereotypes about a protected category, including disability. These slurs are only a subset of hate speech, but most other slurs are appropriated by minorities (Bianchi, 2014) and led to high false positives in a pilot study. Another option was to sample Tweets with a detection algorithm from the computer science literature, but even state-of-the-art methods suffer from low external validity (Arango et al., 2019).

3. There is no evidence of users self-censoring (deleting their posts or accounts, or locking their accounts from public view) in response to reports. There is also no evidence that reports induce their other Tweets to disappear.

4. No observable sanctions were implemented in 6.6% of the accounts I reported, but Twitter provided an update that it had found rule violations. This number is likely biased downward because Twitter does not always send updates, even for reports for which a sanction is observed.

hate. A proxy of the hours spent on Twitter, constructed with the daily number of Tweets and likes, increases by 7.5% (0.042 standard deviations) in the three weeks after reports, but it is not statistically significant.[5] The fraction of hateful Tweets that users post in the three weeks after the treatment, measured using Google's toxicity score, decreases by an insignificant 1.8%.

The third set of results show that reporting has significant spillovers on other users. The main measure of spillover is the activity of the users to whom the Tweets in my sample are replying, which I call "replied users"—86% of Tweets reply or quote a post from another user. Reports increase the time the replied users spend on the platform over the course of three weeks by 10%, or 10 minutes per week. Furthermore, the estimate is stronger among Tweets that attack the other user, rather than, for example, those that are just replies among friends. The effect is 13.4% among those Tweets that were labeled as attacks by human annotators.

Results are robust to alternative measures of user activity and hatefulness, dropping outliers, and specifications with different sets of controls. Together, these findings imply that sanctions induced by reporting do not change the behavior of those who posted the Tweets; content moderation does not seem to moderate users. Reports, however, increase the activity of those attacked by the hateful posts. Hence, the evidence supports the model's prediction that content moderation in a profit-maximizing platform marginally increases the advertising revenue from some users.

Does this evidence mean that moderation increases welfare? Not necessarily. Following Spence (1975), another result from my model is that a platform can, in principle, remove too much or too little content relative to a surplus-maximizing planner. Intuitively, the monopolist caters to the marginal consumer, whereas the planner caters to the average

5. Moreover, the impact on time spent might be biased downward, because Twitter restricts some accounts temporarily (Twitter, 2021d). In these cases, the number of Tweets and likes will be mechanically lower, even if users do not change their behavior.

consumer. From the consumers' point of view, a utilitarian test of whether the platform underprovides or overprovides moderation is whether a small change in censorship, all else equal, increases or decreases consumer surplus.[6] Even if this test ignores externalities, many costs and benefits associated with moderation, such as free speech and direct harms from hateful expressions, occur inside platforms.

I conducted the test in a survey of 3,000 U.S. Twitter users sampled through Luc.id, a widely used online survey panel provider.[7] I shift users' beliefs about the likelihood that Twitter moderates hate speech, and I elicit their willingness to accept (WTA) to stop using social media. I vary the perceived likelihood of moderation using an information-provision design with an active control group (Haaland et al., 2020). I randomize survey participants into two treatment arms that receive different information about the likelihood of moderation among hateful Tweets.

The information provided comes from a random sample of 10,000 Tweets that I collected in August 2020 and classified as hate speech with the help of human annotators. I vary the likelihood of moderation without deception by using different rules to classify hate speech. Under a majority decision rule, in which a post is hateful if most annotators label it as such, Twitter removes 3.6% of hateful Tweets or suspends their authors within one month. Under a consensus decision rule, in which a post is hateful if all annotators label it as such, the likelihood of moderation is 9.1%. Under both rules, the prevalence of hate—that is, the fraction of hateful Tweets—is less than 1%. Both treatment arms receive the same information about hate prevalence, which allows isolating the effect of moderation.

The survey first elicits beliefs about the prevalence of hate speech and the likelihood of moderation with incentives for accuracy and then provides participants with the randomized

6. This test can be generalized to a model of multiple platforms by measuring the change in users' social-media valuation, not just their valuation of a given platform.

7. I reweight observations to match a representative sample of Twitter users based on gender, age, race or ethnicity, region, and political orientation. I also present unweighted results.

information. Respondents are told that some of them will be randomly selected for a small follow-up study that compensates participants to stop using social media (Twitter, Facebook, Instagram, Snapchat, YouTube, Reddit, and TikTok) for one week. I then elicit the WTA to participate in this follow-up, using an incentive-compatible procedure in the form of an iterative multiple price list (iMPL).[8]

I find large misperceptions about hate speech and moderation. Most users overestimate the prevalence of hate speech on Twitter and the likelihood of sanctions. Ninety-six percent of users believe the prevalence of hate speech is above the observed value of less than 1%, with a median of 33%. Eighty-four percent of respondents believe the likelihood of moderation is above 9.1%, with a median of 36%.[9]

Informing participants of a higher likelihood of moderation does not change their valuation of social media. The WTA falls by 15 cents (0.5% or 0.004 standard deviations), from \$33.7 to \$33.6. This result is robust to different specifications and measures of WTA, and I find no evidence that it is explained by inattention or experimenter demand effects.[10] At the end of the survey, I asked participants to repeat the information about the percent of Tweets that get sanctioned. The treatment effect on this recollection was 5.6 percentage points, significantly different from zero (F-statistic = 36) and not statistically different from 5.5, which is the gap between the information provided in both arms, 9.1% and 3.6%. The treatment also significantly shifted the posterior beliefs about the likelihood of moderation on Facebook and there is suggestive evidence that it increased the time that minorities spent on Twitter one week after the survey.

8. In this procedure, participants have to choose if they are willing to participate in the follow-up for different compensation offers. The sequence of offers starts at \$50, and subsequent amounts decrease or increase depending on whether participants accept or reject. The sequence stops until the WTA is classified in 11 intervals, which I then convert into a continuous measure following Allcott and Kessler (2019).

9. Platforms' lack of transparency could be driving these misperceptions. The likelihood of moderation on Facebook remained unknown until a whistleblower revealed internal documents some weeks after my survey (Giansiracusa, 2021).

10. The experiment was ex-ante powered to detect effects of 0.1 standard deviations, and the sample size is more than double what Haaland et al. (2020) recommend for information-provision designs.

Overall, my results suggest that moderation on Twitter is consistent with profit maximization, and they rule out large moderation distortions from the consumers' point of view, holding constant the prevalence of hate speech. These findings have two policy implications. First, cost-benefit analyses of online moderation can emphasize its offline consequences, such as hate crimes. Second, authorities might want to deal with hate speech on social media not by directly regulating moderation, but by supervising platforms' pricing (advertising) policies; Twitter could still be setting its advertising loads suboptimally, leading to inefficient amounts of hate speech.

The paper makes four contributions to a multi-disciplinary literature. First, it provides evidence of the online effects of moderation. A growing body of work in economics focuses on the offline consequences of online content and social-media penetration (Enikolopov et al., 2020; Müller and Schwarz, 2020a,b; Bursztyn et al., 2019; Braghieri et al., 2021). Other work studies government online censorship (Hobbs and Roberts, 2018; Roberts and Roberts, 2018; Chen and Yang, 2019). The computer science literature documents the relationship between content moderation and online behavior (Ali et al., 2021; Rauchfleisch and Kaiser, 2021; Jhaver et al., 2021; Zannettou, 2021), but most of these exercises are non-causal. A challenge with observational studies is isolating the effect of moderation from confounders. For example, Chandrasekharan et al. (2017) find that banning groups on Reddit decreased their former members' activity on the platform, but this could happen both because they are sanctioned or because they find the platform less attractive after the group closures.

Experimentally varying moderation, however, is also challenging due to limited cooperation with platforms. Thus, a second contribution is using social media's infrastructure experimentally, as Levy (2021) did on Facebook, which is useful for independent research. The reporting treatment is similar to other exercises by academics (Carlson and Rousselle, 2020), Governmental organizations (Jourová, 2016; Reynders, 2020), and non-profits (Matias et al., 2015; Center for Countering Digital Hate, 2021) who report content to monitor

platforms' responsiveness. However, these exercises are non-experimental (they contain no control group) and do not analyze the impact on other outcomes beyond the platform's response. Experimental interventions include counter-speech treatments (Munger, 2017, 2021; Siegel and Badaan, 2020), reminders of Twitter suspensions (Yildirim et al., 2021), and censoring hate speech in the lab (Álvarez-Benjumea and Winter, 2018).

A third contribution is combining an information-provision design with a welfare elicitation of social media. Haaland et al. (2020) and Bursztyn and Yang (2021) provide overviews of information-provision designs, and Bottan and Perez-Truglia (2017) and Bursztyn et al. (2020) are some applications. The WTA that I elicit is in the ballpark of other social-media welfare studies such as Mosquera et al. (2020), and Allcott et al. (2020); the median and mean WTA per week were \$15 and \$34, respectively.[11] Providing information required computing other basic statistics, surprisingly scarce in the literature, such as the prevalence of hate speech in a random sample of posts (0.1%-5.6% depending on the measure) and the occurrence of Tweet deletions and user suspensions (2.5%-9.1% among hateful Tweets, within one month).[12] As other surveys find (Anti-Defamation League, 2021), minorities experience more harassment online, but I also find that they are more likely to be sanctioned by Twitter.

The fourth contribution is to develop a simple model of user behavior and platform moderation decisions that captures spillovers between users, using the two-sided market framework of Weyl (2010). Prices allow the platform to determine its amount of hateful and non-hateful content, which clarifies the separation between pricing distortions and moderation distortions as in Spence (1975).[13] Liu et al. (2021) are among the first to model

11. This is after reweighting my sample to match representative U.S. Twitter users. Allcott et al. (2020) find a median and average WTA of \$25 and \$45 per week, respectively. These estimates were for deactivating Facebook over four weeks.

12. Relia et al. (2019) find that 0.5% of Tweets in a sample of 73.42 million posts contained hate speech keywords. Founta et al. (2018) found a 4% prevalence in a random sample of 10,000 Tweets. Facebook (2021) reports a prevalence of 0.05% of hate speech among all views. Few studies report the occurrence of sanctions. An exception is Merrer et al. (2020), who document that 2.34% of accounts are shadowbanned. Seyler et al. (2021) find that 5.1% of accounts from a 2009 sample are suspended.

13. There is evidence that consumers respond to platforms' advertising policies; Huang et al. (2018) traced

moderation decisions and to discuss the implications of different revenue models on platform incentives. One difference with their model is that, in my framework, users respond to the pricing policy (advertising frequency) of the platform.[14] Acemoglu et al. (2021) model online misinformation and show that engagement-maximizing platforms have incentives to create filter bubbles and propagate extremist content.

The next section develops the model. Section 3 provides background information about hate speech, moderation, Twitter, and the data sources that this study uses. Sections 4 and 5 describe the experimental design of both experiments and present their results. Section 6 concludes.

1.2 Model

Users and platform. Users can be one of two types, $\theta \in \{A, H\}$. "Acceptable" users ($\theta = A$) post content that is not subject to content moderation. "Hateful" users ($\theta = H$) post content that is censored by the platform with probability c, the censorship or moderation rate. Users who join the platform experience utility that increases on the time that they spend consuming or posting content.[15]

The utility of spending t minutes on the platform for user i of type $\theta_i = \theta$ is

$$\underbrace{U_i^{\theta}(t; \mathbf{T}, c)}_{\text{Utility from consuming content}} \qquad \underbrace{- t \times w_i(1 + p^{\theta})}_{\text{Time cost}}, \tag{1.1}$$

out a downward-sloping demand curve for a music platform by randomizing ad-loads across consumers.

14. In both frameworks (under an advertising business model), platforms use moderation as a tool to increase revenue. In Liu et al. (2021), moderation increases revenue through increases in the consumer base. However, in my model, prices determine the customer base and moderation increases revenue through the willingness of users to engage with ads, given the customer base.

15. No difference exists between consuming or producing content. In practice, however, users differ substantially in the amount of content they post. On Twitter, few users post the majority of Tweets (Wojcik and Hughes, 2019). Yet, it is not obvious whether users who like posts are less responsible for their diffusion than those who write them. For instance, sometimes Twitter alerts the followers of a user when she likes a post.

where $U_i^\theta(0;\mathbf{T},c) = 0$ for all i. $\mathbf{T} = (T^A, T^H)$ is the aggregate content of the platform and captures spillovers and network effects, where T^θ is the total content posted by θ users. The sign of spillovers is flexible; users can be positively or negatively affected by each type of content. For instance, A users could dislike encountering hate speech, but haters might like trolling A users. The censorship rate c enters the utility function because it reduces spillovers from hateful content (since users see less of it), but users can also obtain direct utility or disutility from c, for example, haters might dislike having their account locked.

The time-cost of t minutes spent enjoying content is proportional to the value of time $w_i > 0$. Moreover, the "price" that users pay is the advertising load p^θ; the time they have to spend watching ads per minute of content consumed. Following Weyl (2010), the platform can set a different price for each type of user.

The time that user i spends enjoying content is t_i^*, which maximizes (1.1) with respect to $t \geq 0$. The aggregate content demand $\mathbf{T}$ is then computed setting

$$T^\theta \equiv \int_{\{i:\theta_i=\theta\}} t_i^* \mathrm{d}i, \quad \text{for each } \theta. \tag{1.2}$$

Since the time spent on the platform by any user is decreasing in the advertising load, one can define the inverse demand functions $P^A(\mathbf{T},c)$ and $P^H(\mathbf{T},c)$, where $P^A(\mathbf{T},c)$ is equal to the p^A inducing T^A given $\mathbf{T}$ and c; similarly for $P^H(\mathbf{T},c)$.[16] In other words, the pricing policy—not moderation—is what allows the platform to choose the amount of content of both types of user.

16. Formally, imposing rational expectations $\widetilde{T^\theta} = T^\theta(p^\theta, c, \widetilde{T^A}, \widetilde{T^H})$, one can invert T^θ in an interior equilibrium point, where $\widetilde{T^\theta} > 0$. This procedure requires demands T^θ to be strictly decreasing in p^θ, which results from imposing Inada conditions on utilities or full-support assumptions as in Weyl (2010).

The platform maximizes profits solving

$$\max_{\mathbf{T},c} \underbrace{a \times \left(\underbrace{P^A(\mathbf{T},c)\,T^A + P^H(\mathbf{T},c)\,T^H}_{\text{Time spent watching ads}} \right)}_{\text{Advertising revenue}} - \underbrace{\phi(\mathbf{T},c)}_{\text{Cost of moderation}}, \tag{1.3}$$

where $a > 0$ is the price per unit of advertising,[17] and ϕ is a function describing the costs of censorship. For instance, Gillespie (2018) documents that moderation is a labor-intensive task, and Kaye (2019) argues that regulatory fines push platforms to remove borderline content.[18]

The first-order condition (FOC) with respect to quantity T^θ is similar to a standard monopoly problem. The FOC with respect to c requires that

$$a \times \left(\underbrace{\frac{\partial P^A(\mathbf{T},c)}{\partial c} T^A + \frac{\partial P^H(\mathbf{T},c)}{\partial c} T^H}_{\text{Change in time spent watching ads}} \right) = \underbrace{\frac{\partial \phi(\mathbf{T},c)}{\partial c}}_{\text{Marginal cost of moderation}}. \tag{1.4}$$

This condition is analogous to the quality decision in Spence (1975); the platform moderates such that the marginal benefit—the value of the marginal increase in the willingness to watch ads—equals the marginal cost. The left-hand side of equation (1.4) clarifies the main trade-off faced by the platform when choosing its moderation policy. Consistently with the observations in Kaye (2019) regarding controversial pages,

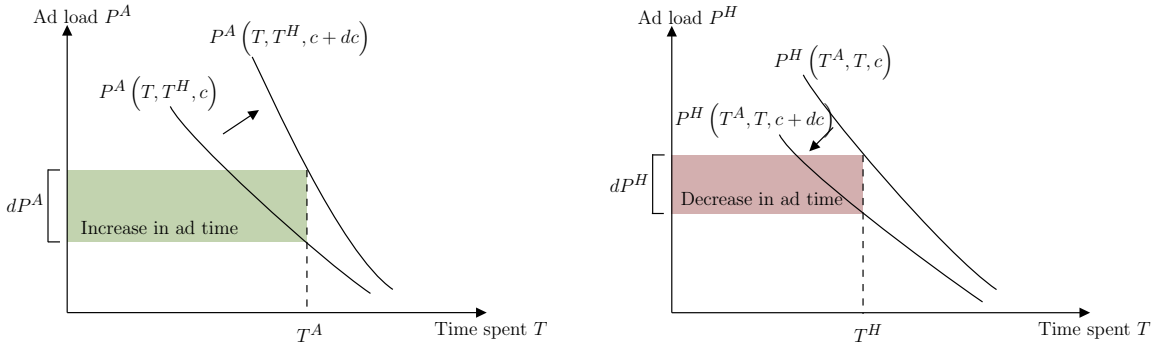
> These kinds of pages seem to put Facebook in a no-win position: If they leave up the page, they anger opponents who see hateful content or disinformation; if they take it down, they offend free-expression advocates who do not think the

17. Incorporating different prices for haters and non-haters or market power on the digital advertising market is possible, but this extension adds little to the results.

18. Platforms might worry about future fines, even if current ones are small; e.g., in 2019, Germany fined Facebook for 2 million euros for violating the NetzDG law (Bundesamt für Justiz, 2019).

rules very clearly articulate hate speech standards.

Figure 1.1 illustrates this trade-off for the case in which $\partial T^A/\partial c > 0$, and $\partial T^H/\partial c < 0$. For a fixed amount of time that users spend on the platform, moderation changes the number of ads they are willing to watch. The platform increases revenue from A users, who dislike hateful content, while it looses revenue from H users, who do not like to be censored. The optimal level of content moderation balances the net change in revenue with the marginal cost of increasing the censorship rate.



(a) Acceptable users **(b)** Hateful users

Figure 1.1: Graphical intuition of the platform's moderation decision

Notes: These figures plot the change in the inverse demands of acceptable and hateful users in response to an increase in moderation, dc, holding quantities fixed, and assuming that the moderation elasticity is positive for A and negative for H. The colored areas are the change in time spent watching ads, which equals the change in ad load dP^θ times the time spent T^θ. The net revenue gains equal the green minus the red area, multiplied by ad prices.

The FOC is, equivalently,[19]

$$-\frac{\partial T^A/\partial c}{\partial T^A/\partial p^A} a T^A - \frac{\partial T^H/\partial c}{\partial T^H/\partial p^H} a T^H = \frac{\partial \phi(\mathbf{T}, c)}{\partial c}. \tag{1.5}$$

We know that the right-hand side is strictly positive (by assumption), and that demand decreases in prices, $\partial T^\theta/\partial p^\theta < 0$. Therefore, at the optimal level of c for the platform it must be that either $\partial T^A/\partial c > 0$, or $\partial T^H/\partial c > 0$, or both. In words, for at least one type of user,

19. This applies the implicit function theorem. For example, letting $\widetilde{T^A} = T^A(p^A, c, \widetilde{T^A}, \widetilde{T^H})$, taking the total derivative implies $0 = \frac{\partial T^A}{\partial p^A} dp^A + \frac{\partial T^A}{\partial c} dc$, so that $\frac{dp^A}{dc} = -\frac{\partial T^A/\partial c}{\partial T^A/\partial p^A}$.

moderation must increase their platform activity, holding constant the aggregate quantities. The derivatives of T^θ with respect to c, one for each type, are the main parameters of interest of my first experiment.

Welfare. The platform-optimal level of censorship could differ from the socially-optimal level. Similarly to Spence (1975), the platform in my model optimizes moderation with respect to the marginal users. The social planner, however, chooses the level of censorship that maximizes total welfare, which includes the impact of moderation on inframarginal consumers. I formalize this argument in Appendix A.1; the platform can moderate more or less than a surplus-maximizing social planner, holding quantities $\mathbf{T}$ fixed.[20] Hence, two distortions exist: the usual monopolist pricing distortion that leads to inefficient quantities and an additional quality distortion.

The goal of my second experiment is to test the second distortion—to evaluate whether Twitter provides too little or too much content moderation from the perspective of the user. I follow the approach of Mosquera et al. (2020) and Allcott et al. (2020) to measure consumer surplus. In practice, I quantify the impact of different levels of censorship on the willingness to accept a monetary reward to pause the use of social media. I ask users to pause the use of social media, not just a single platform, to allow for substitution between platforms as argued in Appendix A.1.

20. Liu et al. (2021) argue that platforms undermoderate in an ad-based business model and overmoderate in a subscription-based business model. However, in their ad-based business model there are no prices, so the platform has to use moderation as the only tool to adjust its quantity. In their subscription-based case there are prices, but the sign of the Spence distortion is determined by their extremeness aversion assumption, which in practice implies that the demand curve becomes steeper in response to a small increase in moderation.

1.3 Background and Data Sources

1.3.1 Twitter and Moderation of Hate Speech

Twitter is a microblogging social media platform. Users of this platform create profiles that display self-reported information such as their name, a short biography, and a profile picture. They also post messages to their profiles called Tweets, which contain a combination of text of up to 280 characters, photos, and videos. Users can follow other accounts to see their Tweets more readily, but they can interact with others without following them. They interact with others' Tweets by giving them a like (or favorite), replying to them, Retweeting (reposting) them, or quoting them.

Like all social media platforms in the Surface Web, Twitter has rules that delimit the content that users are allowed to post. Besides illegal activity, the rules tend to cover hate speech (as well as misinformation, harassment, spam, sexual content, and graphic content). Hate speech has no single legal definition of hate speech (Waldron, 2012; Strossen, 2018). Still, most platforms define it in their rules using common elements such as the concept of protected categories from U.S. anti-discrimination law (Gillespie, 2018). Twitter's hateful-conduct policy (Twitter, 2021b) says, "You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease".

Twitter enforces these rules and moderates content by sanctioning users. Figure 1.2 illustrates the process of content moderation. Twitter can detect content by algorithms, or by the "flagging" mechanism that allows users to report Tweets or accounts for violating the rules. After the content is detected, a team of human moderators or an algorithm decide whether to enforce the rules by imposing post-level or account-level sanctions. The range of sanctions include a combination of removing Tweets from the platform, shadowbanning

(reducing the visibility of) users or Tweets,[21] and suspending or banning users (that is, deleting their accounts). Other sanctions, such as locking accounts, prevent users from posting or liking content and can last from 12 hours to seven days. See Twitter (2021d) for the full list of sanctions.

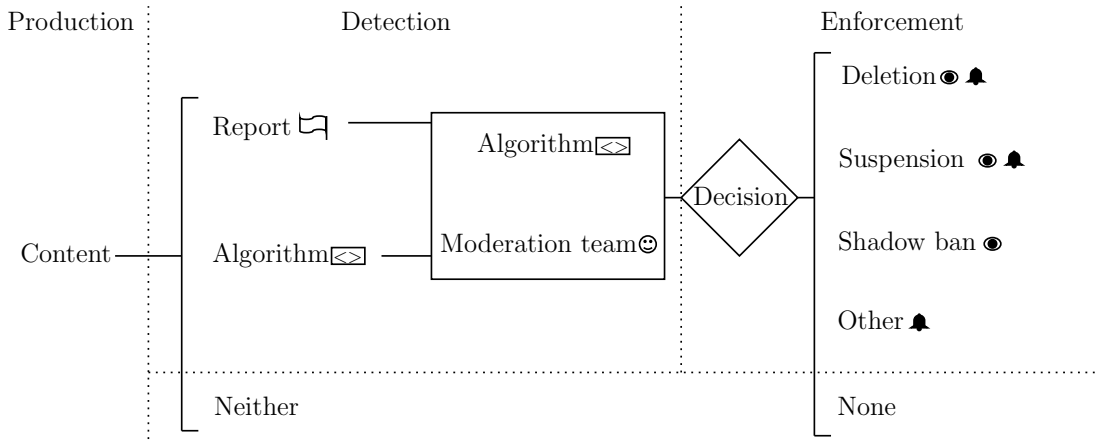


Figure 1.2: Content moderation process

Notes: Eye icons indicate that the sanction is observable to others besides the user. Bell icons indicate that the user receives a notification. This diagram omits interventions at the production stage, such as recent tests in which Twitter asked users if they wanted to review offensive Tweets before posting. It also omits the appealing process, in which users can contest a sanction.

Sanctions differ by their observability, that is, whether they are privately observed by the user or publicly observable, and whether the user is notified about them. When Twitter removes a Tweet, users are notified that they violated the rules, and must remove the Tweet to be able to use the platform again. Twitter replaces the Tweet with a notice that indicates it violated the rules. Anyone with access to the Tweets can see the notice, starting from the moment that Twitter sanctions the Tweet and up to 14 days after the user agrees to remove the post. When Twitter suspends a user, she can no longer log in to the account, and her profile and Tweets are replaced by suspension notices, which seem to last indefinitely. In principle, suspended users cannot create new accounts, but in practice they do. Users

21. Twitter has stated it does not shadowban users (Gadde and Beykpour, 2018), even if it ranks Tweets "to create a more relevant experience" (Twitter, 2021a). However, Merrer et al. (2020) document evidence of shadowbanning.

are not notified when they are shadowbanned, but Twitter sometimes hides their Tweets behind a notice—especially those that reply to another user. Twitter notifies users when their accounts are locked, but whether others can observe this sanction is unknown.[22] Figure A.2 shows examples of public notices and the notifications that users receive.

1.3.2 Measuring Hate Speech

Platforms rely to some extent on algorithms to detect hate speech and enforce their rules. Most of the detection algorithms in the computer science literature share the following procedure (see Fortuna and Nunes (2018) for a review of the literature). The first step is to obtain a training dataset, consisting of a sample of texts—usually social media posts—paired with labels, for example, hate speech or not hate speech. Often, these labels or "ground truth" result from aggregating the opinions of multiple humans or "annotators" into a single category. For example, Davidson et al. (2017) ask three or more crowd workers to annotate each Tweet as "hate," "offensive,", or "neither." Then, they aggregate these annotations into a single label with the majority decision rule, that is, the category chosen by most annotators. The second step is to convert the text into vectors of features with text analysis, reviewed in Gentzkow et al. (2019). The final step is to use machine learning to predict the labels with the features.

One challenge in the hate-speech-detection literature is the algorithms' low external validity; see Arango et al. (2019) and Fortuna et al. (2021). For this reason, this study uses three approaches to classify hate speech and limit measurement error. First, for large-scale tasks, I use the Perspective toxicity score developed by Google. This score is widely used in the industry and as a benchmark in academic articles. It is a number between 0 and 1 that

22. Twitter (2021c) shows examples of notices of locked accounts, but anecdotal evidence suggests accounts are locked without any notice. For instance, in 2020 Twitter locked actor James Woods and his account did not show any notice (Whalen, 2020).

reflects the likelihood that a text is an attack or harassment.[23] Many studies classify posts as hate speech if their toxicity is higher than a 0.8 cutoff (ElSherief et al., 2018; Han and Tsvetkov, 2020; Vidgen et al., 2020). Second, I sample hate speech using keywords, instead of an algorithm, to minimize false positives in the reporting experiment. Third, I use human annotation by MTurk workers to account for measurement error in the first experiment and to increase the interpretability of the information treatment of the second experiment.

1.3.3 Data Sources

Most of the variables analyzed in this paper come from Twitter's API. This data source provides publicly available information about all Twitter users, such as their number of followers, number of accounts they follow, date of account creation, total number of Tweets and likes, and biography. The API provides additional information about users who do not restrict their profile visibility, such as their list of followers and accounts followed, and a collection of up to 3,200 of their most recent Tweets. The API returns detailed information for these Tweets, such as their timestamp, text and media, likes, and Retweets. This source also allows me to sample Tweets by searching for specific keywords or sampling at random 1% of all Tweets. Lastly, I also use this API to detect whether Twitter removes specific Tweets or suspends users, following the procedure outlined in Appendix A.3.1.

Besides the API, I also collect some information manually from the website. Twitter occasionally notifies users when it sanctions an account they previously reported, even if the sanction might not correspond to the reported content.[24] Figure A.4 has a screenshot of some of these updates. I collect this information for the reporting experiment, because it provides evidence of "unobservable" sanctions.

23. The algorithm is a convolutional neural network trained on Wikipedia Talk Pages; see Wulczyn et al. (2017) and Dixon et al. (2018).

24. Twitter says: "You will receive an in-product notification if an action is taken on an account that you recently reported. This action may or may not be related to your report" (Twitter, 2021e).

I also use other APIs. I retrieve the toxicity score of posts from Google's Perspective API. I also obtain a measure of the likelihood that users are bots from the Botometer API (see Yang et al. (2020)). Finally, I retrieve measures of shadowbans from the API of Shadowban.eu, because Twitter does not give an official shadowban measure. This API measures different forms of shadowbanning, for example, whether Twitter hides accounts, Tweets, or replies from search results (see Merrer et al. (2020) for more details). I combine the different measures into a single indicator of whether users are shadowbanned.

Another data source is human annotation; I ask MTurk workers to annotate posts. For example, I follow the approach in Davidson et al. (2017) and ask workers to classify posts as "Hate speech," "Offensive but not hate speech," and "Neither offensive nor hate speech." I assign three workers to annotate each post. I give them Twitter's definition of hate speech for reference, offer a $20 bonus to the five most accurate workers (measured by the inter-annotator agreement), and include attention checks to improve the quality of annotations. Figure A.23 in the Appendix includes screen shots of the instructions. Then, I aggregate workers' annotations into a single label using either the majority decision rule, in which a post is hate speech if two or three workers label it a such, or the consensus decision rule, in which all three workers have to agree.

Lastly, I obtain demographics of representative Twitter users from the American Trends Panel (ATP) of September 2020. The Pew Research Center conducts this nationally representative panel of randomly selected U.S. adults.

1.3.4 Summary Statistics

Accounts and Tweets. According to the ATP, 25% of adults in the U.S. use Twitter. Table 1.1 displays selected summary statistics of Twitter users and their accounts. Twitter users are younger, more educated, and more likely to be Democrats than the general population. Thirty-one percent of them are between 18 and 29 years old, 40% are at least college

graduates, and 35% are Democrats, compared to 18%, 33%, and 30%, respectively, in the overall ATP respondents. The table also shows statistics from a sample of 200,000 Tweets that I collected in August 2020 from the 1% random sample of Twitter's API. On average, the accounts in this sample were five years old, posted 12 Tweets per day, gave 13 likes per day, followed 1,000 users, and had 4,800 followers. Ten percent of these accounts are bots; that is, they have a Botometer score of 0.5 or more.

Table 1.1: Summary statistics

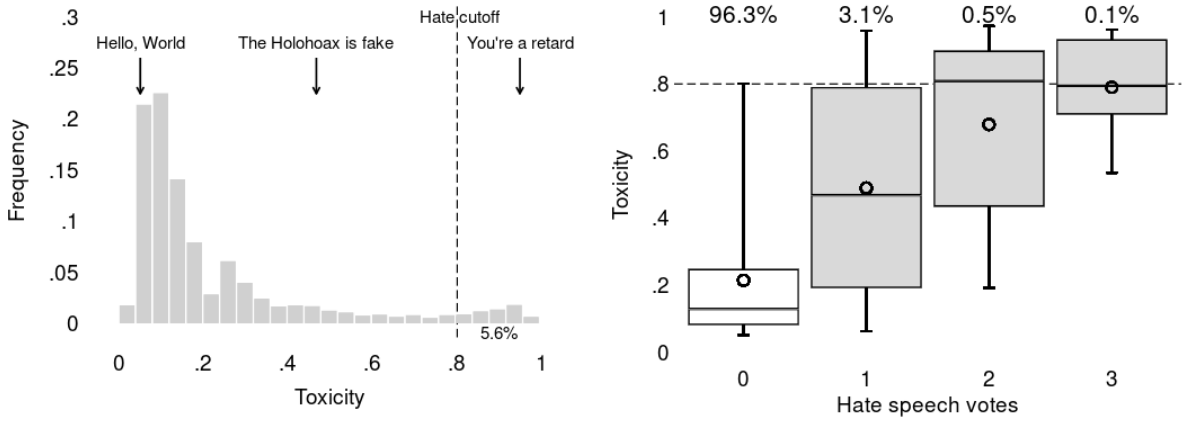
	Mean	Std. Dev	p10	Median	p90	Obs.	Sample
Accounts							
Account years	5.24	3.82	1.17	4.04	11.21	191,835	Random
Tweets per day	12.02	39.26	0.23	4.35	29.62	191,835	Random
Likes per day	13.03	24.08	0.07	3.98	36.16	191,835	Random
Followers	4,804	169,343	15	340	3,167	191,835	Random
Followed	1,071	6,755	45	381	2,078	191,835	Random
Is bot (%)	9.90	29.88	0.00	0.00	0.00	1,000	Random
Age 18-29 (%)	30.99	46.25	0.00	0.00	100.00	2,463	ATP
Male (%)	53.20	49.91	0.00	100.00	100.00	2,464	ATP
White (%)	57.92	49.38	0.00	100.00	100.00	2,464	ATP
College graduate (%)	40.47	49.09	0.00	0.00	100.00	2,464	ATP
Republican (%)	20.72	40.54	0.00	0.00	100.00	2,464	ATP
Democrat (%)	35.16	47.76	0.00	0.00	100.00	2,464	ATP
Tweets							
Is reply or quote (%)	62.53	48.40	0.00	100.00	100.00	201,038	Random
Is toxic (%)	5.61	23.01	0.00	0.00	0.00	201,038	Random
Is hate (%, majority)	0.56	7.47	0.00	0.00	0.00	9,991	MTurk
Is hate (%, consensus)	0.11	3.32	0.00	0.00	0.00	9,991	MTurk

Notes: The random sample indicates a random extraction of 201,308 Tweets from Twitter's API on August 2020. The bot score was computed on a subsample of 1,000 accounts from the random sample of Tweets, due to rate limits from the Botometer API. The ATP sample is a subsample of Twitter users from the Pew Research Center's ATP. The MTurk sample is a random subsample of Tweets that I annotated on MTurk following Davidson et al. (2017).

Prevalence of hate speech. The random sample of Tweets allows me to quantify the percent of Tweets that are hate according to different measures. Using the 0.8 toxicity cutoff, I find 5.6% of Tweets are hate. To compare this number with human annotation, I annotated a subsample of 10,000 Tweets from the random sample. As Table 1.1 shows, less than 1% of Tweets are considered hate speech using human annotation, under both the

majority decision rule and the consensus rule. Thus, hate speech is a low-probability event.

The long-tailed nature of hate is more evident in Figure 1.3a, which plots a histogram of the toxicity score in the random sample of Tweets. The figure also includes the toxicity scores of three example texts: the neutral phrase "Hello, World" (toxicity = 0.05), one phrase related to disability (toxicity = 0.95), and one that denies the Holocaust (toxicity = 0.47). These examples, which are relevant for the reporting experiment, illustrate that the toxicity cutoff adequately separates some slurs from neutral expressions, but it fails to identify more subtle hate. Still, toxicity is closely correlated with human annotation. Figure 1.3b shows the distribution of toxicity scores shifts to the right as more workers label Tweets as hate speech.



(a) Toxicity histogram **(b)** Toxicity by hate votes

Figure 1.3: Toxicity scores and annotation in a random sample of Tweets

Notes: Panel (a) displays a histogram of toxicity scores based on a random sample of 201,038 Tweets from August 2020. The dashed line is the 0.8 toxicity cutoff to classify hate speech; 5.6% of Tweets have a toxicity above that cutoff. The phrases "Hello, World," "The Holohoax is fake," and "You're a retard" have toxicities of 0.05, 0.47, and 0.95, respectively. Panel (b) has toxicity box plots by the number of workers who voted that a Tweet is hateful. The data is from a subsample of 10,000 Tweets annotated by MTurk workers. The boxes indicate percentiles 25, 50, and 75; the circles indicate the means; and the lines indicate percentiles 5 and 95. The percentages at the top indicate the fraction of Tweets by number of votes.

Occurrence of sanctions and reports. Table 1.2 presents the fraction of removals and suspensions in the random sample of Tweets and the different subsamples of hate speech.

Depending on the measure of hate, the fraction of Tweets that Twitter removed or suspended within one month is 2.6% to 9.1%—higher than the 2% in a random sample. These numbers match the statistics recently revealed in Facebook's whistleblower event, that the platform removes 3% to 5% of hateful content (Giansiracusa, 2021). From this table, we can also see that removals are a rare event. I did not measure shadowbans in this sample, but Merrer et al. (2020) document that 2.3% of accounts are shadowbanned. Figure A.9 in the Appendix plots the fraction of sanctions by the type of rule violation; hateful conduct and harassment are the most commonly sanctioned violations in the platform.

Table 1.2: Likelihood of sanctions by subsample

	Random	Hate speech		
		Toxicity ≥0.8	MTurk annotation	
			Majority	Consensus
Removal	0.01	0.1	0	0
Suspension	1.9	2.5	3.6	9.1

Notes: This table shows the fraction of Tweets or accounts that get removed from the platform within 1 month of posting hate speech by each subsample. The random sample of posts is based on 201,038 Tweets and the MTurk annotation is based on a subsample of 9,991 annotated Tweets.

In the second half of 2020, 11% of active accounts were reported according to official Twitter data,[25] and 1% of accounts concentrate the majority of reports (Twitter, 2018). Recently, Twitter's CEO reported that algorithms detect 51% of the content that the platform finds in violation of the rules and that the company's goal is to increase this percentage to 90% (Melendez, 2020). Users can report content even if they are not its targets; in a small study by a nonprofit, 57% of reports were filed on behalf of someone else (Matias et al., 2015).

25. This number results from dividing the total number of accounts reported, from the Rules Enforcement Report (Twitter, 2020b), by the monetizable daily active usage published on the letter to shareholders from Q4 2020 (Twitter, 2020a).

1.4 Reporting Experiment

1.4.1 Experimental Design

Sample. I sampled 6,148 Tweets containing hateful keywords during July and August 2020. I collected the Tweets every day with an algorithm that uses the search function of Twitter's API, which queries a subset of recent English-language Tweets excluding Retweets.[26] I searched posts containing two slurs: one that denies the Holocaust (Holohoax), and a disability slur (retard), the latter constituting 98% of the sample. Both terms are prevalent on social media and considered by many to be hate speech; see Guhl and Davey (2020) and Sherry (2019). Even if some people use the disability slur frequently (Albert et al., 2016), it is precisely the removal of this type of slurs that is controversial and policy-relevant. Moreover, Twitter's hateful-conduct policy covers the Holocaust and slurs that reinforce negative stereotypes about a protected category, which includes disability (Twitter, 2021b).[27]

Because the disability slur has alternative meanings, for example, to retard the progress of something, I refine the search with sentence structures such as, "You are a retard." This refinement captures directed hate speech (ElSherief et al., 2018) and facilitates identifying the targets of hate speech. I also consider multiple misspellings and word distortions to sample Tweets that attempt to bypass detection algorithms. Table A.1 in the Appendix contains the full list of queries used to search Tweets.

After my search algorithm detects a Tweet, it filters users to increase the quality of the sample and to reduce false positives, that is, Tweets that are not hate speech even if

26. The algorithm conducted the search every 20 minutes. This timing allowed the data processing to be spread throughout the day to comply with the API's rate limits.

27. In a pilot study, I included a broader list of slurs about race, ethnicity, religion, gender and sexual orientation. However, the sample contained many false positives, because most slurs are used by the members of the group that they target. Bianchi (2014) refers to this practice as appropriated or reclaimed uses of slurs. The two keywords that I use seem, anecdotally, to have lower false positives.

they contain the slurs. The filter drops users who self-report being under 18 in their profile biographies, those with new accounts (opened less than 2 weeks before the Tweet), inactive users or non-English speakers (with less than 10 posts in English and more than 50% of posts in another language), and bots (those with a Botometer score higher than 0.5). I also exclude users who display their preferred pronouns on their profile biographies,[28] Tweets that enclose the slurs in quotation marks (to capture users who are only referring to the slur), and those in the Holocaust sample who self-report being Jewish in their biographies. Users enter the sample once, so the filter also drops Tweets from duplicate users. This way, every observation in the sample is a user-Tweet pair, and I report users at most once.

At midnight every day, immediately before randomization into treatment, my algorithm checks whether the users or Tweets collected the previous day were removed from the platform; only those that have not been removed at this point enter the final sample. Table 1.3 compares descriptive statistics between the experimental sample and the random sample of Tweets from section 1.3. These samples are quite different. Experimental subjects have more recent accounts, give more likes per day, and are more likely to have posted toxic Tweets in the past. Tweets in the experimental sample are more toxic, as expected. The Holocaust and disability samples are also different; for example, the Tweets and timelines of users from the Holocaust sample have a lower toxicity. Figure A.6 in the Appendix plots the most common topics in each subsample, which I obtained by annotating the Tweets on MTurk. Some common topics include politics, religion, sports, and COVID-19.

Treatment. Figure 1.4 summarizes the experimental design and the timing of the algorithms involved. Every day at midnight, my algorithm randomly splits users or Tweets sampled in the previous 24 hours, who have not been removed from the platform, into a control or a treatment arm. The assignment is stratified by sampling date and slur; every

28. Arguably, these users might be more empathetic and more likely to refer to the slurs rather than use them to attack.

Table 1.3: Characteristics of the reporting experiment sample

	Means			Difference t-statistic	
	Full Sample	Holocaust	Disability	Random-Full	Hol.-Disab.
Observations	6,148	123	6,025		
Accounts					
Account years	3.22	3.29	3.22	40.2	0.2
Tweets per day	11.62	19.69	11.46	2.2	3.7
Likes per day	24.17	33.64	23.98	-32.3	2.1
Followers	634.85	1,436.41	618.49	2.1	1.7
Followed	433.75	554.98	431.27	7.6	1.6
Initial shadow ban	0.71	0.71	0.71		0.1
Tweets					
Word count	15.98	23.98	15.81	-14.1	6.8
Is toxic	0.80	0.06	0.82	-244.3	-22.0
Is hate (MTurk)	0.30	0.43	0.30	-63.6	3.1
Is reply	0.84	0.56	0.84	-48.4	-8.4
Is attack (MTurk)	0.78	0.24	0.79		-14.8
Is quote	0.07	0.02	0.07	7.9	-2.0
Is mention	0.85	0.67	0.85	-42.5	-5.5
Tweet from phone	0.79	0.49	0.80	-9.0	-8.3
Timelines					
Previous toxicity	0.93	0.69	0.94	-28.2	-11.1
Previous disability	0.39	0.15	0.40	-179.1	-5.6
Previous Holocaust	0.10	0.66	0.09	-6.0	21.9

Notes: This tables presents means of characteristics in the reporting experiment sample and subsamples. It also presents t-statistics from tests of difference in means between the random and the experimental samples and between the Holocaust and disability subsamples.

day, half of the Tweets using each slur enter each experimental arm. Users in the control arm do not receive any intervention. The treatment consists of reporting Tweets for violating Twitter's rules against hateful conduct on the next day after they enter the sample, so Tweets can be reported between five and 48 hours after they are posted. Every day, my algorithm assigns the Tweets in the reporting arm evenly to one out of the 11 accounts that I use for reporting. Table A.4 in the Appendix displays summary statistics of the accounts that I used for reporting and Figure 1.5 includes screenshots of the reporting process.[29]

Table A.3 in the Appendix shows that the two experimental arms are balanced in pre-

29. When reporting Tweets, I click "It's abusive or harmful," then "It directs hate against a protected category (e.g., race, religion, gender, orientation, disability)." Due to logistics, 1% of the reported subjects were reported using a different account than the one that was assigned at the moment of randomization.

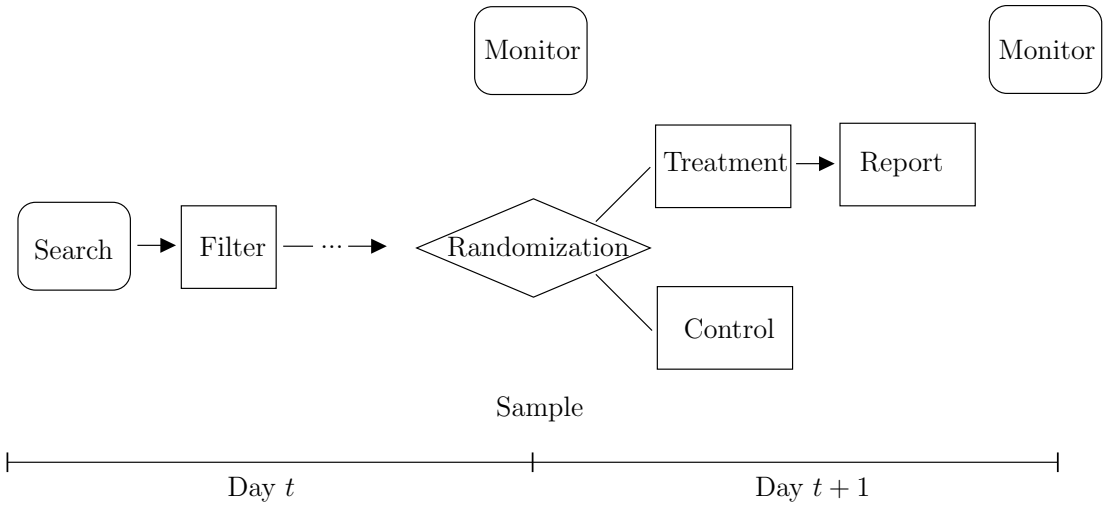
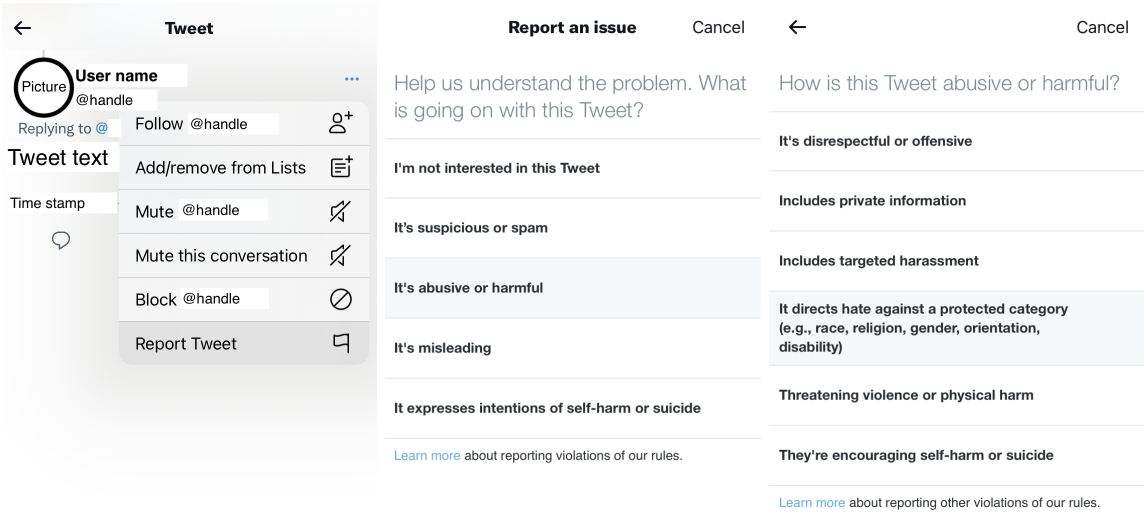


Figure 1.4: Design of the reporting experiment

Notes: Two main programs collect the sample and outcomes. The search program looks for hateful posts every 20 minutes and the monitor program keeps track of user activity every day. Randomization takes place at the beginning of every day with the sample of users collected the previous day.



(a) Tweet options **(b)** Reporting screen 1 **(c)** Reporting screen 2

Figure 1.5: Procedure to report Tweets

treatment characteristics. Normalized differences—for each characteristic and all of them jointly—are well below the 0.25 value that Imbens and Rubin (2015) suggest. This balance confirms randomization was successful. I did not report 3.26% of the Tweets that were assigned to the reporting arm; that is, there is one-sided non-compliance. Three percent of

Tweets in the reporting arm disappeared after treatment assignment and before I could report them, because users deleted them or deleted or protected their account, or because Twitter deleted the Tweets or suspended the users. Additionally, I did not report eight Tweets (0.26% of the Tweets in the reporting arm) that were clearly not hate speech.[30] Because of this one-sided non-compliance, the estimates can be interpreted as an intention-to-treat (ITT), but I also give instrumental variable estimates that account for non-compliance.

Reports are an instrument for content moderation, that is, for receiving any sanction from Twitter. First, Twitter uses reports to detect content and enforce its rules, which implies the relevance and monotonicity conditions of instrumental variables hold. Second, reports only affect user behavior through their effect on sanctions, so the exclusion restriction holds if sanctions are perfectly observed. To the best of my knowledge, Twitter does not notify users they have been reported.[31]

Outcomes. I measure two types of outcomes: first-stage outcomes are the sanctions that Twitter enforces on users and second-stage outcomes are the users' activity on Twitter, their hatefulness, and spillovers to the activity of others. These outcomes allow testing whether reports influence moderation, whether Twitter's sanctions moderate users, and whether sanctions affect other users. I construct these outcomes with data that my algorithm collects every day. I gather users' cumulative number of Tweets, likes, accounts followed, and followers. I also collect the 100 most recent Tweets of each user (posted within 24 hours), and select 20 Tweets at random per user to compute their toxicity score by calling Perspective's API.

30. For example, one user quoted some people using the disability slur to refer to him or her. Other users posted the Holocaust-denial term quoting a study that was published around those dates (Center for Countering Digital Hate, 2021).

31. Some users have received notifications from Twitter saying their posts were reported. According to the survey of section 1.5, 9% of users have received a notification that someone reported their Tweets. However, users seem to receive these notifications only when an account from Germany reports content, due to the Network Enforcement Act. Figure A.3 in the Appendix has a screenshot of one of these notifications.

I also measure whether Twitter sanctions the Tweets in the sample or their authors. Three sanctions are observable: Tweet removals or deletions, user suspensions, and shadowbans. I measure these outcomes as an absorbing state; that is, once users receive a sanction, they remain sanctioned. By construction, at the time of entering the sample, none of the Tweets have been removed by Twitter and none of the users are suspended; however, 71% of users are initially shadowbanned.

I measure activity on Twitter as the time that users spend posting or liking Tweets, which corresponds to t in the model of section 1.2. I do not directly observe time spent, but I construct a proxy using the number of Tweets that users post (that is, the statuses count object from the API) and the number of likes that they give (that is, the favorites count of the API). I then approximate the total number of words that users wrote and read during the period, by multiplying the Tweets and likes times the average number of words per Tweet in the random sample of Tweets, which is 13.81. Then, I convert words into time by using the average reading and typing speeds that have been documented in the literature.[32]

The main measure of hatefulness is the fraction of Tweets with a toxicity score higher than 0.8, but I consider alternative measures for robustness. Spillovers focus on the time spent by the users to whom the Tweets in the sample are replying ("replied users"); 86% of Tweets in the sample are replies to others. I focus on replied users because, arguably, users mentioned in a Tweet are more likely to notice sanctions related to the Tweet than others. Figure A.5 illustrates a reply to another Tweet.

Empirical strategy. This paper reports cross-sectional estimates of the effect of reporting users on different outcomes, three weeks after treatment assignment. I focus on first-stage

32. I use the words per Tweet from the random sample, as opposed to the value from the experimental sample, because this is the value that I pre-registered, before the experimental sample existed. The average typing speed on a desktop computer is 51.56 words per minute (WPM) according to Dhakal et al. (2018). The average typing speed on a mobile device is 36.2 WPM (Palin et al., 2019). Elliott et al. (2019) estimate a reading speed of 179 WPM that is constant across different devices and screen sizes. I obtain the device of a user from the source object of Tweets; I consider the device to be a desktop when the source is "Twitter Web App" and mobile for all the other sources.

and ITT estimates because, as the next subsection shows, I find evidence of unobservable sanctions which means that reports violate the exclusion restriction. In other words, reports affect outcomes not only through their impact on observable sanctions, but also through unobservable sanctions. Thus, I estimate regressions of the form:

$$Y_i = \alpha + \beta Z_i + \delta X_i + \varepsilon_i, \tag{1.6}$$

where i indexes user-Tweet pairs, Y_i denotes first-stage or second-stage outcomes, Z_i denotes treatment assignment (reports), and X_i is a vector of controls. I estimate specifications without controls, controlling for stratum—sampling date and slur—fixed effects, and adding controls from the rich set of pre-treatment characteristics of Table A.3. I select controls with a two-step method using lasso as suggested in Urminsky et al. (2016) with the methodology of Belloni et al. (2014). Regressions use robust standard errors unless noted otherwise.

I also estimate dynamic treatment effects, which is possible because my algorithm collects outcomes every day after users enter the sample. I use the efficient estimator proposed by Roth and Sant'Anna (2021), which is robust to heterogeneous treatment effects. Because reports are randomized every day, the design satisfies their assumptions of random treatment timing and no anticipation.[33] I use their method to obtain event-study estimates, in which the event date is the number of days since a report. I construct the estimates on balanced panels but also report the treatment effect on attrition. The results use their Neyman-style pointwise confidence intervals and the sup-t confidence bands of Montiel Olea and Plagborg-Møller (2019).

33. These assumptions hold within each stratum (sampling date and slur). However, since the Holocaust denial slur has few observations per day, I compute the estimators within each sampling date, pooling observations from both slurs.

1.4.2 Results

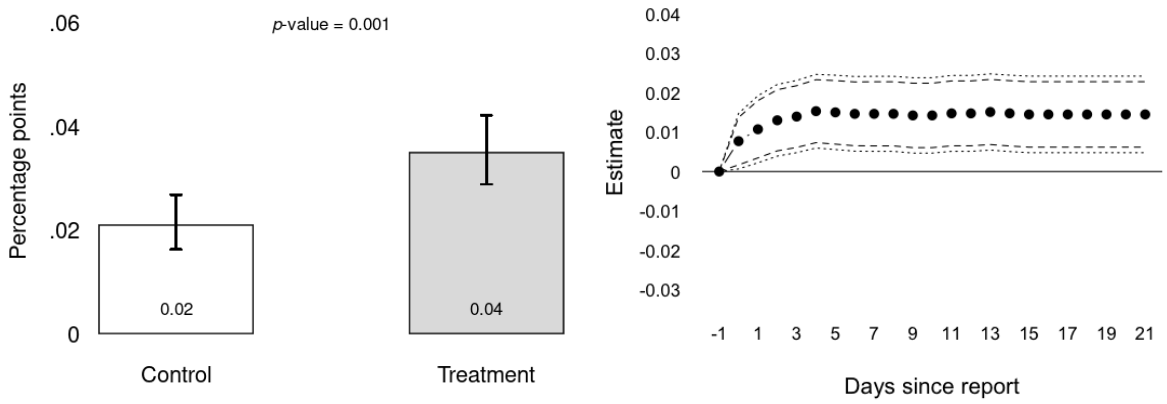
Sanctions. Reporting Tweets increases the likelihood that Twitter deletes them. Figure 1.6a shows the impact of assignment to treatment on the likelihood that Twitter removes the Tweets in the sample. Twitter removed 2.1% of the Tweets in the control arm within three weeks (21 days) after they entered the sample, and it removed 3.5% of them in the treatment arm. The treatment effect is 1.4 percentage points, which is 0.08 standard deviations, or a 66% increase.[34] The p-value of the difference in proportions is 0.001, and the F statistic from a regression of deletions on treatment assignment is 11.01.[35] Figure 1.6b displays dynamic treatment effects over event time; that is, the number of days since assignment to treatment. The dependent variable indicates whether Twitter removed Tweets at or before each event date. This figure shows that reports induce Twitter to remove Tweets within the first four days.

Table A.6 in the Appendix shows estimates of the effect of reporting on other Twitter sanctions and user self-censorship. Reporting does not significantly influence the other observable sanctions; that is, suspensions or shadowbans. The table also displays insignificant effects on the likelihood of users deleting their own posts or accounts, or protecting their accounts (making them private) within three weeks after reporting. Moreover, it shows reporting does not change the likelihood that other Tweets in the users' profiles go missing, which includes self-removals and Twitter removals.[36] The null effects persist after adding strata fixed effects and other controls, and the size of all estimates is below 0.033 standard

34. These numbers include cases in which Twitter required the removal of a Tweet but the user did not remove it within the three weeks. Eleven percent of Tweets were not removed by users in the control arm, and 5% were not removed in the treatment arm.

35. These estimates keep all users, even those whose accounts were deleted. Results are unchanged if we drop them. Results from a two-stage least-squares regression that uses treatment assignment as an instrument for reports are the same.

36. These numbers include the Tweets that users post after the sampling date and up to three weeks after the end of the sampling period. For these Tweets, distinguishing user deletions from Twitter deletions was not possible due to the API's rate limits.



(a) Means **(b)** Cumulative dynamic treatment effects

Figure 1.6: Likelihood that Twitter removes a post

Notes: Panel (a) displays the mean and 95% confidence intervals of the likelihood that Twitter removes a Tweet in the three weeks after reporting by treatment arm. The p-value is from a test of proportion differences. Panel (b) presents cumulative dynamic treatment effects of the likelihood of deletions, point-wise confidence intervals (dashed), and sup-t simultaneous confidence bands (dotted). Dynamic effects use the estimator from Roth and Sant'Anna (2021).

deviations.

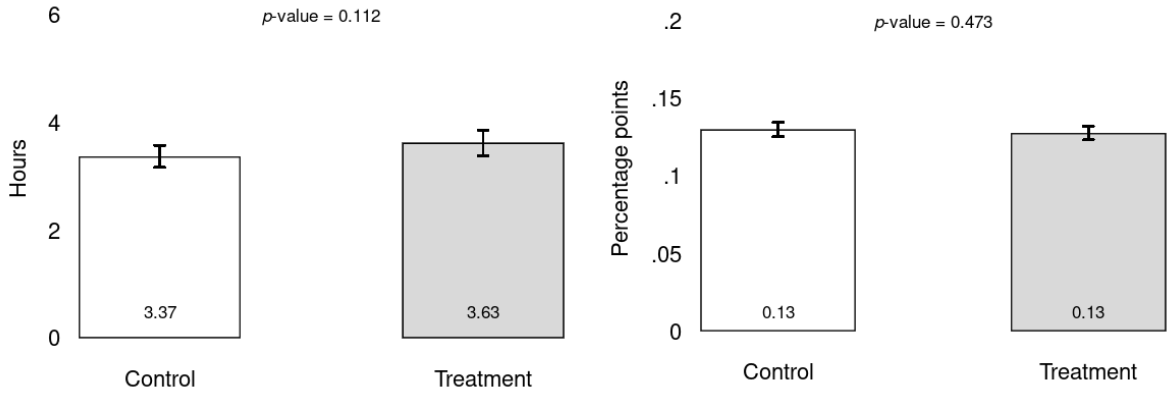
Evidence of unobservable sanctions exists, however. Twitter sent updates informing me it found that 270 (8.8%) out of the 3,074 accounts on the reporting arm violated the rules, within three weeks of the reports (see Figure A.4 for an example). One hundred fourteen (42%) of these updates were not accompanied by Tweet deletions, user suspensions, or shadowbans, which means 6.2% of reports led to an update but not an observable sanction. Some unobservable sanction is likely in these cases, such as accounts being temporarily locked (see Figure A.2f for an example). Moreover, that percentage likely understates the true number of unobservable sanctions, because Twitter does not always send updates whenever it imposes a sanction. For instance, Twitter sent me updates only for 13.4% of the 1,162 accounts in the reporting arm that received an observable sanction. Overall, I received updates on 12.52% of my reports. As a benchmark, an exercise conducted by the European Commission (Reynders, 2020) observed that Twitter sent an update on 26% of reports filed by general users.

Figure A.10 provides additional evidence of unobservable sanctions; it plots daily treatment effects on the number of hours since the last post, computed at midnight. The treatment effect is positive and pointwise significant around day 10 after reporting, although not significant with the simultaneous confidence bands. This figure suggests that reporting slightly increases the gap in between posts, which indicates that users might have had their accounts locked, although the daily number of hours since last post might not reflect locking periods of less than 24 hours.

Activity. Reporting does not significantly decrease user activity on Twitter. Figure 1.7a displays the treatment effect on the number of hours that users spend posting and liking Tweets in the three weeks after reporting. Both treatment and control spent around three and a half hours, and the treatment effect is 0.25 hours (5 minutes per week), which is a 7.5% increase or .042 standard deviations. This effect, however, is not significant at conventional levels, because the p-value of the difference in means is 0.11. Figure A.11a in the Appendix shows that treatment effects remain flat throughout the period. Table A.7 shows regression estimates using alternative measures of activity: Tweets and likes separately, a winsorized measure of time spent online removing the top and bottom percentiles, and an extensive-margin measure (the fraction of days that users post, like, or follow someone). The results remain unchanged using these alternative measures; if anything, the effect on Tweets is positive and significant at the 10% level under some specifications. Moreover, these estimates are mechanically biased downward since Twitter might temporarily lock user accounts.

Hatefulness. Reporting does not significantly decrease the likelihood of posting hate on Twitter. Figure 1.7b shows that the fraction of hateful Tweets (toxicity bigger than 0.8) that users post in the three weeks after the treatment is the same for both experimental arms. The treatment effect is -0.02 percentage points of hateful Tweets, which is a 1.7% decrease (-0.02 standard deviations). Figure A.11b shows a decrease in hatefulness in the first three

to five days after reporting (pointwise significant), but the effect returns to zero by the end of the three weeks. Table A.8 considers other measures of hatefulness; two extensive-margin measures (whether users post any Tweet with toxicity ≥ 0.8 or they repeat the slur), the average toxicity, and the average severe toxicity (another measure developed by Google). None of these measures yield significant effects, and the treatment effect is less than 0.011 standard deviations across all variables using different specifications.



(a) Hours spent on Twitter **(b)** Fraction of hateful posts

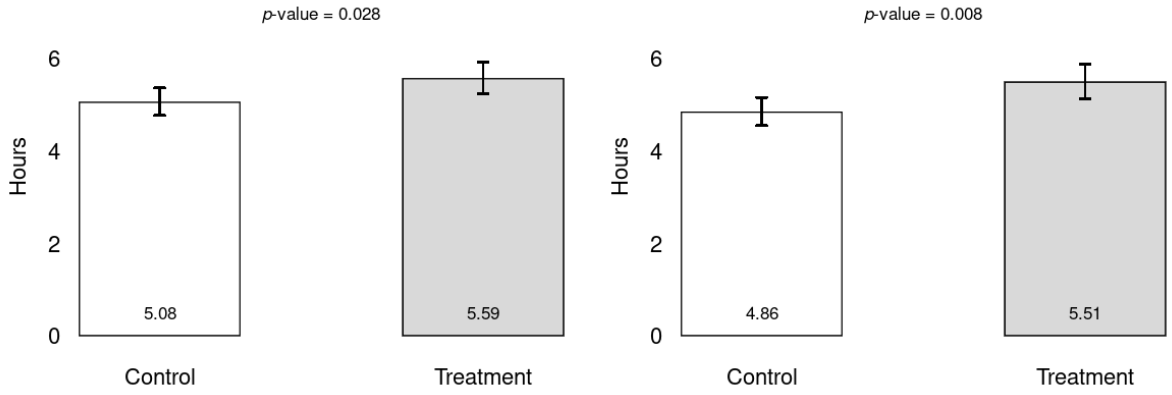
Figure 1.7: Hours spent on Twitter and fraction of hateful posts

Notes: This figure displays means and 95% confidence intervals of outcomes in the three weeks after reporting by treatment arm. Hours spent is calculated using statuses and favorites. Hateful posts are those with toxicity higher than 0.8. The p-value is from a test of difference in means.

Spillovers. Even if reporting does not seem to moderate the authors of the Tweets, that is, decrease their activity or hatefulness, it impacts other users. Figure 1.8a shows that reporting increases the time the replied users spend Tweeting and liking by 0.51 hours, which is 10 minutes per week, 10%, or 0.064 standard deviations (p-value = 0.028). The treatment effect seems persistent; Figure A.12 shows the cumulative effect increases continuously after the reporting day.

Although many of these replies are attacks, some are replies between social media friends. As specified in the pre-analysis plan, I asked MTurk workers to read the context of both

posts and classify whether the replies in my sample were attacks on the replied user. Under the majority decision rule, in which Tweets are attacks if the majority of workers agree, 87% of replies were attacks on others. Figure 1.8b shows that the effect of reporting is stronger among attacks; it is 0.65 hours (13 minutes per week, a 13.4% increase or 0.08 standard deviations, p-value $= 0.008$). Table A.10 shows estimates using the same alternative measures of activity as above. Results remain significant at the 5% level across specifications considering Tweets and likes separately or winsorizing time spent.[37] Hence, reporting seems to increase the activity of those users that are attacked by the Tweets in the sample.



(a) All replies **(b)** Attacks

Figure 1.8: Spillover on the time spent of users replied by the posts

Notes: This figure displays means and 95% confidence intervals of the time spent on Twitter in the three weeks after reporting by treatment arm. Panel (a) includes all users replied by the Tweets. Panel (b) includes users that were attacked by the Tweets, according to MTurk annotators. The p-value is from a test of difference in means.

Attrition. In this experiment, attrition occurs because accounts go missing after treatment assignment; 7% of them were missing after three weeks. Attrition happens when users delete their own accounts or Twitter suspends them. Given that the previous results showed no treatment effect on account suspensions or on the likelihood that users delete their ac-

37. The findings are also robust to dropping those users who were replies to more than one Tweet in the hateful sample. More than 93% of the replied users corresponded to a single user in the hateful sample, so concerns about SUTVA violations are minimal.

counts, finding no differential attrition by treatment arm is not surprising. Table A.11 shows insignificant treatment effects on the likelihood that users leave the sample at the end or on any day of the three weeks after users enter the sample. Figure A.13 shows dynamic treatment effects on attrition; the effect is not significant pointwise or with the simultaneous bands.

Heterogeneity. The pre-analysis plan specified two dimensions for the heterogeneity analysis, besides attacks: by slur (Holocaust vs disability) and by human annotation (among the hate sample). I report these results in Figure A.14 in the Appendix due to their low informational content.[38]

1.4.3 Interpretation

The previous results indicate that reports instrument for sanctions, particularly Tweet removals and, potentially, unobservable sanctions. Moreover, the treatment did not decrease user activity or the likelihood of posting hate within three weeks; reports did not moderate users. The effect on the users' activity is insignificant, which I interpret as a low elasticity of time spent with respect to moderation among the users in my sample; $\partial T^H/\partial c \approx 0$ in the notation of the model of section 1.2.

Yet, reports spill over to other users; they increased the amount of time that the attacked users spent posting and liking. I interpret these findings as evidence of a positive elasticity of the time spent of some users in my sample with respect to moderation; $\partial T^A/\partial c > 0$ in the notation of the model of Section 1.2.

38. The experiment is not powered to detect the effect on the small Holocaust sample. The heterogeneity analysis by human annotation was intended to capture measurement error (false positives) in the sampling of hate speech. More than false positives, these labels seem to capture heterogeneity due to the subjective nature of hate speech. Thirty percent of Tweets in the sample were labeled as hate speech by the majority of annotators, 61% were considered offensive, 1.6% were not considered offensive or hate, and the remaining did not have a majority label. Hence, splitting the sample between "hate" and "not hate," as preregistered, captures the difference between hateful and offensive Tweets.

Three main mechanisms may explain why reports impacted the replied users. First, the reported users could have changed their behavior or their interactions with the replied users. Second, if Twitter removed the Tweets, the replied users could have noticed the legends that Twitter placed on the Tweets, as in Figure A.2a. Third, if the replied users also reported these Tweets, Twitter could have sent them an update on their reports, as in Figure A.4.

Regarding the first mechanism, the results in subsection 1.4.2 rule out that the reported users substantially changed their behavior. Additionally, Figure A.15 shows an insignificant effect on the likelihood that the users in the sample mention the replied users again within three weeks. Hence, the evidence in favor of this mechanism is weak. The same is true for the second mechanism. Table A.12 shows that the treatment effect on deletions is smaller in the sample of replies relative to the full sample, and insignificant in the sample of Tweets that attack others.

As for the third mechanism, the percentage of reports for which I received an update and found no observable sanction is similar in the full sample, among replies, and among attacks (6.2%, 6.4%, and 6.5%, respectively). Hence, Twitter may have imposed an unobservable sanction (e.g., locking accounts) on the users who attacked others, and the attacked users who reported these Tweets may have received an update about the sanction.[39]

How does reporting affect monetization? I obtain a back-of-the-envelope estimate as follows. The treatment increased by 10-15 minutes per week the time that reported users and replied users spent liking and posting. The advertising load on a small sample of 50 Tweets was one ad per four regular Tweets. Assume this number translates into an ad load of 0.25 minutes per minute of content consumed. Twitter's Ad website has a default bid of $0.21 per six-second video advertisement.[40] Ignoring effects on others, the treatment

39. This hypothesis is difficult to test without access to internal data, because user reports are unobservable. Moreover, whether users would reveal that they reported a particular Tweet in a survey is unclear, even if reporting is common (indeed, the next section shows that one-third of users have reported content).

40. This price is for the general audience of U.S. adults. The ad price did not change when I tried targeting an ad to the list of users in the sample.

amounts to a $5.25-$7.88 increase in ad revenue per week per report.[41]

1.5 A Test of Overprovision or Underprovision

1.5.1 Experimental Design

Sample. I recruited 3,027 respondents in September 2021 through Luc.id, a widely used online marketplace that matches researchers with survey providers (Coppock and McClellan, 2019; Bursztyn et al., 2020). I pre-screened participants to select English speakers who live in the U.S., are over 18 years old, are willing to provide their email, self-report using Twitter, and pass a basic attention check. After the pre-screen, participants entered the online survey and had to answer demographic questions. The survey also asked them for their Twitter handle (optionally), which I used to get their account creation date, Tweet counts, and like counts. Sixty-four percent of participants provided a handle, and 74% of the handles were valid. This results in a sample size of 1,427 respondents, which satisfies the recommendation of Haaland et al. (2020) of 700 respondents per treatment arm.

Table 1.4 compares the characteristics of the sample with representative adult Twitter users from the ATP survey and with accounts from the random sample of Tweets. My survey undersamples users in the 18-29 age range, college graduates, and politically Independents, and oversamples white respondents and Democrats.[42] Users who provided their Twitter handle have an older account, and fewer Tweets and likes per day relative to accounts in a random sample of Tweets.

Afterward, the survey asked questions about social media use, online harassment, hate speech, and Twitter sanctions. These questions provide further insights about the previous

41. Besides being a rough estimate, this calculation is based on a selected sample and ignores equilibrium effects, so it does not imply that Twitter would like to increase reports. Moreover, these numbers do not consider the marginal costs of moderating.

42. I pre-registered introducing quotas to match representative Twitter users on gender, age, race or ethnicity, region, and political orientation, but relax the quotas to obtain the desired sample size was necessary.

Table 1.4: Characteristics of the welfare experiment sample

Panel A: Demographics, N= 3,027		
	Means (Survey)	ATP-Survey t-stat.
Age 18-29 (%)	24.48	3.01
Male (%)	53.88	-0.34
White (%)	68.19	-5.04
College graduate (%)	31.68	4.89
Republican (%)	22.76	-1.34
Democrat (%)	52.89	-9.41
Panel B: Twitter accounts, N= 1,427		
	Means (Survey)	Random-Survey t-stat.
Account years	7.93	-23.65
Likes per day	2.34	29.37
Tweets per day	1.54	41.76

Notes: This tables presents means of characteristics in the welfare experiment sample. It also presents t-statistics from tests of difference in means between the ATP or the random sample of Tweets, and the experimental samples.

experiment. Figure A.16 shows that the API-based measure of time spent on Twitter correlates closely with users' self-reported hours, so it is a good proxy measure. Table A.13 includes additional statistics. For instance, 32% of users have reported content for violating the rules, 10% have had a Tweet removed, and 5% have been suspended. Moreover, the experience on the platform differs by minority status, which I define based on religion (Jewish, Muslim, Buddhist, Hindu, or other), sexual preference (not heterosexual), gender (other than man or woman), and race (other than white). Consistent with other surveys (Anti-Defamation League, 2021), minorities are more likely to experience harassment online, to self-report seeing more hate speech in their feed, and to report content. However, they also receive more sanctions and reports, which, to the best of my knowledge, has not been documented before.[43]

43. This finding is related to the literature on racial biases in detection algorithms; see, for example, Cowgill and Tucker (2019).

Treatment. Figure 1.9 summarizes the experimental design. I use an information provision treatment with an active control group (Haaland et al., 2020). After the baseline questions, I randomize survey participants into two treatment arms that receive different information about the likelihood of moderation among hateful Tweets. The information provided comes from the annotated random sample of 10,000 Tweets. To vary the likelihood of moderation without deception, I use different decision rules to classify hate speech. As Table 1.2 shows, 3.6% of hateful Tweets are removed or their authors are suspended within one month of the post under the majority decision rule, that is, if most annotators agree. That percentage changes to 9.1% under the consensus rule, that is, if all annotators agree. Half of participants are randomized into the low-moderation arm (3.6%) and half into the high-moderation arm (9.1%). The treatment is stratified by whether respondents are male, minorities, and have been sanctioned by Twitter, and whether they provided a Twitter handle.

After randomizing participants, I inform them, for transparency, of the rule that I use to classify hate. As pictured in Figure A.7, I tell them that a crowd-sourced team of annotators identified hate speech using 10,000 Tweets, and that a Tweet is hate speech if [most/all] annotators label it as hateful. I then elicit their beliefs about (1) the prevalence of hate speech in this sample and (2) the fraction of Tweets that are removed or suspended within one month. These elicitations are incentivized, because they know that one participant with the closest guess will get a $50 Amazon gift card.

After the elicitation, I provide information about the likelihood of moderation, as displayed in Figure 1.10. I also hold constant the prevalence of hate speech in both arms, by telling respondents that less than 1% of Tweets are classified as hate (recall that 0.56% Tweets are hate under the majority rule and 0.11% are hate under the consensus rule). The message also shows that other popular platforms, such as YouTube, Facebook, and Reddit, have a similar prevalence of hate, according to different sources (Kennedy et al., 2020; Vid-

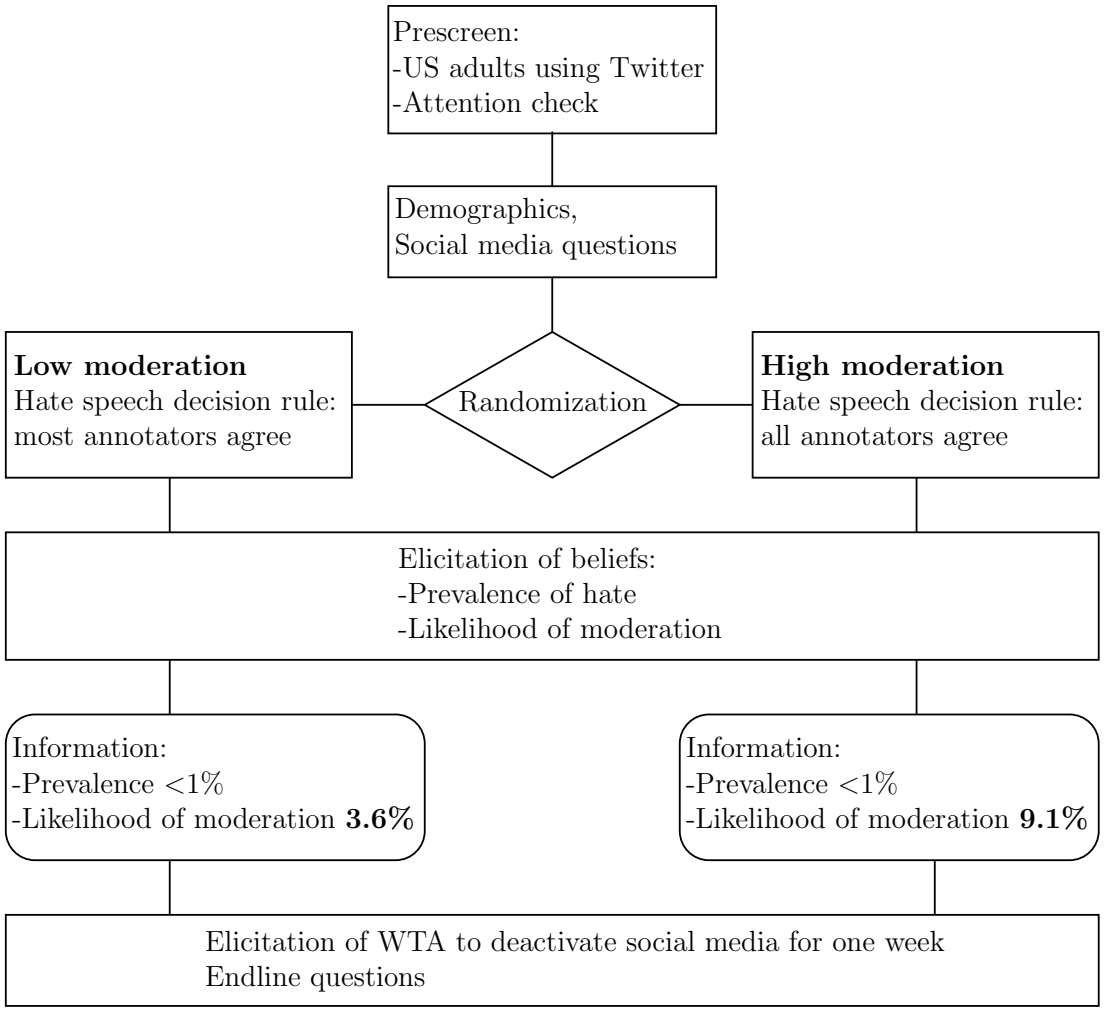
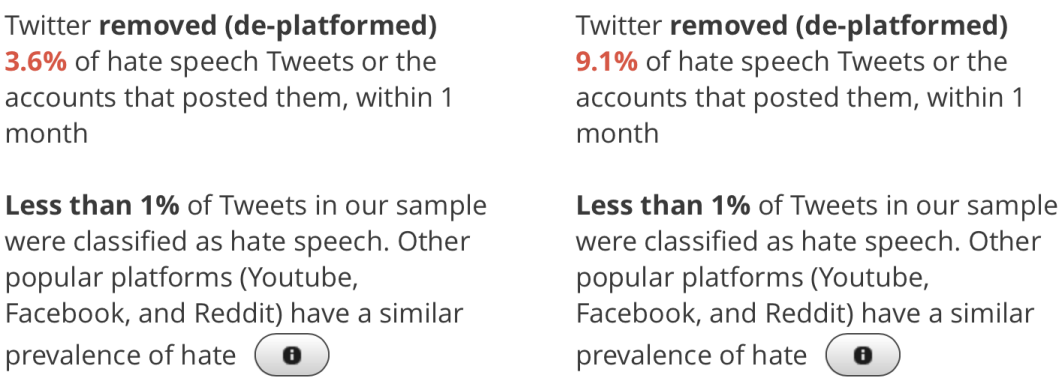


Figure 1.9: Design of the welfare experiment

gen et al., 2020; Facebook, 2021). They can consult the sources by clicking a button on this screen.

Table A.5 shows that both experimental arms are balanced on pre-treatment characteristics. The table also rules out that changing the decision rule to classify hate influences the participants' concept of hate; the treatment has no effect on the belief about the prevalence of hate or the likelihood of moderation.[44]

44. This finding is similar to what other studies obtain, such as Bottan and Perez-Truglia (2017), who argue that changing the source of information does not have an impact on participants who do not have expertise on the data.



(a) Low moderation **(b)** High moderation

Figure 1.10: Information provision by treatment arm

Outcomes. There are two outcomes of interest. Based on the results of section 1.2, the main outcome is the willingness to accept (WTA) to stop using social media, that is, Twitter, Facebook, Instagram, YouTube, Snapchat, TikTok, and Reddit, for one week. I first tell participants that the research team will conduct a small follow-up study that compensates some participants to deactivate their social media for one week. I inform them that similar studies have been conducted in the past (Hunt et al., 2018; Mosquera et al., 2020; Allcott et al., 2020). I then elicit their WTA with an iterative multiple price list (iMPL, see Harrison et al. (2005); Andersen et al. (2006)).[45] Subjects have to decide whether they are willing to stop using social media for different Amazon gift card offers. The first offer is for \$50, and subsequent amounts increase or decrease until the WTA is placed in intervals that go from $(-\infty, 0]$ to $[100, \infty)$ and increase by \$10, as Figure A.8 illustrates. I transform these intervals into a continuous measure using the triangular distribution procedure from Allcott and Kessler (2019).

This elicitation is incentivized. I inform respondents that a computer will randomly choose some eligible participants whom the research team will contact for the follow-up.[46]

45. The iMPL has two advantages over a regular multiple price list. First, it induces monotonicity on responses by construction. Second, it saves time by omitting reduntant questions.

46. Following Allcott et al. (2020), I did not tell participants the likelihood of being selected into the

If the participant is selected, the computer will also choose one of her answers at random. If the answer is "yes," the research team will ask her to stop using social media for one week and pay the offered amount. If the answer is "no," the participant will not be asked to stop using social media. This information is truthful; I recontacted 50 participants at random in October 2021 and implemented the follow-up study.[47]

The second outcome of interest is the API-based time spent on Twitter one week after the survey, which I compute for the participants who provided valid Twitter handles following the procedure outlined in section 1.4. At the end of the survey, I ask questions to measure attention, experimenter demand effects, and posterior beliefs.

Empirical strategy. The empirical strategy consists of OLS regressions of outcomes on an indicator of treatment status. All estimates use robust standard errors. I run regressions without controls, controlling for stratum fixed effects, and a specification adding controls as in Urminsky et al. (2016). As pre-registered, I report estimates of the main outcomes reweighting observations to match the ATP on first moments of gender, age, race or ethnicity, region, and political orientation, but I also report unweighted estimates. I obtain the weights using the maximum entropy approach of Hainmueller (2012).

1.5.2 Results

Misperceptions about hate speech and moderation. Most users overestimate the prevalence of hate speech on Twitter and the likelihood that Twitter sanctions hateful content. Figure A.17 displays histograms of beliefs among respondents. Ninety-six percent of

follow-up; previous research has shown that, at least on Becker-DeGroot-Marschak elicitations, informing participants can bias WTA estimates.

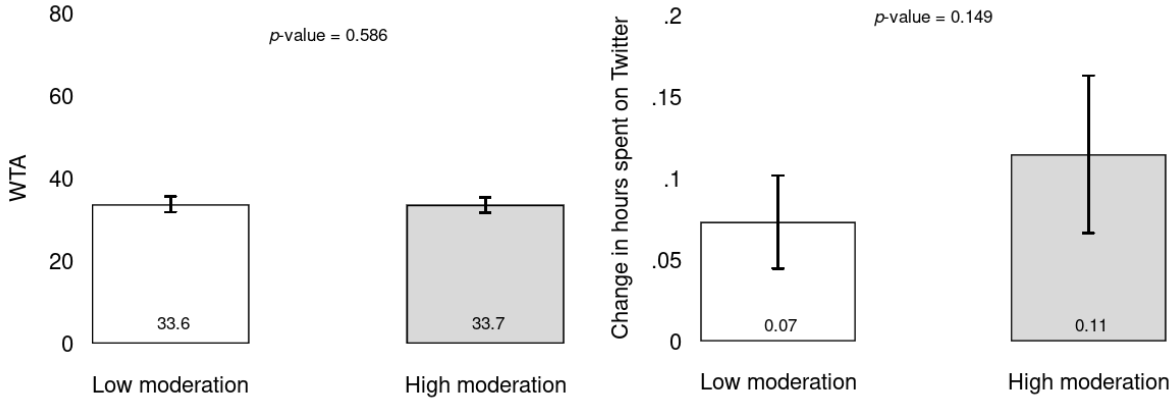
47. Thirteen participants replied to the recontact email. Seven of them had been randomized into the deactivation treatment, and six to the control group. I asked participants in the deactivation arm to upload screenshots of the time-tracking app of their phones as proof of deactivation, as in Hunt et al. (2018). Five out of seven participants self-reported that they had stopped using social media, and four submitted the screenshots.

Twitter users overestimate the prevalence of hate speech, that is, their belief is above 1%, and 84% guess a moderation rate above the higher 9.1% value. These results add another example to the literature on misperceptions about others (Bursztyn and Yang, 2021).

There are several explanations for these facts. An "echo-chamber" argument is that users might not notice what happens outside their curated feeds, which they personalize with the help of Twitter's algorithms. Consistent with this argument, I find that 74% of users believe that the prevalence of hate in the random sample of Tweets is higher than what they see in their feed. Platforms' lack of transparency might also contribute to misperceptions. Even Facebook, which publishes a substantial amount of information (Facebook, 2021), informs only about the prevalence of hate speech but not about the likelihood of moderation (Bradford et al., 2019). The only information about the likelihood of moderation, between 3 to 5% of hateful content, was revealed thanks to the recent whistleblower incident (Giansiracusa, 2021).

WTA to stop using social media. Providing information about a higher likelihood of moderation has little effect on the users' social-media valuation. Figure 1.11a displays the treatment effect on the WTA to stop using social media during one week. The average WTA was \$33.6 in the low moderation arm, and \$33.7 in the high moderation arm. The treatment effect is -15 cents per week, which is 0.004 standard deviations, or a 0.5% decrease. The null effect is not just on average; Figure A.18 in the Appendix shows that the cumulative distribution function of WTA is the same for both arms. Table A.14 presents regression estimates with alternative measures of social-media valuation. As in Allcott and Kessler (2019), I assume a uniform distribution of WTA beyond the endpoints instead of the triangular distribution. I also use -\$50 and \$150 for the endpoints as benchmarks, or a take-it-or-leave-it dummy for the first \$50 offer. The results remain unchanged using these alternative measures.

Activity. The information provision treatment has an positive but insignificant effect on the time that users spent on Twitter one week after the survey. Figure 1.11a plots the effect on the number of hours spent by users who provided their Twitter handle. The effect is 0.04 hours, which is 2.4 minutes (57% increase relative to the low-moderation arm, or 0.077 standard deviations).[48]



(a) Means of WTA **(b)** Means of hours spent

Figure 1.11: WTA to stop using social media and hours spent on Twitter

Notes: Panel (a) displays the mean and 95% confidence intervals of the WTA to stop using social media one week by treatment arm. Panel (b) presents the mean and 95% confidence intervals of the hours spent on Twitter one week after the survey. The p-values are from a test of difference in means and observations are reweighted to match Twitter users from the ATP on observables.

Posterior beliefs, attention, attrition, and experimenter demand. Respondent inattention cannot explain the previous null results; providing information significantly shifts participant's recollection of the information provided and their posterior beliefs about moderation. At the end of the survey, I asked participants to repeat the moderation rate that I gave them, and I incentivized the closest answer with a $50 gift card. Figure 1.12 plots the effect on the respondents' recollection of the moderation information. Sixty percent of par-

48. Table A.15 shows estimates using the same alternative measures of activity as in section 1.4; Tweets and likes separately, winsorized hours, and an extensive-margin measure of the fraction of days in which users post or like. The effect remains insignificant with these measures across specifications. Figure A.19 confirms that dynamic treatment effects remain flat throughout the week post-survey.

ticipants recalled a number within one percentage point of the true value.[49] The treatment effect on this recollection is 5.6 percentage points (53% or 0.425 standard deviations, with an F-statistic of 36), not statistically different from 5.5 (p-value = 0.907), which is the gap between the high moderation rate (9.1%) and the low moderation rate (3.6%).

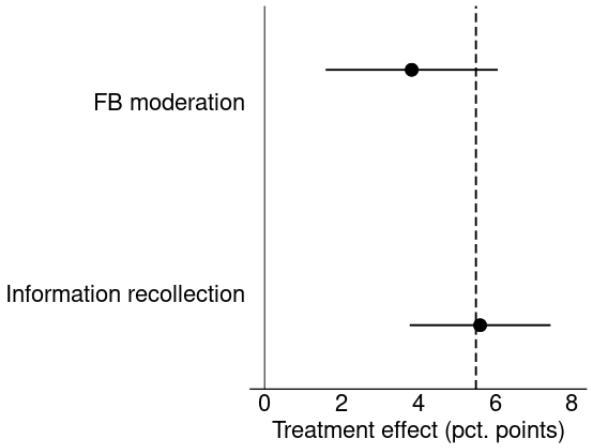


Figure 1.12: Posterior beliefs about moderation on Facebook and attention check

Notes: This figure presents coefficients and 95% confidence intervals of OLS regressions on an indicator of the high-moderation arm. FB moderation is the users' beliefs about the fraction of posts or users that Facebook moderates. The attention check is the participants' recollection, at the end of the survey, of the information provided about the moderation rate on Twitter. The dashed line is at 5.5 percentage points, the difference between the high moderation rate (9.1%) and the low moderation rate (3.6%). Observations are reweighted to match Twitter users from the ATP on observables.

Figure 1.12 also plots the treatment effect of information on users' beliefs about the likelihood of moderation on Facebook. This follows the recommendation of Haaland et al. (2020), of measuring posteriors by asking post-treatment beliefs about a related but different variable. The average belief of the moderation rate on Facebook was 19% for users in the low-moderation arm and 22.9% in the high-moderation arm. The treatment effect was 3.8 percentage points (20% or 0.16 standard deviations, with an F-statistic of 11.2).

Additionally, there is no evidence of differential inattention, attrition, or experimenter demand effects by treatment arm. Table A.16 presents insignificant treatment effects on inattention, measured as the absolute difference between participants' recollection and the

49. Because of left-digit bias, many participants in the low-moderation arm remembered 3%.

information provided. Thirty-four participants (1.1% of the sample) completed the pre-screening questions but did not finish the survey, and Table A.16 shows null treatment effects on attrition under different specifications. Following Allcott et al. (2020), the last part of the survey asked a question to test for experimenter demand effects: "Do you think the researchers in this study had an agenda?" Similar to that study, 57% of respondents in both arms thought I had no particular agenda or were not sure. Figure A.20 shows insignificant treatment effects on the responses to that question.

Heterogeneity. I do not find substantial heterogeneity of the effect on the WTA across most of the pre-registered covariates, including minority status (as defined above), whether participants have experienced a sanction on Twitter, and whether their beliefs are above or below the median moderation belief of 33% of hateful Tweets. The exception is the time spent on Twitter after the survey. Figure A.21 in the Appendix shows suggestive evidence that minorities spend more time on Twitter when they receive the high moderation information. The treatment effect in this subsample is 0.054 hours (three minutes, 100% increase relative to the control group, 0.17 standard deviations, p-value = 0.03).[50]

1.5.3 Interpretation

The previous results indicate that providing information about a higher moderation rate shifted users' beliefs, but had little impact on their social-media valuation. Taken at face value, these results mean that Twitter does not moderate too much or too little from the consumers' point of view, for a fixed prevalence of hate speech. One explanation for this finding is that Twitter internalizes the impact of moderation on users' willingness to pay for the platform, which requires that marginal and inframarginal users respond similarly to

50. The treatment effect among minorities is significant at the 10% without reweighting observations. Figure A.21 also shows large point estimates on the subsample of users who have been sanctioned and those with high prior beliefs, although these are noisily estimated.

sanctions.

Another option is that users do not directly care about moderation, holding constant the hate they encounter. Indeed, it is possible that the experiment did not change users' perceptions about hate in their own feed. In that case, users could have differentially updated their beliefs about how effective the algorithms are at hiding content without moderating. This is consistent with platforms providing a wide range of tools that allow users to customize their experience. For instance, Twitter allows users to mute and block accounts and words, and to hide sensitive content from their feeds.

One challenge to the interpretation of these findings comes from the welfare discussion in Allcott et al. (2020). They argue that users might misperceive Facebook's value, and thus the WTA might overstate consumer surplus. These value misperceptions could explain why increasing perceived moderation did not impact users' WTA. Another challenge is that the treatment not only shifted users' beliefs about moderation on Twitter; it also impacted beliefs about moderation on other platforms (at least Facebook). Based on Appendix A.1, the correct measure of the change in consumer surplus is to consider current social media users, not just current Twitter users. Table A.17 in the Appendix shows that results are unchanged after reweighting observations to match representative social media users, or without reweighting.

There is also suggestive evidence that the treatment increased minorities' time spent on Twitter. Given that these individuals are more likely to experience harassment online (Table A.13), this is consistent with the finding from the previous experiment that reporting increases the activity of the targets of hate speech.

1.6 Conclusions

Simple economics explain why it makes sense for profit-maximizing social media companies to ban some of their customers or restrict their content: because this increases the willingness

to pay of marginal users. In an advertising-driven business model, platforms remove content only if this increases the time that some users spend consuming content, and hence interacting with ads. I find evidence consistent with this implication, by running a natural field experiment in which I report content that violates Twitter's rules against hateful conduct. Reports increase Tweet removals and, potentially, unobservable sanctions, and they do not decrease user activity or hatefulness. Yet, the targets of hateful posts increase their activity after the reports. While this treatment provides some evidence of the behavioral effects of moderation, further work is needed to understand repeated sanctions, different classes of platform interventions, or the effects of moderation on other types of content.

In terms of policy, both sides in the discussion of how to regulate platforms often mention a tension between profit maximization and optimality of content moderation. While platforms can, in theory, remove too little or too much content relative to a surplus-maximizing planner, this study finds no evidence of distortions from the consumers' point of view. There are, however, two caveats to these findings. First, consumer surplus ignores the costs that hate speech imposes outside platforms. Hence, an avenue for future research is to examine the costs and benefits of the real-world consequences of content moderation. Second, even without moderation distortions, imperfect competition between platforms likely leads to pricing distortions, so they might be setting the ad loads of haters or non-haters suboptimally. These distortions can be empirically confirmed by future work.

CHAPTER 2

ESTIMATING THE DISTASTE FOR PRICE GOUGING WITH INCENTIVIZED CONSUMER REPORTS

2.1 Introduction

Emergencies like natural disasters or pandemics create ideal conditions for prices of essential products to increase. There is typically an increased demand paired with an inelastic short-run supply or even supply disruptions (Cavallo et al., 2014). In response, many states implement anti-price gouging laws that restrict price increases. Despite their wide adoption, these laws remain controversial among economists.[1] In perfect competition, artificially low prices can create shortages and cause markets to clear through other margins such as queues or search efforts (Becker, 1965; Barzel, 1974; Weitzman, 1991). However, these policies may improve allocative efficiency under imperfect competition.[2]

An additional factor that complicates welfare evaluations about these laws is that individuals may get disutility from others voluntarily trading essential goods at high prices. For example, individuals may find the prices to be unfair (Kahneman et al., 1986) or the transactions to be repugnant (Roth, 2007).[3] This disutility results in an externality, increasing the social cost of raising prices during a crisis.

This paper proposes a proxy measure of individuals' distaste toward price gouging and provides evidence on its mechanisms. We conduct what we call an Incentivized Reporting Experiment (IRE) in which we measure individuals' willingness to pay to report price goug-

1. Thirty-four states prohibit either increases above pre-crisis prices, 10-20% price increases, or "unconscionable" price increases. Surveys of economists' opinions on anti-price gouging laws can be found in `https://www.igmchicago.org/surveys/pricegouging` and `https://www.igmchicago.org/surveys/prices-of-medical-supplies`.

2. It is well known since at least Pigou (1920) that price controls can restore efficiency with monopolies (Bronfenbrenner, 1947).

3. The economic definition of a repugnant transaction is when third parties disapprove it and wish to prevent it even if participants are willing to take part in it (Roth, 2007, 2015).

ing. This method exploits that authorities rely on consumer reports to enforce anti-price gouging regulations, as with many crimes (Akerlof and Yellen, 1994). Through the lens of a model, we argue that reporting decisions reflect how much individuals expect to change their disulity from third-party transactions with their report and how much they value punishing sellers that post high prices. Thus, researchers can use this method to measure externalities in other settings that also rely on reports to enforce rules, or even more general settings where consumer actions (e.g., boycotts or negative reviews) impact others.[4]

We operationalize the (pre-registered) framed field experiment as a nationally representative survey distributed by a survey company, CloudResearch.[5] We develop an algorithm that combines text analysis and image recognition to make a list of PPE listed on Amazon. We randomize subjects into treatments where they make incentive-compatible choices between receiving a gift card and reporting a seller from our list to the Department of Justice. We randomize subjects across two goods: face masks and hand sanitizer. We also randomize whether subjects have the option to report a seller who charges a low ($7.50 - $10) or a high price ($27.50 - $30). Both price ranges represent increases from pre-crisis levels (12-70% and 310-400%, respectively).

We choose the seller at random from the pool of listed sellers and we do not give individuals the seller's information. Hence, reporting decisions reflect only a desire to prevent third party transactions or punish sellers and not other confounders such as the possibility of getting compensation from the seller or reducing own search costs in the future. We use the responses to estimate the subjects' Willingness to Pay to Report (WTPR) sellers, our measure of distaste towards price gouging.

4. For example, Ba (2018) studies the willingness to pay to report police malfeasance in Chicago. Our method offers an alternative to Ba's that does not depend on the existence of naturally occurring exogenous variation in the costs of reporting. We thank Julio Elías for pointing out the similarity with consumer reviews.

5. We sampled participants through Prime Panels, an aggregate of online panels (Chandler et al., 2019). The survey is nationally representative on first moments, but not other moments.

Next, a complementary donation experiment with the same subjects helps tease out some of the mechanisms underlying the distaste toward price gouging. In our model, one key component of this distaste is the external payoff that consumers get from third party transactions. In turn, there are two main reasons why individuals obtain external payoffs. First, they may want to increase the access of others to the product. Second, individuals might have a distaste for firm profits or a disutility from pricing deviations from marginal cost (e.g., due to social norms). To tease out between the two, we ask subjects to choose between a $5 gift card and having us donate PPE we purchase from a seller to a hospital. As before, we randomize whether we buy from a high or low-price seller. Since we hold the quantity of PPE donated fixed, donation rates that decrease with higher ask-prices are consistent with a distaste for firm profits.

We provide four sets of results. First, individuals take costly actions to enforce price-ceilings. Eighty-percent of subjects choose to forgo money to report sellers in the lower-price range. On average, the willingness to pay to report sellers who charge the lower-price range was $4.78.[6] At the same time, seventeen percent of individuals are willing to pay to prevent us from reporting sellers. This polarization is similar to the one of Elías et al. (2019) who also find that some individuals strongly oppose paid kidney transactions while others are in favor of them.

Second, the WTPR is increasing in the price that the seller charges, as indicated by our theoretical framework. A one-percent increase in the ask-price increases the WTPR by 0.17%. This increase reflects a shift in the whole willingness to pay distribution, substantially reducing the polarization of our results. The reporting behavior is similar for both hand sanitizer and face masks.

Third, in contrast to the reporting behavior, the underlying mechanism driving the behavior is good-specific. Donation rates decrease by 24% when we buy the PPE from higher-

6. Consider that the price per survey response in these survey companies is around $1.25 for a 10-minute survey.

priced sellers, but only for hand sanitizers; face mask donations are unaffected by the seller's price. This finding suggests that there is a distaste for profits or deviations from markup norms in the case of hand sanitizer transactions, but not masks transactions.

Finally, half of the subjects who are willing to pay to report sellers are also willing to forgo the $5 gift card to have us donate PPE from a price gouging seller. This result suggests that individuals simultaneously internalize the desire to complete transactions and prevent them from occurring. They are against the transaction when other consumers pay for it but in favor when it is the experimenters who pay for it on behalf of a hospital. Hence, one cannot simply partition the population into those who want to transact and those who find the transaction undesirable.[7]

Our experiment captures a natural setting. Using observational data from actual price gouging consumer reports filed with different attorneys general, we document that complaints were on the rise during our study period and that the products we chose were prevalent in these complaints. The complaints contain wording that is associated with a distaste for transactions at illegally high prices, such as "take advantage of people" and wording associated with punishment, such as "hold accountable." Moreover, our results are robust to experimenter demand concerns and other confounders such as quality and attention differences.

This paper contributes to three strands of literature. First, we contribute to the literature on price gouging and price-control regulations. Cavallo et al. (2014) document lower product availability but sticky prices following natural disasters, consistent with a model of "consumer anger" against price increases. In the context of COVID-19, Cabral and Xu (2020) argue that seller reputation might explain why larger and older sellers engage less in price gouging. Chakraborti and Roberts (2020a,b) document increased consumer search following anti-price

7. We thank Al Roth for pointing out this insight.

gouging regulations.[8] Dworczak et al. (2019) develop a model in which the planner does not observe individuals' rates of substitution between a good and money, and find that when there is high dispersion in values for money it might be optimal to impose price controls even if it induces rationing.

Our results suggest that price gouging generates an externality, which a market designer might want to include in welfare calculations of anti-price gouging regulations (Rotemberg, 2008). For example, it might be possible to implement the same allocations in imperfect competition with price controls and subsidies. However, price ceilings might have higher welfare if there is distaste for firm profits since subsidies increase them. Moreover, our results suggest that the underlying mechanism differs depending on the type of product, so a one-size-fits-all policy might not be appropriate in response to emergencies.

Second, we contribute to the literature on fairness and third-party punishment by adding field context to the subject's decisions. A large experimental literature shows that third-parties frequently impose punishments for unfair economic behavior in the laboratory (Fehr and Fischbacher, 2004; Henrich et al., 2006). Some more recent natural field experiments show that altruistic punishment is rare, does not increase with the severity of the violation (Balafoutas and Nikiforakis, 2012; Balafoutas et al., 2016).[9]

Our finding that the mechanisms driving reporting differ by good suggests that not all third-party punishment is driven by altruism. Indeed, our model suggests that a distaste for firm profits drives punishment for price gouging hand sanitizer. This finding supports self-report evidence that the decision to report in third-party punishment games is due to

8. Beatty et al. (2020) provide similar evidence.

9. In contrast with Balafoutas and Nikiforakis (2012), we find that the vast majority of subjects are willing to punish others. Balafoutas et al. (2016) suggest that their finding in Balafoutas and Nikiforakis (2012) may be due to concerns about counter-punishment. Consistent with this claim, we find that the propensity to punish decreases the cost of punishment. Moreover, our finding that the WTP to report is increasing in the seller's price is consistent with the punishment decision depending on the severity of the harm inflicted by the seller (Carlsmith et al., 2002; Cushman, 2008; Ginther et al., 2016) and suggests that the tasks used in Balafoutas and Nikiforakis (2012); Balafoutas et al. (2016) may have suffered from a flat payoff problem (see Harrison (1992)).

anger in response to violations of social norms in some cases (Fehr and Fischbacher, 2004; Fehr et al., 2002; Fehr and Gächter, 2002; Herrmann et al., 2008). Finally, we contribute to the literature on repugnance by obtaining a revealed preference proxy measure of the distaste toward third-party transactions. Identifying repugnance requires a setting where individuals willingly engage in illicit transactions and third-parties can restrict the choice set of the potential transactors. This is challenging outside of the laboratory since many repugnant transactions are prohibited by law. For this reason previous studies primarily use hypothetical vignettes to study repugnance (see Ambuehl et al. (2015) and Elías et al. (2019)).[10] A notable exception is Stüber (2021) who uses a multiple price list to document that subjects permit low payments for bone marrow or stem cells, but pay to prevent the use of high monetary incentives. We introduce a revealed-preference method of measuring the taste for others transactions, which can be used in other settings that rely on reports for enforcement. Additionally, the only other paper that we are aware of that formally defines and models a taste for others transactions is Ambuehl et al. (2015). This model, however, relies on an observer misjudging the welfare of a third-party transaction. In contrast, our model does not rely on consumer misperceptions to generate repugnance.

Relatedly, Kahneman et al. (1986) argue that community standards of fairness restrict profits attainable by firms; consumers judge firm prices relative to reference levels. Rotemberg (2005, 2011) develops models of consumer anger and firm altruism, where consumers want their sellers to feel altruism toward them. Individuals also judge firms with respect to a reference level. Anderson and Simester (2010) provide experimental evidence of consumer anger along these lines. In our model, the reference level of repugnance is endogenous and depends on the market's distribution of prices.

The remainder of our paper proceeds as follows. Section 2.2 describes the setting and

10. Clemens (2018) uses exogenous variation in migration of guest workers, a job commonly regarded as repugnant, and analyzes the impact of migration of different outcomes (e.g., debt) as loose conditions to test for repugnance. Sullivan (2021) elicits revealed preferences for organ transplants for cats and relates these incentives to hypothetical preferences for human organ donations.

institutional context. Section 2.3 introduces our theoretical framework. Section 2.4 describes the subjects and experimental design. Section 2.5 describes the empirical results and Section 2.6 argues for their external validity. Section 2.7 concludes.

2.2 Setting

2.2.1 Observational Data Sources

In addition to the data generated by our experiment (which we describe in Section 2.4), we use data from two other sources. First, we use the Rainforest API to obtain information about search results and individual product characteristics from surgical face masks and hand sanitizer listings on Amazon. Each search reviews roughly 10,000 results for face masks and 1,800 for hand sanitizers. We combine an image recognition machine-learning algorithm and text analysis to filter unrelated products from the search results and to convert prices from different presentations to common units (12 fl oz. for sanitizer, 50 pack for masks). According to our algorithm (see Appendix B.2 for more details), only 6.3% of face mask search results were surgical face masks and 52% of sanitizer search results were hand sanitizer products.[11] Our algorithm, while precise, introduces measurement error relative to selecting products by hand, so the prices that we obtain should be taken with caution.[12]

We also use a database of actual price gouging complaints that consumers filed with Attorneys General from 6 different states, which we obtained with Freedom-Of-Information-Act (FOIA) requests.[13] Most states required individuals to fill a form that had at least two

11. Many results in the face mask category were cloth masks, which we distinguish from surgical masks since the medical community has pointed out differences in their effectiveness (MacIntyre et al., 2015). Many results in the hand sanitizer search were e.g., soaps.

12. Our product classification algorithm has an accuracy of over 0.95. We rely on a large-scale algorithm since we needed to detect sellers that are not easily detectable by manual search (e.g., Cabral and Xu (2020) use a sample of 14-17 hand sanitizers and masks) and we needed results in real time since many products were quickly removed by Amazon and new versions were continuously appearing.

13. Utah, South Carolina, Wisconsin, Idaho, Missouri and Illinois. We filed FOIA requests with every state and the DOJ, but we only received information from these states.

sections. In the description of the complaint, individuals included information about the seller, product and price. There was also a section that asked individuals for their suggested remedy (e.g., whether they wanted compensation, refund or something else). We machine-read and parsed the text from these two sections and obtained close to 1,900 observations.

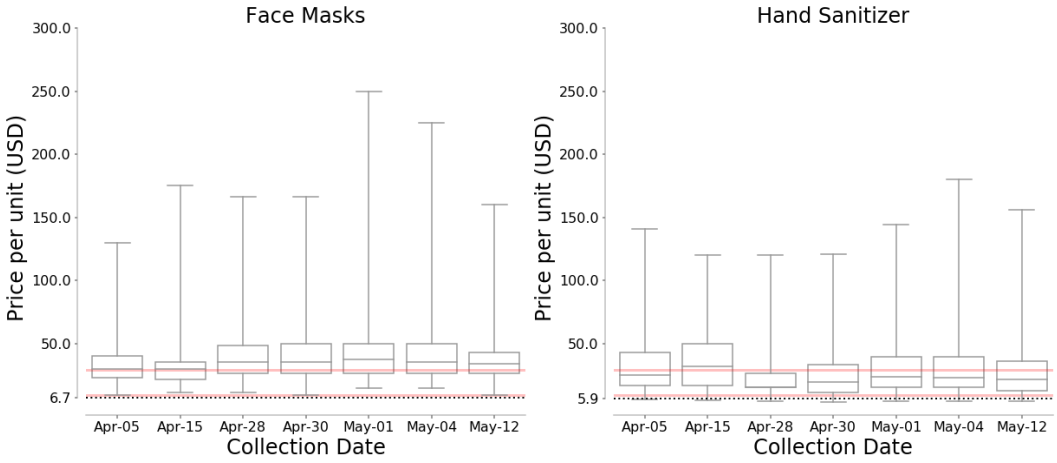
2.2.2 Context

The experiment occurred on April 30th and May 1st, three months after the first confirmed COVID-19 case in the United States (Holshue et al., 2020). At this time, the demand for PPE outpaced production capacity. Ninety-percent of U.S. mayors reported PPE shortages and one-third of medical facilities urged donations of personal masks to make up for the insufficient supply (Kamerow, 2020).[14] The sharp increase in demand led to dramatic price increases. Cabral and Xu (2020) document that, between January and March 2020, mask and sanitizer prices were equal to 2.72 and 1.8 times the 2019 prices, respectively. Within our sample, we observe an average price ratio of 6 for face masks and 5.3 for hand sanitizers, as compared to December 2019 prices that we obtained from camelcamelcamel.com. These pre-crisis prices correspond to December prices of 5 sanitizers and 2 face masks that we collected by hand (see Table 2.1).[15] Figure 2.1a shows that the price distribution remained stable throughout our sample period, before and after our experiment, and exhibits large dispersion.

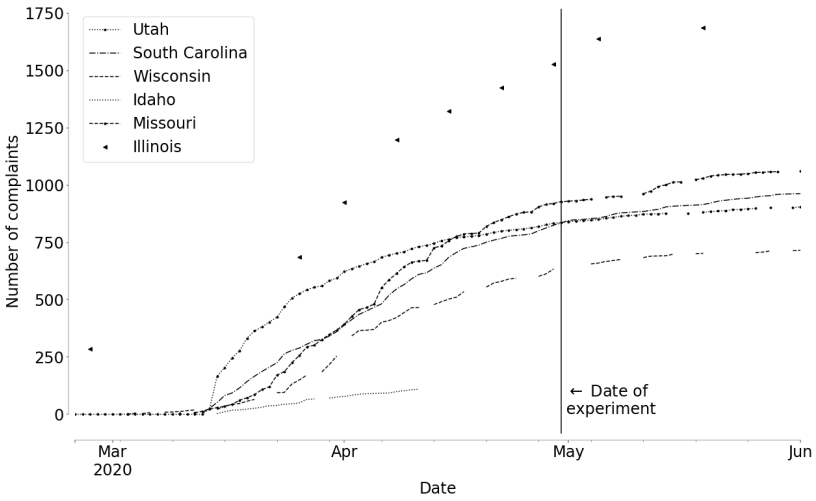
In response to these price increases, Amazon removed over half a million items with excessive prices (Amazon, 2020). State-level emergency declarations triggered price controls

14. While the CDC initially discouraged mask-wearing due to concerns about shortages, they reversed this guidance on April 3rd, 2020 (Dwyer and Aubrey, 2020). The change in guidance was about a month before our experiment began.

15. The difference between our price ratios and those in Cabral and Xu (2020) could be due to the different sample periods covered; they cover dates between January 15th and March 15th, while we cover April and May. Anecdotally, there was a substantial increase in demand between those dates. Moreover, our sample does not include historical price data; the API only provides real-time data. Obtaining historical data of products that were not taken down at this time was extremely hard, which is why we had a very small sample of pre-crisis prices.



(a) Amazon Price Distributions on Dates Close to the Experiment



(b) Cumulative price gouging complaints to state Attorneys General

Figure 2.1: Observational price gouging and complaint data

Notes: Panel (a) displays the price distributions on Amazon on dates around our experiment. Boxes contain quartiles of the distributions and the whiskers represent the 1st and 99th percentiles. The pink lines correspond to the price range in our experiment and the dashed lines correspond to the December prices. Panel (b) displays the cumulative sum of complaints for all types of goods from the beginning of the pandemic to June 2020 for states that responded to FOIA requests.

on goods "necessary for survival" in thirty-four of these states (see the maps in Figures B.1-B.3 in the Appendix for more information). These laws prevented either any increases above pre-crisis prices, 10-20% price increases or unconscionable price increases. Although

Table 2.1: Personal protective equipment prices in April and May

	Product	N	Price Ratio	Price	Lowest Price	Highest Price
April	Face Masks	1,862	5.63	37.74	5.40	349.25
			(4.80)	(32.13)		
	Hand Sanitizer	2,251	5.33	31.46	3.49	210.00
			(4.52)	(26.68)		
May	Face Masks	1,122	6.35	42.56	5.99	349.50
			(5.41)	(36.24)		
	Hand Sanitizer	986	5.32	31.38	3.49	220.15
			(5.03)	(29.69)		

Notes: Table displays summary statistics for the prices of PPE sold on Amazon between April 5th and May 12th. The prices are normalized to the units of the goods considered in the experiment. The price ratio column displays the average price of the PPE relative to the December price, which was calculated using the data of 4 products obtained from the price-tracking website camelcamelcamel.com. This is $6.70 for face masks and $5.90 for hand sanitizer. Standard deviations appear below the means in parentheses. Data scraped from Amazon on April 5th, April 15th, April 28th, April 30th, May 1st, May 4th, and May 12th 2020.

there is no federal law against price gouging, Executive Order 13910 issued on March 23rd prohibited the resale of PPE "at prices in excess of prevailing market prices."

Following the Executive Order, the Department of Justice (DOJ) announced a task force to combat hoarding and price gouging of different products, including sanitizing products and PPE. Individuals could report price gouging practices to their attorney general or to the Department of Justice's National Center for Disaster Fraud (NCDF).[16] The NCDF requests complainants identify themselves along with the accused, and provide as much information as possible about the transactions. At this point, the complaint is filed and investigated. Individuals found guilty of price gouging face steep fines and up to ten years in prison.

While there is no information about the total number of price gouging complaints received by the DOJ, the states in our sample of complaints had received roughly 1,000 complaints each by the time of our experiment, and they continued to rise afterward. Figure 2.1b plots the evolution of complaints filed in 6 different states. 13% of complaints in our sample include the word "mask" and 10% of them include the word "sanitizer". We summarize the

16. See `https://www.justice.gov/disaster-fraud/webform/ncdf-disaster-complaint-form`.

text in our sample of complaints using an unsupervised machine-learning algorithm (latent Dirichlet allocation, LDA) that detects topics automatically from a document.[17] On Table 2.2 we can see that complaint descriptions mostly concern products (e.g., eggs, meat, PPE and toilet paper). On the other hand, consumers refer to "lowering prices", "take advantage of people", "hold accountable" and "fair prices" in the section of the forms that asks about their suggested solution to the complaint.[18] For example, a (selected) complaint filed with the Idaho AG explains that a fair resolution for the complaint is: "I think they should be fined. I don't want a refund. I want justice."

Table 2.2: Topics from latent Dirichlet allocation model

Topic	Prevalence	Top terms
Description		
1	41.4%	egg, dozen, lb, pound, meat, beef, grocery, grind beef, hamburger, dozen egg
2	31.3%	mask, sanitizer, hand, hand sanitizer, bottle, amazon, wipe, lysol, oz bottle, seller
3	27.3%	paper, toilet, toilet paper, gas, station, gas station, towel, charmin, paper towel, gas price
Solution		
1	36.0%	normal, low price, paper, price normal, toilet, toilet paper, difference, desist, bring, cease
2	33.8%	company, gas, raise, complaint, raise price, fix, gas price, report, seller, control
3	30.2%	advantage people, community, accountable, food, hold accountable, fair price, grocery, check, change, issue

Notes: The table includes topics from price gouging reports filed to the AGs of Idaho, Illinois, Missouri, and Wisconsin. There are 1890 complaints in our sample (68 from ID, 102 from IL, 1271 from MO, and 449 from WI). "Description" is the field where consumers detail the reason why they are submitting the complaint. "Solution" is the field where consumers express any relief/solution that they are requesting. We only have solutions for 488 complaints. Missouri did not include a field to detail the requested solution. We exclude from the analysis common English stop words and lemmatize the words using the Hunspell dictionary. Top terms are calculated by sorting words according to the $\Pr(topic|word)$. We decided on three topics for parsimony.

17. See Gentzkow et al. (2019) for an overview of LDA topic models and some applications to economics. See Table B.2 in the Appendix for unigrams and bigrams used in complaints.

18. Tables 2.2 and B.2 show qualitatively similar wording to the one used in surveys about individual attitudes toward economic policies, see Stantcheva (2020).

2.3 Theoretical Framework

We present a simple model to motivate our experimental design and to argue that price gouging complaints contain information about repugnance and individuals' desire to prevent or punish illegal or unfair transactions. The model is a game in which $M > 2$ producers with constant marginal cost normalized to zero attempt to sell a product after a crisis to one of two consumers.[19]

The timing of the game is as follows. First, producers choose prices simultaneously. Second, one consumer is randomly selected to meet with a random seller. The consumer receives a price offer and decides whether to buy one unit of the product and whether to report the seller for price gouging to the authorities. The authorities charge reported sellers a fine $\kappa \geq 0$ and remove them from the pool of available sellers before they can match with the second consumer. Reports do not result in refunds to match our experimental design. The second consumer then matches with a random available seller and also decides whether to buy and whether to report. At the moment of making their choices, consumers ignore if they are the first or the second ones to meet with a seller.

Consumers experience utility u from buying the product and a cost c_r from reporting. Both of these parameters are privately observed and drawn from a distribution of types G. Consumers also get a common external payoff $e(p) \geq 0$ if the other consumer buys the product at price p. The function $e(p)$ is decreasing in price to capture a distaste for others transacting at high prices.[20]

As we show in Appendix B.1, the average willingness to pay to report price offer p is

19. Assuming a constant marginal cost is not essential, but it simplifies the exposition. Indeed, in the short run marginal costs might be steep. Perhaps this analysis is not valid for the extremely short-run with perfectly inelastic supply, but, as Figure 2.1b shows, price gouging complaints occur during extended periods of time. In this time frame, it is natural to assume that supply can be adjusted.

20. An example of such a function is $u^e - m(p)$, where u^e is a payoff that a user gets when the other consumes the product and m is a distaste for markups as in the model of pricing under fairness concerns of Eyster et al. (2021). Adding a distaste for markups in consumers' own transactions—in which case their utility from buying the product is $u - m(p)$—leaves the results unchanged.

equal to the expected change in the external payoff net of reporting costs:

$$WTPR(p) = \underbrace{\frac{1}{2M}\left[\mathbb{E}(e(p)q(p)) - e(p)q(p)\right]}_{\text{Expected change in external payoff}} - \overline{c_r}, \tag{2.1}$$

where $\mathbb{E}$ denotes expectation with respect to the equilibrium price distribution, $\overline{c_r}$ denotes average reporting costs and $q(p) \equiv 1 - G_u(p)$ is the individual expected demand. The expected change in the external payoff as a result of reporting equals the probability that the report makes a difference $(1/(2M))$ times the change in the payoff from the other consumer transacting at random prices rather than at price p. This willingness to pay is decreasing in seller price: consumers prefer forcing the other to meet a randomly chosen seller rather than a confirmed expensive one. We can interpret $e(p)q(p)$ as the (ex-ante) repugnance of a transaction: it is the surplus that an individual gets from a third party transaction times the probability that it happens. Hence, the willingness to pay to report captures individuals' repugnance as well as their desire to prevent or punish illegal or unfair transactions, which we capture in this model in the reporting costs.[21]

In Appendix B.1, we also characterize the equilibrium price distribution. When reporting is not available, all sellers charge the monopoly price p^m. With reporting, all sellers charge a price p^*, which is lower than the monopoly price and decreasing in the fine κ. Hence, authorities can keep prices down by allowing consumers to report.

Finally, in Appendix B.1 we also argue that disentangling the mechanisms underlying the external payoffs $e(p)$, one of the determinants of the WTPR, is essential to understand the welfare implications of different policies. For instance, if consumers exhibit a distaste for firm profits or markups, subsidies might bring a lower surplus than price controls, even if they achieve the same quantity. The literature on consumer anger has obtained similar results; for instance, see Di Tella and Dubra (2014).

21. We assumed that reporting costs are not a function of prices, but we can relax this assumption to incorporate, for example, utility that individuals get from reporting illegal transactions at different prices.

2.4 The Experiment

A survey company, CloudResearch, recruited 1,418 participants from the United States for the experiment. The company selected these participants to match the U.S. census on race, Hispanic origin, age, and gender.[22] Panel (a) of Figure 2.2 illustrates the flow of the experiment, and an exact copy of the survey appears in Appendix Section B.4. The experiment begins with questions related to purchasing behavior. We then elicit the willingness to pay for PPE and ask subjects to report the lowest PPE price they consider excessive.

Table 2.3: U.S. adult sample description

	Full Sample	Treated Sample	US Pop
Female	52.95	52.91	51.00
Age 18-34	27.94	27.82	32.10
Age 35-54	36.07	36.59	31.30
Age 55+	36.00	35.59	36.60
White (non-Hispanic)	63.72	63.84	62.30
Black	12.09	12.15	12.96
Hispanic	16.61	16.53	16.41
Asian	5.77	5.61	5.96
Other race/ethnicity	2.36	2.44	2.37
Less than HS	2.08	1.94	10.60
HS/GED	15.36	15.31	28.32
Some college/Associate degree	31.97	31.70	27.77
Bachelor's Degree	30.79	30.77	21.28
Graduate Degree	19.81	20.27	12.04
Income < \$50,000	37.53	37.38	43.70
\$50,000 ≤ Income < \$100,000	37.46	37.10	30.00
\$100,000 ≤ Income	25.02	25.52	26.20
Sample Size	1,439	1,391	

Notes: The table describes the demographic characteristics of the respondent sample and compares them to the Vintage 2019 national population estimates from the Census Bureau `https://www.census.gov/data/tables/time-series/demo/popest/2010s-national-detail.html`. The survey company selected participants to match the U.S. census on race, Hispanic origin, age, and gender. The sample over-represents high-education and median income subpopulations based on self-reported information.

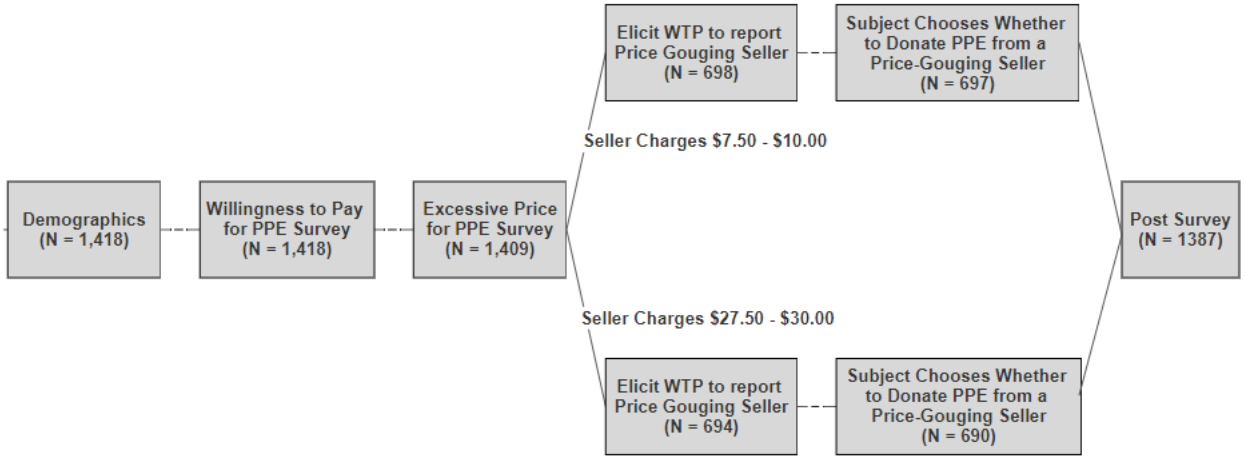
22. We pre-registered 1,200 observations. CloudResearch automatically added 218 observations to match the target characteristics we requested prior to the experiment. The characteristics of our subjects is shown in Table 2.3. Treatment balance is shown in Table 2.4. There is some imbalance in education, but controlling for education dummies does not change the coefficients in our regression models, suggesting that this chance imbalance did not affect our results.

Table 2.4: Treatment balance

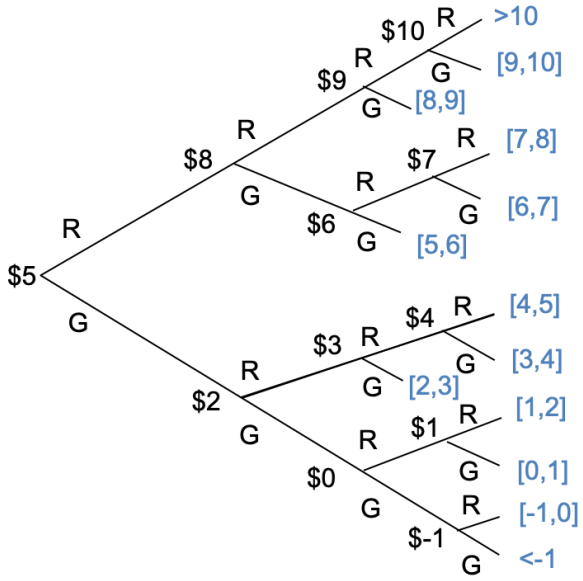
	Hand Sanitizer		Face masks		
	\$7.5-\$10	\$27.5-\$30	\$7.5-\$10	\$27.5-\$30	F Test p-value
Age	46.15	45.48	47.32	47.44	
	(17.02)	(16.6)	(17.5)	(16.59)	0.35
Female	0.52	0.55	0.51	0.54	
	(0.5)	(0.5)	(0.5)	(0.5)	0.69
White	0.63	0.66	0.64	0.63	
	(0.48)	(0.47)	(0.48)	(0.48)	0.81
Black	0.12	0.11	0.13	0.12	
	(0.33)	(0.32)	(0.33)	(0.33)	0.95
Hispanic	0.19	0.16	0.16	0.16	
	(0.39)	(0.37)	(0.36)	(0.37)	0.70
Asian	0.05	0.03	0.05	0.07	
	(0.21)	(0.18)	(0.22)	(0.25)	0.21
Other race/ethnicity	0.04	0.05	0.04	0.03	
	(0.19)	(0.22)	(0.2)	(0.18)	0.70
Less than high school	0.02	0.02	0.02	0.01	
	(0.15)	(0.13)	(0.15)	(0.12)	0.79
High school or GED	0.13	0.2	0.18	0.11	
	(0.34)	(0.4)	(0.38)	(0.31)	0.00
Some college/associate degree	0.32	0.34	0.27	0.34	
	(0.47)	(0.48)	(0.44)	(0.47)	0.11
Bachelor's degree	0.34	0.27	0.3	0.33	
	(0.47)	(0.44)	(0.46)	(0.47)	0.16
Graduate degree	0.19	0.18	0.23	0.21	
	(0.39)	(0.38)	(0.42)	(0.41)	0.26
Income $<$ \$50,000	0.38	0.38	0.35	0.38	
	(0.48)	(0.49)	(0.48)	(0.49)	0.83
\$50,000 $\leq$ Income $<$\$100,000	0.46	0.48	0.49	0.45	
	(0.5)	(0.5)	(0.5)	(0.5)	0.69
\$100,000 $\leq$ Income	0.26	0.24	0.25	0.27	
	(0.44)	(0.43)	(0.43)	(0.44)	0.82
Sample Size	349	346	348	348	

Notes: Table shows the mean and standard deviations of each variable separately by treatment with standard deviations in parentheses below each mean. The F Test P-value shows the p-value from an F test with the null hypothesis that the means of each variable are equal across all treatments.

After the surveys, we assigned subjects into treatments using a 2×2 completely randomized between-subjects design (treatment is balanced across almost all demographics except education, see Table 2.4). The treatments varied the type of PPE subjects would consider independently with a seller's ask-price. Half of the subjects considered a lower price range



(a) The structure and flow of the experiment



(b) Willingness to pay to report decision tree

Figure 2.2: Experimental design

Notes: Panel A reports the flow of the experiment. Randomization occurs after the excessive price survey. Within each of the seller price treatments, subjects are randomly split into considering either hand sanitizer or face masks. Numbers in parentheses represent sample sizes at that stage of the experiment. Panel B displays the decision tree subject's faced during the willingness to pay to report the experiment. All subjects began with the decision between a $5.00 gift card and reporting a seller. Subsequent decisions depend on the subject's choice.

($7.50 to $10) and a higher price range ($27.50 to $30). Both price-ranges constitute illegal price-increases under many price gouging regulations. Within each price-range we evenly split subjects into treatments that consider 12 FL oz / 355 ML hand sanitizer or 50 count

disposable face masks. We use two different types of PPE to investigate good-specific heterogeneity in the willingness to pay to report or the mechanisms. We revealed pre-crisis prices (December 2019) were \$5.90 for hand sanitizer and \$6.70 for face masks of equivalent presentations, to homogenize the points of reference. We also provided a picture of the goods to prevent subjects from confusing disposable face masks with the more expensive N95 face masks.[23] Following Kuziemko et al. (2015), we undertook several steps to ensure the sample's validity. First, we only allowed participants with U.S. IP addresses and launched our survey on a workday morning. Second, we included a CAPTCHA to exclude potential robots. Third, we told respondents that payment was contingent on survey completion. Finally, we included attention checks.

2.4.1 Willingness to Pay for Personal Protective Equipment

The survey told subjects about an algorithm we created to track PPE on Amazon. We offered to notify them if the delivery of a similar product was available in two weeks or less. If they wanted to be notified, they could select the maximum price that they were willing to pay for each of the products. 44% of subjects responded that they wanted to be notified. At the end of the survey, we provided subjects with a link to a randomly chosen product from our list at or below their maximum willingness to pay (see Figure B.8 in the Appendix B.4). Following our pre-registration plan, we winsorized the data at the 99th percentile. While this procedure is not incentive-compatible, it still gives some information about valuations and beliefs about the price distribution. Moreover, it gives a lower bound to the WTP for the product, since participants do not have incentives to bid above their WTP (by doing so they risk getting information that will not be useful for them).[24] Throughout the paper

23. On April 2nd, 2020, Amazon prohibited the sale of N95 face masks on their platform (Rey, 2020).

24. Unfortunately, only 1.2% and 6% of respondents who received a link to a listing of face masks and sanitizer in their price range clicked on it. While this could indicate low consequentiality of our WTP elicitation (few respondents actually willing to "track" products at the price range they report), it could also be due to respondents fearing to lose their progress in the survey (if the Amazon link opens in the

we refer to this quantity as willingness to pay for the PPE, with the caveat that it is a lower bound. This downward bias does not affect our results, since we use this variable to argue that subjects are willing to purchase the products at the price ranges considered in the experiment.

2.4.2 Excessive Prices for Personal Protective Equipment

To compare our incentivized measure of willingness to pay to report with hypothetically stated measures about the subject's distaste for the third-party transactions, we asked subjects to tell us the lowest price they considered to be excessive for both goods. Individuals use numerous adjectives to describe prices in the gouging context, e.g. abusive, unfair, exorbitant or excessive. While all these terms have some normative content and could trigger differentiated concepts in subjects' minds, we chose to use excessive as it is commonly used in laws (see, for instance, Giosa (2020)) and describes a situation in which the price is unexpectedly high without placing undue emphasis on potential ill intention of the seller. Hence, this exercise allows us to compare an objective measure such as the WTPR with the subjective language that is commonly used in regulation.

2.4.3 Eliciting willingness to pay to report

We elicited the subject's willingness to pay to have us report a randomly chosen seller for price gouging to the Department of Justice using an iterative multiple price list (iMPL). The procedure confronts subjects with an array of paired options and asks them to make a single choice within each pair. At each step, the program asks subjects which of the following two options they prefer:

1. We **report** an Amazon seller to the **Department of Justice National Center for**

same Window). There could be some unobserved fraction of respondents who simply copied the link without clicking on it.

Disaster Fraud. This Department is in charge of preventing price gouging for critical supplies. We will report one seller in our list who charges between [$7.50 - $10.00,$27.50 - $30.00] for one [12 FL oz. / 355 ML hand sanitizer, 50 count disposable face masks].

2. You receive a $[Value] Amazon Gift card.

All respondents first decide between reporting a seller to the DoJ and a $5 Amazon gift card. If the subject chooses to report, her next decision is between an $8 gift card and reporting the seller. If instead, she selects the money, her next decision is between a $2 gift card and reporting the seller. We continue increasing or decreasing the gift-card amount; Panel (b) of Figure 2.2 displays the iMPL's decision tree. When the differences in values between the last choice and refined choice dropped below $1, the program stopped. We randomly select one in every 10 subjects and randomly implement one of the subjects' decisions (including both reporting and donating).[25]

Variation in the gift card amount maps into variation in c_r through the lens of our model, and it allows us to measure each subject's willingness to pay to report (WTPR). The WTPR can fall into one of thirteen intervals: $(-\infty, -1]$, $(-1, 0]$, $(0, 1]$, $(1, 2]$, ..., $(9, 10]$, and $(10, \infty)$.[26] Following our pre-registration, we either present the portion of subjects falling within a WTPR interval or set the WTPR value to be the maximum of the interval, 11 in the case of the $(10, \infty)$ interval. We also follow the triangular distribution approach from Allcott and Kessler (2019) as a robustness check.

25. The iMPL imposes strict monotonicity and enforces transitivity (Gonzalez and Wu, 1999). The method's main advantages are transparency to subjects and avoiding framing effects. However, it provides interval responses rather than an exact WTPR. We elected not to use a method providing exact WTPR's due to concerns of a flat payoff problem (Harrison, 1992). Out of the 1,200 subjects originally planned for the experiment, we selected 120 to implement their choices. From these, 57 were to receive a gift card, 60 to report a seller and 3 to donate a product. The reporting numbers were: 19 and 17 lower and higher-priced hand sanitizers, 14 and 10 lower and higher-priced face masks, respectively. There were 2 hand sanitizer donations in the high-priced range and 1 in the lower-priced range. Donations were sent to the Knapp Center For Biomedical Discovery KCBD , located at 900 E 57TH ST Chicago, IL 60637-1428.

26. To administer the negative WTR, we offered subjects the choice between reporting and earning $1 or not reporting and earning $0.

To ensure consequentiality, we chose goods subject to price gouging legislation. Furthermore, we informed our subjects that our algorithm detected sellers who charged prices between five and fifty dollars in the months before the experiment, so both treatments had the same support. Whenever our algorithm identified sellers charging in a price range at which a subject chose to report, we reported the seller to the NCDF. Thus, report decisions exposed sellers to the threat of steep fines or incarceration.

We do not give participants any information about the seller other than the price. By doing this, we restrict the possibility that they might obtain some direct benefit of reporting, such as reducing their own search costs in the future or obtaining a refund from the seller (as many consumers in our sample of complaints look for).[27] Additionally, this prevents participants from reporting the seller by themselves and still get the gift card, especially since we are saving them the costs of filling out the report form.[28]

2.4.4 Donation Experiment

After the reporting experiment, subjects decided between a $5 Amazon gift card and Donating PPE to a hospital listed in `getusppe.org`, an organization that allocates PPE donations to health care workers. Moreover, we tell subjects that we purchase the PPE from a randomly chosen seller at the price range. The item considered in this step of the experiment matches the iMPL in the type and seller price range. Our treatment thus keeps constant the quantity of PPE donated and varies only the price at which we buy the product. If individuals have a distaste for profits/deviations from markup norms, the fraction of individuals donating with the low price should be higher than with the high price. We presented the

27. Since there are thousands of search results, the possibility of reducing their own search cost by reporting a random seller is insignificant. However, many other consumers might still match with that seller, so they can still reduce others' search costs, as in our model.

28. Participants could still search for a seller by themselves and report it, but this is true across our treatments and gift-card amounts. Moreover, since there are thousands of noisy search results (see Section 2.2.1), searching is costly

question to respondents as:

1. We **buy** from a seller and **donate** to a site listed in `getusppe.org`. This organization coordinates donation of Personal Protective Equipment to health care workers. We will buy one [12 FL oz. / 355 ML hand sanitizer, 50 count disposable face masks] from a seller in our list who charges between [$7.50 - $10.00,$27.50 - $30.00].

2. You will receive a **$5** Amazon gift card (code to redeem it at the end of this survey).

We ensured consequentiality by verifying that `getusppe.org` had a demand for both types of PPE. Whenever a subject in our sample was randomly selected to have their donation decision implemented, we purchased the items and donated them to a hospital listed in `getusppe.org`.

In the final part of the experiment, we asked subjects questions that checked their comprehension of the experiment and their beliefs about quality differences between differently priced goods.

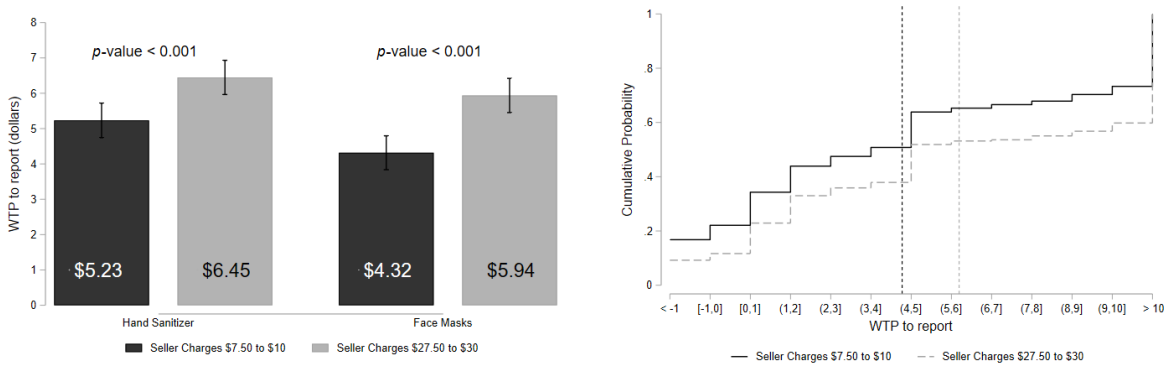
2.5 Results

2.5.1 Willingness to Pay to Report

Our first goal was to test whether consumers take costly actions to oppose price gouging. We find that 78% of them are willing to forgo compensation to report sellers who charge in the low-price range.[29] On average, respondents forgo over $4.8 to report sellers. Moss et al. (2020) find that most respondents report median earnings of $6-$9 per hour on Cloud Research. Our finding translates to subjects giving up 53-80% of their hourly wage.

29. Note that individuals do not need to view themselves as the only person who can report the seller, as long as they believe that their reports marginally increase the probability that a seller is punished.

Consistent with our model, the WTPR is increasing in the ask-price.[30] Figure 2.3a and Table 2.5 show that increasing the price range from \$7.50-10.00 to \$27.50-30.00 increases the WTPR by \$1.22 and \$1.60 for hand sanitizer and mask, respectively. The economic significance of the treatment effect is substantial as it amounts to slightly over 20% of the pre-pandemic prices of both categories and implies an elasticity of WTPR to the ask-price of 0.17.[31] The effect size is 0.3 standard deviations. Moreover, these results are robust (and even stronger) when we follow the procedure in Allcott and Kessler (2019) to assign WTPR values from the multiple price list (see Figure B.7 and Table B.4 in the Appendix).



(a) Average willingness to pay to report by seller price

(b) Distributions of willingness to pay to report by seller price

Figure 2.3: Willingness to Pay to Report

Notes: Panel (a) displays the average willingness to report sellers for price gouging at different prices separately by PPE type with 95% confidence intervals. Panel (b) presents the cumulative distribution function of willingness-to-report price gouging of either good by seller price. The vertical lines represent the average WTPR at each seller price. Kolmogorov-Smirnov p-value of 0.00003 for face masks and 0.0009 for hand sanitizer. p-values of 0.8224, 0.9989 for face masks and 0.8521, 0.9986 for hand sanitizer, for the H0 that the distribution of WTPR under high prices first and second-order stochastically dominates the distribution with low prices, using the Bootstrap tests from Abadie (2002).

The average effect underlies a more dramatic shift in the WTPR distribution. Figure 2.3b shows an increase of subjects willing to forgo the maximum potential gift card and

30. As a robustness check, Appendix Table B.3 displays the treatment effects on the probability of selecting to report in the first decision. This result can be seen as purely a between-subject estimate of the treatment effect. Consistent with our within-subject WTP measure, we find that subjects are more likely to choose to report in the first decision when the seller's price is higher.

31. Elasticity estimate calculated using the midpoint of the seller's price range.

	(1) WTPR	(2) WTPR	(3) WTPR	(4) WTPR
Seller Charges 27.50 to 30	1.42	1.42	1.40	1.23
	(0.25)	(0.25)	(0.25)	(0.36)
Face Masks		-0.71	-0.75	-0.92
		(0.25)	(0.25)	(0.35)
Seller Charges 27.50 to 30 × Face Masks				0.34
				(0.50)
Constant	4.77	5.13	7.12	7.19
	(0.17)	(0.21)	(0.79)	(0.80)
Elasticity Estimate	0.17	0.17	0.17	0.15
Controls	NO	NO	YES	YES
R-Squared	0.023	0.029	0.047	0.047
Observations	1,391	1,391	1,391	1,391

Table 2.5: Willingness to pay to report

Notes: This table shows regressions of individual willingness to pay to report on treatment dummies. Heteroskedasticity robust standard errors in parentheses. Controls include race indicators, gender indicator, age, income, education, and whether they chose to track either product, has purchased either item on Amazon in the past and whether they have Amazon Prime. Elasticity estimate calculated using the midpoint of seller price range.

a substantial reduction in individuals expressing indifference or a desire to pay to prevent reporting. The distributions of WTPR for both prices are statistically different (Kolmogorov-Smirnov p-value < 0.001 for face masks and for hand sanitizer). Moreover, we cannot reject that the distribution of WTPR under the high prices first and second-order stochastically dominates the distribution under the low prices (p-values of 0.8224, 0.9989 for face masks and 0.8521, 0.9986 for hand sanitizer).[32]

Figure 2.4 also shows the distribution of WTPR to be polarized for the low-price range treatment arms; subjects have polarized preferences toward moderate price gouging. 17% of subjects are willing to forgo one dollar or more to avoid punishing these sellers. This negative willingness to pay to report is consistent with our theoretical framework; it could be driven either by deriving negative utility from punishing sellers or by considering the repugnance of a given price to be much lower than the market average. We found such respondents in

32. We use the Bootstrap tests from Abadie (2002) with 100,000 bootstrap samples.

both price ranges, but higher-priced sellers are substantially less likely to be protected by our subjects. The polarization of the distribution of the WTPR reflects a polarization that is similar to what Elías et al. (2019) find in the context of kidney donations; some people strongly opposing the transaction and some strongly in favor of it.

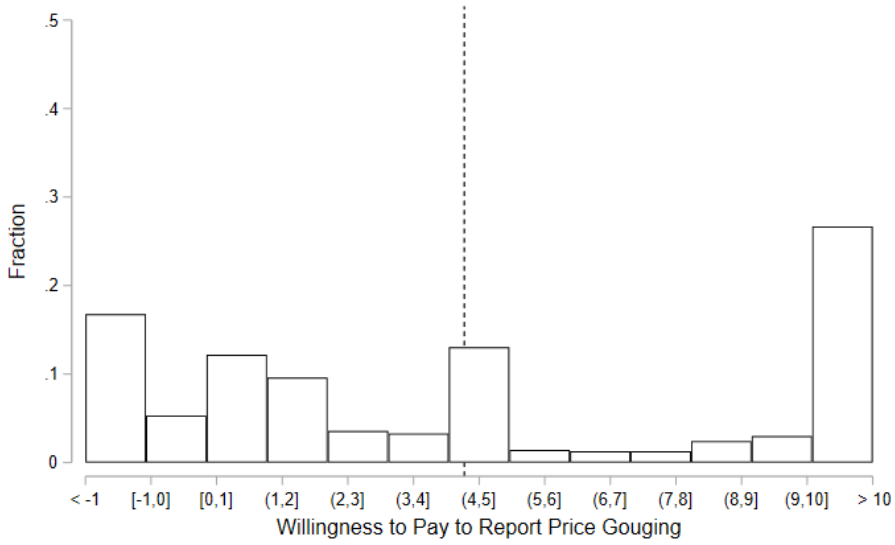


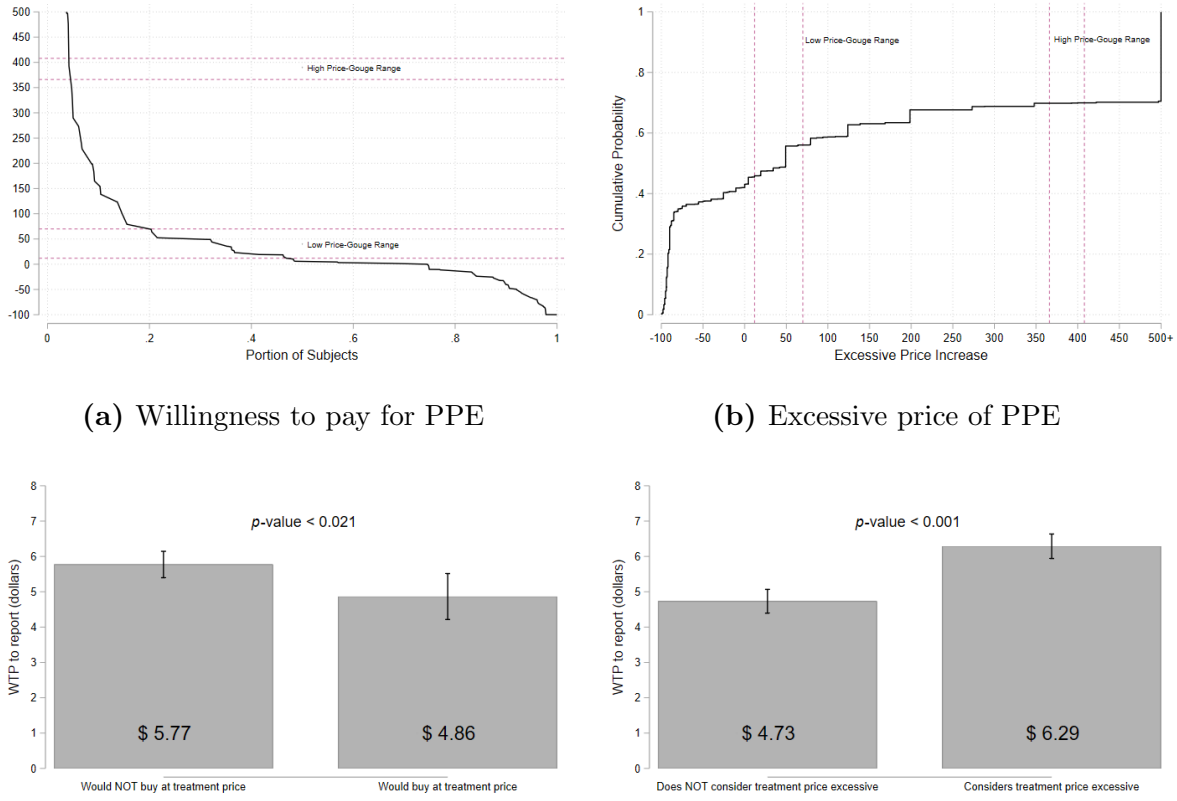
Figure 2.4: Histogram of willingness to pay to report at the low price

Notes: This table displays the distribution of willingness to pay to report for all subjects in the low price range treatments. The vertical line represents the average WTPR at the low-price range.

Since at least 50% of subjects are willing to purchase PPE at prices in the lower price range, the decision to punish sellers implies that subjects find these transactions repugnant (Roth, 2007).[33] That is, subjects prevent voluntary transactions between third-parties. Figure 2.5a shows the portion of subjects willing to pay for either type of PPE at different percent changes from the December price. Almost half of the subjects are willing to buy the goods from "low-price" sellers, while at least five percent are still willing to buy from

33. As we argued above, it is unlikely that individuals receive any direct benefit (other than moral benefit) from reporting sellers, since we match them with a random seller chosen from a large pool. This means that they cannot claim any refund or expect to face lower prices or search costs in the future because of this decision.

"high-price" sellers.[34]



(a) Willingness to pay for PPE

(b) Excessive price of PPE

(c) WTPR by willingness to Buy PPE

(d) WTPR by seller price

Figure 2.5: Heterogeneity in willingness to report by survey responses

Notes: Panel (a) shows the portion willing to pay for either type of PPE at percent changes from pre-crisis prices. Horizontal lines denote the treatment price ranges. Panel (b) displays the CDF of self-reported excessive prices for either type of PPE at percent changes from pre-crisis prices. Vertical lines denote the potential seller price ranges. Data in Panels (a) and (b) are winsorized at the 99th percentile. Panel (c) shows the average WTPR split by whether the subject reported a WTP exceeding the minimum seller price they consider and 95% confidence intervals. Panel (d) displays the average WTPR split by whether the subject reported that they found values in the seller's price range excessive and 95% confidence intervals. Estimates pool subjects across all treatments and exclude subjects who did not report a WTP or excessive price for the PPE considered in their treatment.

34. With the caveat, as we argued above, that our measure of WTP for the PPE is biased downwards.

2.5.2 Underlying Motives

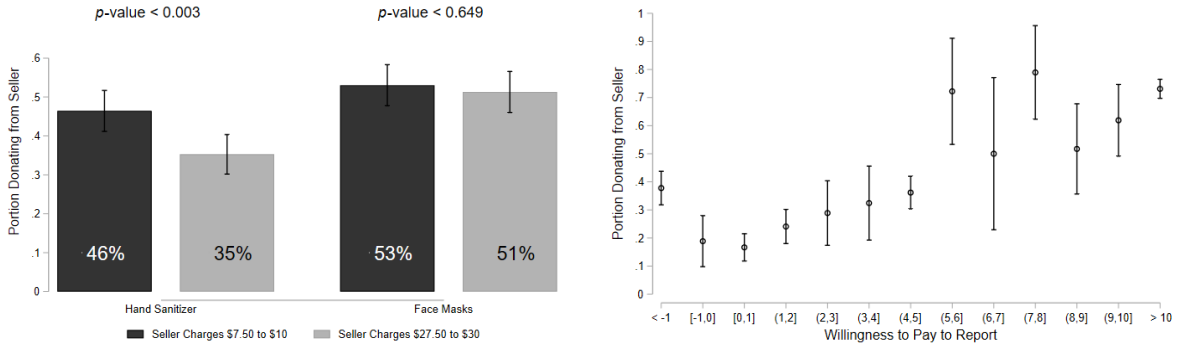
Over 43% of participants are willing to forgo the five dollars to have us donate the PPE (see Table 2.6). Since all subjects completed both tasks, we can use the within-person relationship between these choices to check for consistency between donation and reporting decisions. Nearly 50% of subjects who were willing to pay positive amounts to report sellers were also willing to donate. Figure 2.6b reports a generally positive association between WTPR and donations. The notable exception to this pattern is that subjects who are willing to pay to prevent us from reporting sellers have donation rates twice as large as those who express a WTPR of zero ($p < 0.001$). Their donation rate is less than the average of all subjects who are willing to pay to report price gouging, and comparable to subjects willing to pay \$2 to \$5 to report sellers.

Table 2.6: Propensity to donate

	(1) Donate	(2) Donate	(3) Donate	(4) Donate
Seller Charges 27.50 to 30	-0.06	-0.06	-0.06	-0.11
	(0.03)	(0.03)	(0.03)	(0.04)
Face Masks		0.11	0.11	0.06
		(0.03)	(0.03)	(0.04)
Seller Charges 27.50 to 30 × Face Masks				0.09
				(0.05)
Constant	0.50	0.44	0.40	0.42
	(0.02)	(0.02)	(0.09)	(0.09)
Elasticity Estimate	-0.08	-0.08	-0.07	-0.13
Controls	NO	NO	YES	YES
R-squared	0.004	0.017	0.033	0.035
Observations	1,386	1,386	1,386	1,386

Notes: This table displays the effect of treatments on the willingness to the propensity to donate. Omitted category is hand sanitizer sold for \$7.50 to \$10.00. Controls include race indicators, gender indicator, age, income, education, and whether they chose to track either product, has purchased either item on Amazon in the past and whether they have Amazon Prime. Heteroskedasticity robust standard errors in parentheses. Elasticity estimate calculated using the midpoint of seller price range.

Using exogenous variation in the seller price, we find that the mechanism driving the repugnance toward price gouging is good specific. The donation rates for subjects considering



(a) Propensity to donate by seller price and PPE

(b) Relationship between WTPR and propensity to donate

Figure 2.6: Propensity to donate PPE from price gougers

Notes: Panel (a) displays the average willingness to report sellers force price gouging at different prices separately by PPE type with 95% confidence intervals. Panel (b) plots the average portion of subjects choosing to donate PPE within every willingness to report bin. This figure pools both seller prices and types of PPE.

hand sanitizer decrease by 0.22 standard deviations (24%) when we purchase the good from a higher priced seller. Conversely, subjects who consider face masks are uninfluenced by seller price (see Figure 2.6a and Table 2.6).

Through the lens of our model, these results provide evidence of distaste for firm profits or markup deviations from norms with hand sanitizer, but not face masks. This result is striking, since the willingness to pay to report face masks was at least as responsive to seller price as the one of hand sanitizers (see Figure 2.3a and Table 2.5).[35]

There are many potential explanations for the difference in mechanisms between products. This could be driven by a difference in how individuals perceive masks to be of a higher necessity than hand sanitizer (see the discussion in Tobin (1970) of how basic necessities might trigger distributional concerns). It could also be the case that masks are perceived to have a higher spillover on other individuals—and so the "external" component of repugnance

35. As a robustness check, Appendix Table B.5 displays the treatment effects on the probability of donating based on the first decision made in the iMPL. Consistent with our within-subject WTP measure, we find that subjects are more likely to donate when they chose to report in the first decision.

is higher.[36]

This heterogeneity across products is also present in the observational data of price gouging complaints. We computed the sentiment scores of the text used in complaints, which is a text-analysis method of evaluating whether a given corpus contains positive or negative language.[37] As Figure B.4 shows, the description field of both mask and sanitizer complaints contains a similar sentiment: we cannot reject the hypothesis that the sentiment distributions are equal. This is reasonable, since complaint descriptions typically include factual information about the seller and the circumstances of the report (e.g., price, location, presentation). However, the language used in the suggested solution fields seems to be quite different: we reject equality of distributions between masks and sanitizers and we cannot reject first and second-order stochastic dominance. As we mentioned in Section 2.2, the suggested solution field of complaints tends to contain more normative views of what should be done to the seller. Our sentiment analysis shows that people use more negative language when suggesting solutions about mask complaints versus sanitizer complaints.[38] While this result does not provide any evidence for why we observe different mechanisms, it suggests that individuals have distinct attitudes in response to price gouging in each of these products.

36. Note that these masks were surgical masks and at their main contribution is preventing the spread to other people. Hence, the external component in this case might be high relative to hand sanitizers, since their main benefit is to prevent direct contagion. We thank an anonymous referee for the suggestion and the reference. Our results also support Pedersen et al. (2018)'s claim that third-party punishment arises because of the interdependency between the punisher's and victim's welfare.

37. See Gentzkow et al. (2019) and Algaba et al. (2020) for an overview of the use of sentiment analysis in Economics. We used the 'sentimentR' package; see Naldi (2019) for a description and comparison with other sentiment lexicons.

38. As pointed out by one of the referees, this might sound counter-intuitive. If there is distaste for firm profits in the case of sanitizers, shouldn't people express more resentment in their complaints? One explanation for why people use more negative language in the case of face masks is that they are angrier, not at the sellers for making a profit, but at the fact that people are not getting access to face masks. Indeed, WTPR is more responsive to seller price in the case of face masks, which is consistent with people having a stronger sentiment.

2.5.3 Heterogeneity

The heterogeneity in WTPR across individuals' willingness to buy the goods at the posted price or the perception of "excessiveness" is reported in panels (c) and (d) of Figure 2.5. The WTPR is higher when the ask-price exceeds individuals' willingness to pay as well as when the price range is considered to be excessive ex-ante. Figure 2.5b displays the CDF of self-reported excessive prices for either type of PPE at different % changes from pre-crisis prices. Only 40% of respondents would consider prices in the lower price range excessive while more than 70% deem prices in the higher price range excessive.

We find that a non-negligible portion of the sample finds pre-pandemic prices excessive—close to 40%. One possible driver of this finding is the social norm or preference of individuals for having firms engage in pro-social activities during emergencies (which may include lowering prices). Indeed, Marcelo et al. (2020) document many different examples of firms participating in non-profit or Corporate Social Responsibility (CSR) initiatives during crises.[39] This finding could also be due to the hypothetical and subjective nature of the question of what "excessive" means. The WTPR might provide a revealed-preference version of "excessive" and be more suitable for policy (e.g., to define which price increases are considered to be price gouging). Indeed, there is a close relationship between both measures, since individuals who consider a price range excessive have a 32% higher WTPR than those who don't.

In our study pre-registration, we also posited that the salience and prevalence of the emergency as measured by the number of reported deaths in the state as well as whether or not price gouging was locally forbidden could affect our results. We found no evidence that the number of deaths affected WTPR but the propensity to donate did increase. Regarding local legislation, tests for respondents in states without any anti-gouging laws lost statistical

39. He and Harris (2020) give examples of CSR in the context of Covid. Similarly, Uber offered free rides to passengers during natural disasters in the example above rather than offering rides at non-surge prices.

significance due to the reduced sample size but the results remain qualitatively unchanged. We report these results in the Appendix due to their small informational content.[40]

2.6 Robustness and Generalizability

Regarding internal validity, there are four potential confounders to our results. First, there might be experimenter demand effects that incentivize individuals to align their responses to what they perceive to be our desired results. To reduce this possibility, we provide full anonymity to our participants (de Quidt et al., 2019). We coded the survey to embed bonus payments to avoid asking for any identifying information.[41] Moreover, the heterogeneity observed in Panel (a) of Figure 2.6 suggests that any experimenter demand effect would need to be good specific, which is unlikely. Second, the treatment might be too subtle for individuals to notice. We asked individuals an attention question at the end of the survey, in which they had to report the price range that they were assigned.

Approximately 14.6% of participants did not remember correctly the price that they were given. This affects disproportionately individuals who were assigned to the high-price range. Table B.6 shows that 25% of individuals in the high price range tend to misremember the price range that they were given (vs 5% in the low-price range); that is, they report incorrectly that they were assigned the lower price range. This means that our results, if anything, are biased toward zero, since some people in the upper price range believe that they were assigned the lower prices. Indeed, Table B.7 displays the treatment effects for the willingness to pay to report and donations for the overall sample and only attentive subjects. When we restrict the analysis to attentive subjects, we find a higher WTPR and a more substantial reduction in donations when considering high-priced face masks. In other words, the lack of attention makes our results overly conservative, if anything.

40. The heterogeneity analysis mentioned in the pre-analysis plan appears in Appendix tables B.11, B.12,B.13, B.5, B.14, B.15, and B.16.

41. Many field experiments compensate participants by sending gift cards to their email address.

Third, individuals might perceive that products with higher prices differ in other ways as well from products with lower prices (e.g., differences in quality, shipping dates, etc.). We tell individuals that our algorithm has found products in the previous weeks with prices from \$5 to \$50 with similar shipping dates. At the end of the survey, we ask the subjects whether they agree with the statement that products in the upper price range have a higher quality than products in the lower price range. Roughly 20% of individuals agree with the quality statement. Table B.8 shows that treatment status has mostly insignificant impact on quality beliefs—hence, quality differences do not explain our treatment effects by price.[42]

Lastly, individuals might also be oppose accepting money in exchange for reporting a seller. For instance, Roth (2007) argues that some exchanges become repugnant when money is incorporated into the transaction. While we cannot rule this out, there is at least a partial rate of substitution between cash payments and reporting or donating since WTPR and donation rates are responsive to our treatment. Moreover, this would only bias our estimates towards zero since higher cash payments would also entail a higher "cash repugnance." An individuals valuation from reporting sellers would thus be higher than what they reveal through cash incentives.

We use List (2020)'s SANS conditions to understand the experiment's generalizability to the target population of the entire United States. We selected our subjects to match the U.S. on race, Hispanic origin, age, and gender. However, the survey over samples subjects with a high-school education and under samples subjects with less than high school or more than a four-year degree. We reweight our data to match U.S. population moments to learn about the external validity of our estimates (Hotz et al., 2005). Reweighting does not materially

42. As an anonymous referee pointed out, Table B.9 shows that the WTPR is lower (even negative) for subjects who believe that quality increases with price. This suggests that our WTPR is biased down vs a measure that controls for quality differences. It is worth pointing out two things. First, this result is noisy, since we have only 300 individuals who equate price with quality. Second, even if we take these results at face value this does not challenge our main results, since we argue that individuals take costly actions to report sellers. Since the WTPR that we estimate is biased downwards, this is still true.

change the results.[43]

The completion rate after the randomization is 98%. There are also no motivational or incentive differences across treatments that materially affect attrition. Nevertheless, we use the non-parametric approach in Manski (1989) to derive treatment effect bounds with our data. Our results persist, with less precision, when using the bounding approach.

Regarding the naturalness of the experiment, we use a framed field experiment (see Harrison and List (2004)). Price gouging legislation activates during declared states of emergency. While atypical, we are operating in precisely the setting to which we wish to generalize. The text analysis of our sample of actual price gouging complaints (Section 2.2) shows that complaints about face masks and hand sanitizers were common. The iMPL may be unnatural to subjects, but we are comparing choices made in the iMPL to consequential choices made by thousands of individuals outside of the experiment.

Moreover, Berry et al. (2020) shows that choices made using within-person elicitations are congruent with decisions in more natural take-it-or-leave-it offers. Since the experiment takes place online at the subject's own pace, subjects are free to seek information that would aid in their decision-making. The donation experiment mimics actions taken by private companies during other natural disasters (Uber, 2016). Further work should attempt to understand the WTPR for goods that do not have positive externalities and focus on understanding what drives the differences in mechanisms across goods.

2.7 Conclusions

In this paper we propose an incentivized reporting experiment (IRE). Using our theoretical model, we argue that reporting a seller for price gouging contains information about a distaste for voluntary transactions between third-parties, as well as expected benefits from punishing

43. However, we cannot evaluate unobservable differences between our subjects and those who would never participate.

the seller. Based on this, the IRE elicits the willingness to pay of individuals to report (WTPR) a seller to the authorities for price gouging. While there is some correlation between the elicited WTPR and stated-preference measures of whether prices are considered to be excessive, our revealed-preference approach might be more useful to determine the price ranges sanctioned by price gouging laws. Beyond this application, IREs could be used to study repugnance toward activities in which enforcement requires that illicit activity is reported to the authorities.

The IRE allowed us to show that most individuals value reporting price increases of face masks and hand sanitizers during the first wave of COVID-19, although there is some polarization. Individuals also respond to the seller's price and increase their willingness to pay to report when facing more expensive sellers. The documented measure implies opposition to transactions that some participants would find beneficial and thus presents a consequential example of repugnant transactions in the field. Our results are consistent with prior studies on repugnance, such as Elías et al. (2019), who also find that individuals are willing to tolerate inefficiencies in order to reduce repugnance and that they have polarized preferences. Moreover, the experiment shows that raising the price of essential products during emergencies has economically significant negative externalities on third-parties. While this is not an argument for or against anti-price gouging laws, this complicates any welfare evaluation of these policies.

A choice between a $5 gift card and having us donate an item of PPE purchased from a price gouger clarifies the underlying motivation behind the opposition to large price increases during emergencies. We find evidence for distaste for seller profits in the case of hand sanitizers but a higher priority for others' consumption when it comes to face masks. The fact that individuals may obtain negative payoffs from profits in the case of some products suggests an additional welfare cost of policies such as subsidies—that potentially increase profits—versus price controls. Moreover, further research should understand what drives the

heterogeneity across products.

CHAPTER 3

CASH: A BLESSING OR A CURSE?[1]

The use of cash has received considerable attention from policymakers and academics who, many times, have expressed their negative assessment of its role. Many argue that restricting cash usage would diminish criminal activities including tax evasion, see e.g., Rogoff's book "The curse of cash", and the ensuing debate on the costs and benefits of a "war on cash" (e.g., Bundesbank (2017)). A concrete policy that was recently carried out along these lines was the demonetization in India (see Chodorow-Reich et al. (2018); Lahiri (2020)). However, despite the relevance of the issue, the scholarly debate on the issue is scant, a situation also lamented by Sands (2017). First, data on cash usage and its relation with illegal activities have been the subject of very few scholarly analyses (e.g., Wright et al. (2017); Gandelman et al. (2019); Schneider (2017)). Second, to the best of our knowledge, there are no estimates of the social benefits of curbing cash-related crimes. Finally, although a preliminary quantitative assessment of the private costs of banning cash is given in Alvarez and Lippi (2017) and Briglevics and Schuh (2020), the scope of those results is limited to households who have access to both means of payments, while in actual economies cash is the only payment instrument for many households and the adoption of alternative technology is costly. Our contribution is an attempt to tackle these issues upfront, featuring three ingredients that are essential to discuss the issue rigorously: detailed micro data, explicit identification of the causal effects of cash on illegal activities, and an explicit model of both the costs and benefits associated to cash usage.

In this paper, we present a welfare analysis of the consequences of restricting the use of cash, accounting for both social benefits and private costs. Our application considers the case of Mexico for three reasons. First, the availability of detailed data sets on the

1. This article was published in the Journal of Monetary Economics, Volume 125, January 2022, Fernando Alvarez, David Argente, Rafael Jimenez, Francesco Lippi, Cash: A Blessing or a Curse?, Pages 85-128, Copyright Elsevier (2021).

access and use of cash, by both households and firms, allows us to document cash usage in the country with high precision. Second, we take advantage of two recent policies that aimed to restrict the use of cash in the country to study the impact of cash on criminal activities and informality. Third, the availability of estimates about the cash-credit elasticity of substitution, as estimated for the case of Mexico using experimental data in Alvarez and Argente (2020a) and observational data in Alvarez and Argente (2020b). Quantifying the substitution elasticity is key to assess the consequences of cash elimination to cash-only households, who are still a non-negligible fraction of the population.

We begin by documenting several facts on the use of cash in Mexico. Using a variety of household-level surveys and firm-level surveys, we show that although more than 90% of transactions in Mexico are paid in cash, mixed users – individuals who have access and use both cash and cards – are prevalent and widespread in Mexico. More than 50% of households in Mexico are mixed users. We show that cash is the most important payment method even for mixed users; approximately 80% of their expenditures are paid in cash. The prevalence of cash among these users is relevant because policies restricting the use of cash could impact mixed users if cash and cards are not perfect substitutes.

In order to measure the social benefits of restricting the use of cash, we estimate the impact of two policies that lead to a reduction in cash usage in Mexico on outcomes such as crime, informality, and tax evasion. First, we study a policy that changed the payment method of the conditional cash transfers program in Mexico (Prospera). Approximately 1.3 million beneficiaries of the program received a debit card between 2009 and 2015. The policy aimed to increase financial inclusion in the country and discourage the use of cash. Indeed, the analysis of this policy in Bachas et al. (2017) and Higgins (2019) shows that, after the rollout of cards, both the number of ATM transactions and the prevalence of POS terminals increased drastically. The fact that the implementation of this policy was staggered across randomly selected localities allows us to use an event-study design to estimate the

implications of a reduction on the availability of cash on several outcomes. We find that the policy had a small significant effect on theft and on robberies. On the other hand, we do not find an impact of the rollout of debit cards on homicides, informality, or local tax revenue.

Second, we consider the impact of a policy that reduced the regulatory requirements to implement ATM-sharing agreements between banks (either commercial or development). This policy was implemented by the Bank of Mexico in October 2014 and resulted in the gradual adoption of agreements throughout the period of study (2014-2019) between different banks to share their ATM infrastructure. Each agreement reduced the fees of ATM operations, such as balance checks and withdrawals, and thus provides plausible exogenous variation for cash holdings of clients of the agreeing banks. Because agreements occur between banks at the National level, a natural empirical strategy is a shift-share design that exploits the differential exposure of municipalities to these common agreements. In particular, municipalities that have a large presence of banks in an agreement will be more exposed to it relative to municipalities with a small presence of these banks. In our preferred specification, we observe an increase in the growth rate of ATM withdrawals after an agreement in municipalities that have a higher exposure relative to those with lower exposure. Consistent with the results of the first policy experiment, we find an impact of the policy on thefts and robberies, particularly on those where pedestrians are the victims. The policy had no impact on the homicide rate or on the total number of informal workers.

Using the observed patterns of cash usage and the estimated elasticity of crime to cash, we turn to the estimation of the effects of restricting cash usage to households, which includes a complete ban on cash. The essential ingredients of our model are a general utility function that considers goods paid in cash as a different good than those paid with credit. To analyze the welfare effect of policies restricting cash payments, we start with an initial situation where agents face the same price for goods paid in cash and goods paid in credit. Starting from this situation, we consider the effect on agents' welfare under several alternative scenarios, such

as a full ban on goods paid in cash as well as other intermediate policies such as limits on the value of cash payments and taxes on the goods paid in cash. We parameterize the model using observations for individuals grouped over different income groups and considering their consumption over several categories of goods using the National Survey of Household Income and Expenditure (ENIGH).

We also consider the social benefits that each policy may bring by reducing the prevalence of criminal activities. Given that a reduction in cash caused a statistically significant reduction in theft and robberies in the two policies we studied, we focus on these two crimes. We rely on victimization surveys to measure the prevalence of these crimes and calculate their direct costs, measured as the fraction of GDP that is stolen. We quantify the indirect costs, measured by the deadweight losses of the crimes, which include tangible costs (e.g., preventive police cost, judiciary costs) as well as the intangible costs (e.g., psychological costs for the victim) drawing from the economics of crime literature.[2]

The private losses of a 40% tax on cash are approximately 6% of GDP. A key parameter for this result is the elasticity of substitution between cash and credit, which in our baseline analysis is approximately equal to $\eta = 5$. We show that the magnitude of these welfare losses of restricting the use of cash is high for a wide range of parameter values, including a doubling or tripling of η. On the other hand, the deadweight losses of cash-related crimes gives an upper bound for the social benefits of eradicating theft and robberies, which is about 1.3% of GDP. Even if we consider the social benefits of eradicating all crime (approximately 3% of GDP using UK data) these benefits are half of the costs associated with a 40% tax on cash and less than a third of those associated with a full ban on cash, which are approximately 10% of GDP.

The remainder of this paper is organized as follows. In Section 3.1, we present several

2. Examples of these estimates can be found in Price (2000); Albertson and Fox (2008); Heeks et al. (2018) and the summary of the main estimates for the tangible and intangible indirect costs collected in the meta study by Wickramasekera et al. (2015).

stylized facts of the use of cash in Mexico. Section 3.2 studies the impact of the Mexican government's rollout of debit cards to the beneficiaries of its conditional cash transfers program on outcomes such as criminal activities and informality. In Section 3.3, we study the impact of newly established ATM-sharing agreements between banks on the same outcomes. Section 3.4 develops a simple model to quantify the private costs of taxing the use of cash and calculates the social benefits of elimination cash-related crimes. Section 3.5 concludes.

3.1 Empirical Facts

We start our analysis by documenting cash usage in Mexico. We rely on four detailed data sources. First, the National Survey of Financial Inclusion (ENIF), which provides a detailed description of the use of payment methods at the household level. We complement this evidence with the National Survey of Household Income and Expenditure (ENIGH), which allows us to estimate the share of expenditures in cards by type of good. We then use the Financial Inclusion Databases (BDIF) collected by the National Banking and Securities Commission (CNBV) to characterize the access to payment methods in Mexico. The data set includes information of the bank branches, ATMs, point-of-sale terminals (POS), bank accounts and debit and credit cards at the municipality level. Lastly, we have use the National Survey of Firms' Financing (ENAFIN), a confidential data set provided by the Mexican Statistical Agency (INEGI) that includes information of the payment methods accepted by the firms in Mexico. The data allow us to determine the share of firms in the economy that only take cash as a payment method and their characteristics.

We then analyze the adoption of cards in the Prospera program and the implementation of ATM-sharing agreements between banks. In the analysis of the rollout of cards in Prospera, we use the administrative data of the program obtained through a freedom-of-information request to the Federal Institute for Access to Information (INAI). We obtained information of ATM-sharing agreements and the associated percent reduction in fees from the CNBV.

We use data from the National Employment Survey (ENOE) and the State and Municipal Public Finances (EFIPEM) to analyze the impact of these two policies on informality and tax collection. Lastly, to analyze the impact of these policies on crime we use i) Statistics of Registered Deaths collected by INEGI, ii) Registered Crimes collected by INEGI and iii) Criminal Incidence from the Executive Secretariat of the Public Security National System (SESNSP). We provide more details of each these data sets in Section C.5.

3.1.1 Cash is the most important payment method in Mexico

Cash is the main method of payment in Mexico. According to the National Survey of Household Income and Expenditure (ENIGH), a national representative survey collecting information on households' expenditures and means of payment, around 90% of payments are conducted in cash.[3] Table 3.1 shows that across all income groups and type of goods, cash is the most common means of payment. The table shows that cash is most commonly used by households in the bottom tercile of the income distribution.

The prevalence of cash can in part be explained by the lack of access to financial infrastructure. There are only 1.5 branches per 10,000 adults in Mexico, which represents a level below countries with a similar GDP per capita. The lack of financial infrastructure is more pronounced in rural municipalities where only 8% of rural municipalities have an ATM; 99% of municipalities with more than 50,000 inhabitants have an ATM.[4]

3. Figure C.2 shows that approximately 95% of respondents of the National Survey of Financial Inclusion (ENIF) reported cash as their most frequent payment method for transactions below 20 USD and 87% for transactions above 20 USD. Figure C.3 shows this fraction is similar for payments in sectors.

4. On average, each adult withdraws cash 17.3 times per year. Figure C.4 shows that the ATM and POS transactions are mainly concentrated in the cities (darker colors) whereas rural populations have less than one ATM on average.

Table 3.1: Share of Expenditures Paid in Card by Type of Good

Note: The table reports the share of expenditures paid in card by type of good. The table also shows the share of consumption by type of goods. The source is the National Survey of Household Income and Expenditure (ENIGH), which was conducted from August 21st to November 28th, 2018. The information is based on a diary of daily expenditures collected along with the survey. Households are asked to report the payment method they use for each good as well as the total amount spent on each. The table splits households into terciles according to their reported income. Columns (1)-(3) show the share of expenditure in payment methods other than cash. Columns (4)-(6) show the share of consumption by type of good and for each income group.

	Share Card			Share Consumption		
	Bottom	Middle	Top	Bottom	Middle	Top
	(1)	(2)	(3)	(4)	(5)	(6)
Food, Alcohol, Tobacco	0.001	0.004	0.050	0.557	0.491	0.376
Housing	0.002	0.001	0.028	0.037	0.032	0.025
Utilities	0.002	0.005	0.058	0.074	0.089	0.100
Education, Culture, Recreation	0.003	0.008	0.102	0.021	0.033	0.046
Cleaning	0.004	0.013	0.052	0.055	0.049	0.059
Personal Care	0.004	0.013	0.097	0.075	0.074	0.064
Communication and Vehicle Services	0.005	0.012	0.062	0.052	0.081	0.112
Clothing	0.010	0.024	0.164	0.059	0.066	0.072
Domestic Utensils	0.021	0.043	0.201	0.005	0.005	0.007
Transport	0.031	0.028	0.162	0.016	0.027	0.066
Health	0.033	0.018	0.152	0.031	0.028	0.036
Domestic Appliances	0.052	0.067	0.130	0.015	0.019	0.027
Entertainment	0.159	0.168	0.311	0.003	0.005	0.009

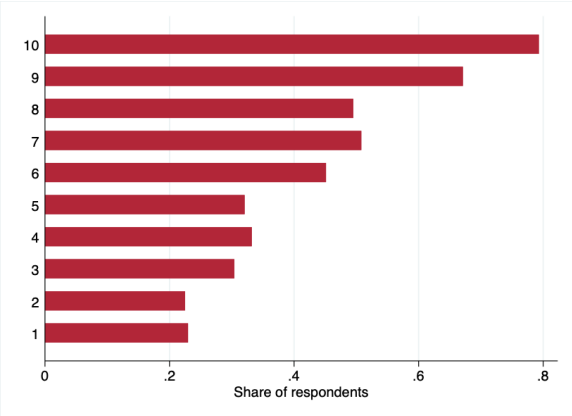
3.1.2 Mixed users are widespread in Mexico

However, the lack of financial infrastructure is not the most prevalent reason for why people do not have access to bank accounts or cards.[5] In fact, the majority of the Mexican population (70 %) have access to at least one financial product (i.e. a bank account, some form of formal credit, retirement savings, etc.) and half have at least one debit or credit card. This means that the majority of the Mexicans are mixed users, individuals who have access

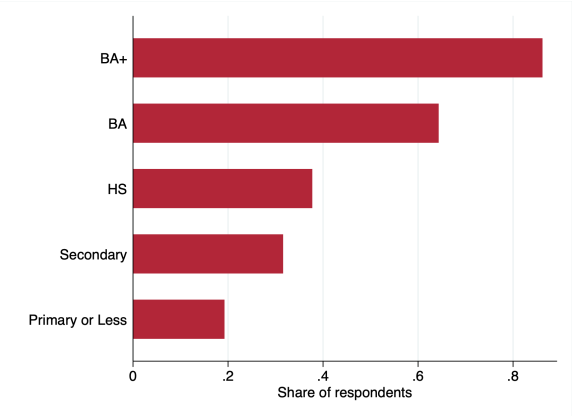
5. When those who do not have a bank account were asked in the National Survey of Financial Inclusion (ENIF), "what is the reason you do not have a bank account?" 33% respond that they do not have enough earnings, 27% respond that they do not need it, 11% respond that they do not meet the requirements. Less than 3% respond that they do not have an account because the bank is far. Similarly, when those who do not have a card were asked in the National Survey of Financial Inclusion (ENIF), "what is the reason you do not have a card" 32% respond that they do not like debt, 26% respond that they do not need it, 23% respond that they do not meet the requirements. Less than 2% respond that they do not have a card because the bank is far. We present these statistics in Figure C.5.

and use both cash and cards. Panel (a) in Figure 3.1 shows the fraction of people that have access to either a debit or a credit card by income decile. Not surprisingly, as the level of income or the level of education increases, the likelihood of having access to debit or credit cards also increases. Panels (c) and (d) show a similar pattern for bank accounts.

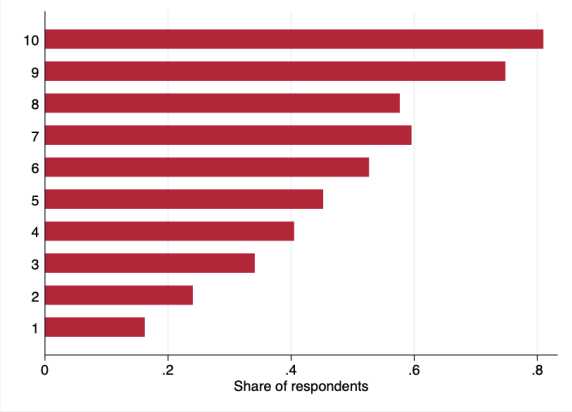
Figure 3.1: Access to Cards and Bank Accounts



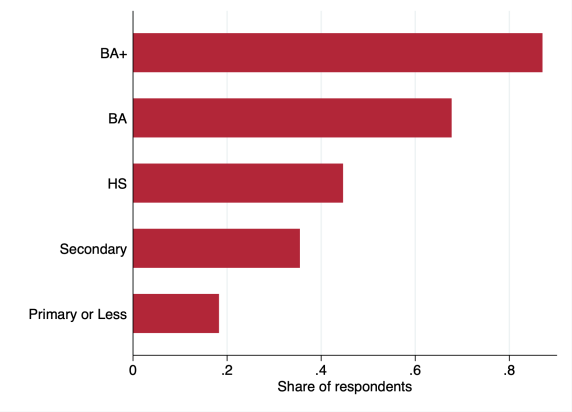
(a) Own a Card (Income Decile)



(b) Own an Account (Education)



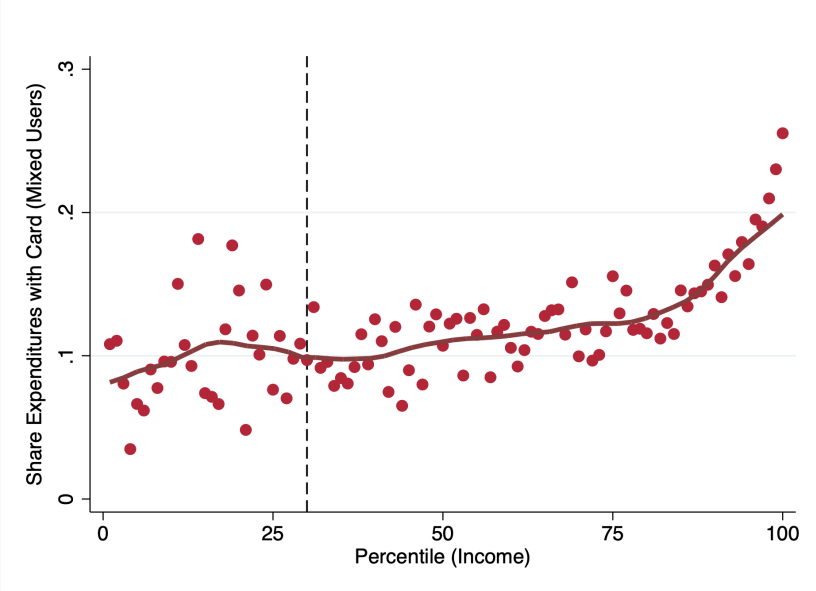
(c) Own an Account (Income Decile)



(d) Own an Account (Education)

Note: The figure shows the share of households in Mexico who have used a debit card in the last three months (from the time they were surveyed). Panels (a) and (b) shows the share of households by income deciles and education respectively. Panels (c) and (d) show the share of households who have a bank account by income and education. The data comes from the 2018 National Survey of Financial Inclusion (ENIF).

Figure 3.2: Share of Payments with Card - Mixed Users



Note: The figure shows the share of expenditures paid with payment methods other than cash for each percentile of the income distribution of Mexico. The sample of households include only those who reported making at least one payment with a method other than cash. Payment methods other than cash include debit cards, credit cards, transfers, and mobile payments. The source is the National Survey of Household Income and Expenditure (ENIGH), which was conducted from August 21st to November 28th, 2018. The information is based on a diary of daily expenditures collected along with the survey. Households are asked to report the payment method they use for each good as well as the total amount spent on each.

3.1.3 Cash is the most important payment method for mixed users

Despite the prevalence of mixed users, cash accounts for the majority of payments in Mexico. Among mixed users, cash is also the most used method of payment. Figure 3.2 shows the share of payments with card for those with either a debit or a credit card. The figure is calculated using ENIGH and shows that mixed users pay approximately 21% of their total expenditures in card. Table 3.2 shows that, among mixed users, goods such as food and housing are more likely to be paid in cash. On the other hand, mixed users are more likely to pay with cards for goods used for entertainment (e.g. TV, DVD players, radios, video games, musical instruments). Panel (a) of Figure C.2 shows that approximately 90% of respondents of the National Survey of Financial Inclusion (ENIF) reported cash as their

most frequent payment method for transactions below 20 USD and 70% for transactions above 20 USD. Figure C.7 shows that approximately 85% of people respond that cash is the most frequent payment method when paying for their taxes, services, and transportation.

Table 3.2: Share of Expenditures Paid in Card by Type of Good: Mixed Users

Note: The table reports the share of expenditures paid in card by mixed users and by type of good. The table also shows the share of consumption of mixed users by type of goods. The source is the National Survey of Household Income and Expenditure (ENIGH), which was conducted from August 21st to November 28th, 2018. The information is based on a diary of daily expenditures collected along with the survey. Households are asked to report the payment method they use for each good as well as the total amount spent on each.

	Share Card Mixed	Share Consumption Mixed
Food, Alcohol, Tobacco	0.159	0.337
Housing	0.083	0.025
Utilities	0.178	0.093
Education, Culture, Recreation	0.246	0.054
Cleaning	0.136	0.069
Personal Care	0.326	0.056
Communication and Vehicle Services	0.198	0.103
Clothing	0.455	0.076
Domestic Utensils	0.466	0.034
Transport	0.352	0.086
Health	0.372	0.044
Domestic Appliances	0.344	0.034
Entertainment	0.653	0.014

Mixed users might prefer cash if they are likely to be victims of credit card related crimes. Panel (a) Figure C.8 shows this is not the case; very few have been victims of identity theft, card cloning or fraud. Alternatively, it is possible that most people receive their wages in cash, thus, increasing the likelihood they spend their earnings in cash. Panel (b) shows that those who report owning a card are also more likely to receive their payments directly into their bank accounts and, as a result, are less likely to use cash in order to evade taxes.

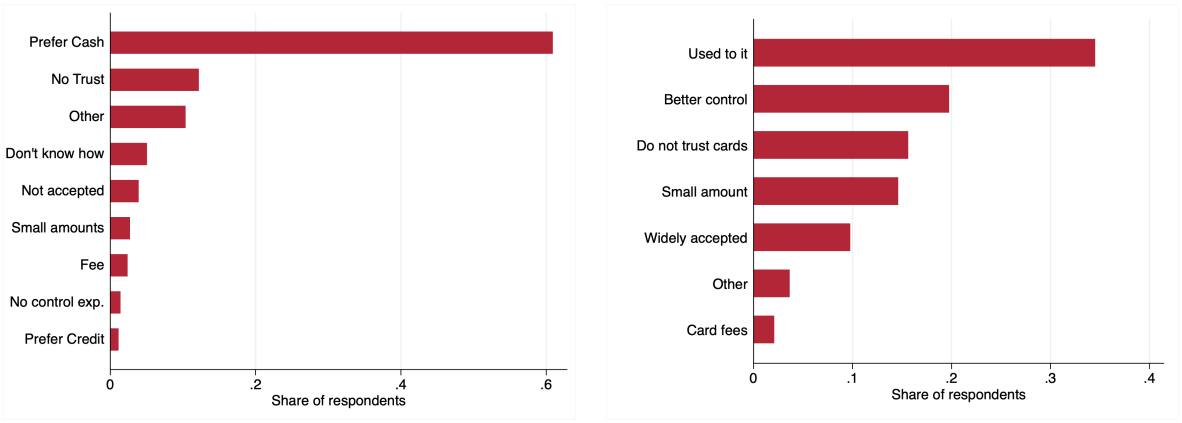
Another alternative is the size of the informal sector. There are approximately 4 million firms in Mexico, 99.8% are small and medium enterprises (52% of GDP and 72% of total employment). Only 43.3% are formal and account for 77.48% of GDP. Given that informal firms can be found by tax authorities if they accept means of payments other than cash, it

is unlikely that they accept another payment method. However, according to the National Survey of Enterprise Financing (ENAFIN), even among formal firms, cash is the most important payment method. Figure C.9 shows the share of firms that accept credit or debit cards in the formal sector according to their size and sector. The figure shows that, even among registered firms that pay taxes, most firms take only cash. This is true across sectors and for large (more than 100 employees) and micro firms (6-10 employees). The figure also shows that, consistent with the share of payments in card by mixed users obtained for household survey data (ENIGH), the share of payments received in debit and credit cards among formal firms is approximately 25%.[6]

However, few mixed users report the fact businesses do not accept cards as the primary reason why they do not use a card. In fact, when those who own a card were asked in the National Survey of Financial Inclusion (ENIF), "why don't you use your card?',' more than 60% respond they prefer cash. Panel (b) of Figure 3.3 shows that when the same people were asked, "why do you prefer cash?", 35% respond that they are used to it, 20% respond that it allows them to have better control of their finances, 15% respond that they only make payments in small amounts, 15% respond that they do not trust cards, 10% respond that they use cash because it is widely accepted, 2% respond that they want to avoid card fees, and the rest had other reasons.

6. When firms are asked, "what are the reasons you do not accept cards as payment method?" The most common answer for large firms is that they prefer transfers since they receive large payments. Micro firms, on the other hand, respond that they prefer cash. For micro firms, 17% respond they prefer cash since they receive payments of small amounts and 16% respond that it is too costly for them to accept cards as a payment method. Across sectors, the reasons firms do not accept card are consistent. For large amounts, firms prefer transfers. For small amounts, firms prefer cash. Across all sectors, an important share of firms state their preference for cash. These results can be found in Figure C.11 and Figure C.10. Interestingly, cash is also a very important payment method use by firms to cover their inputs and payrolls. Figure C.12 shows that almost 40% of large firms pay for their inputs in cash. More than 35% of micro firms pay for their payroll in cash. Figure C.13 shows that a large share of manufacturing and construction firms in the formal sector pay for their payrolls in cash. In the commerce and services sector, more than 30% of firms pay for their inputs in cash.

Figure 3.3: Mixed Users: Why Do You Prefer Cash



(a) Why Don't You Use It? (b) Why Do You Prefer Cash?

Note: Panel (a) shows the responses of households to the question "why don't you use your debit card?". Panel (b) shows the responses of households to the question "Why do you prefer cash?." The sample of households report owning a debit or a credit card. The data comes from the 2018 National Survey of Financial Inclusion (ENIF).

3.2 Prospera: Card Adoption

To study the impact of a reduction of the use of cash on crime, informality, and tax evasion, we take advantage of a large shock to consumers' adoption of debit cards in Mexico. Between 2009 and 2012, the Mexican government disbursed more than one million debit cards as the new payment method for its conditional cash transfer program, Prospera. By 2015, the program had distributed approximately 1.3 million debit cards. The program, previously known as Progresa (1997-2002) and Oportunidades (2002-2014), was a conditional cash transfer program targeting poor households with an estimated income per capita lower than the minimum necessary to acquire the basic food basket and whose social-economic conditions hinder the development of their members in terms of nutrition, health and education.[7]

The program provided cash transfers every two months. These transfers were conditional

7. By 2008 the size of the program stabilized after reaching one-fourth of Mexican households and covering virtually all municipalities. In 2015, the last year of our sample, the program represented approximately 1.6% of Mexico's national budget (equivalent of 0.4% of GDP) (Dávila Lárraga, 2016).

on attendance to a scheduled appointment with health services and enrolling children in school as well as encouraging them to attend school on a regular basis. The size of the payment depended on the compliance of these co-responsibilities and on the characteristics of the family; it averaged US$150 per two-month payment period during the years of the card rollout.

Before the card rollout, each beneficiary had a savings account at National Savings and Financial Services Bank (Bansefi), a government bank created to promote savings and financial inclusion. Benefits were deposited in this account and beneficiaries could choose to withdraw any amount at any point in time. Nonetheless, Bachas et al. (2017) report that, prior to the debit card rollout, 90% of beneficiaries made one trip to the bank per payment period, withdrawing their entire transfer. This is because the benefits could only be withdrawn at a Bansefi branch which are on average 5 kilometers away from an urban beneficiary household. The debit card rollout allowed beneficiaries to withdraw their benefits from any ATM and to pay using their Visa debit card at any business accepting cards as payment method. Thus, the program considerably reduced the travel costs of beneficiaries increasing the number of ATM withdrawals (Bachas, Gertler, Higgins, and Seira, 2017) and the number of times they pay in POS terminals using their cards (Higgins, 2019).

The rollout of debit cards was implemented at the locality level, a geographical unit smaller than a municipality.[8] At each treated locality (those chosen for the debit card rollout), all beneficiaries obtained a debit card during the same payment period. Because the payments were disbursed every two months, the administrative data from Prospera identifies at this frequency the timing of the rollout and the number of beneficiaries in each locality.

Bachas et al. (2017) argue that Prospera officials did not target localities with particular attributes because they wanted to test their administrative procedures for the rollout

8. On average, there are approximately 60 localities in each municipality (median 30).

on a quasi-representative sample. This is somewhat inconsistent with Higgins (2019). In particular, on the one hand, Bachas et al. (2017) show that the timing of the treatment is uncorrelated with pre-treatment wages, prices, POS terminals, bank branches, ATMs, debit cards, credit cards, beneficiary savings, number of ATM withdrawals, program beneficiaries, or whether the party in power at the municipal level corresponds with the party in power at the national level. On the other hand, Higgins (2019) argues that Prospera determined that it was only worthwhile to distribute debit cards in urban localities with sufficient ATM infrastructure. He writes "the timing of the shock is not correlated with levels or trends in locality-level financial infrastructure or other observables (conditional on being included in the rollout), but the initial selection of which localities to include in the rollout is correlated with locality characteristics." Thus, to be conservative, in what follows our identifying assumption relies on the work by Higgins (2019): conditional on being included in the rollout, the timing of when a locality received the card shock is uncorrelated with locality-level unobservables or other trends.

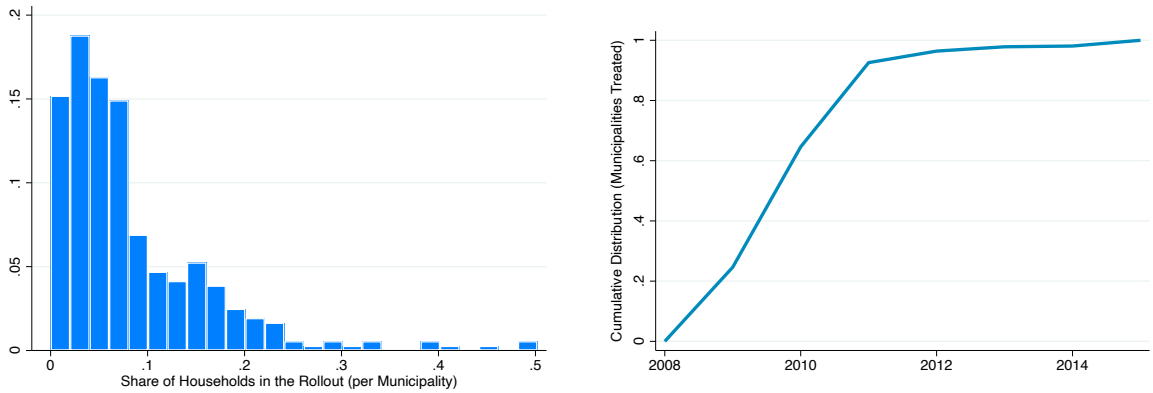
In contrast to Higgins (2019), who only focuses on urban localities, we include all treated localities with at least one thousand inhabitants, a total of 966 localities and 418 municipalities. In treated municipalities, approximately 40% of localities were included in the rollout and more than half of beneficiaries received a card. Panel (a) in Figure 3.4 shows that within a treated municipality, approximately 8% of all households received a card. Panel (b) shows that the majority of cards were distributed before 2012. Nonetheless, we extend our analysis to 2015 since we observe in the data that some municipalities are included until that year.[9]

3.2.1 Event Study

We begin our analysis by studying how the rollout affected cash-related outcomes in several localities and municipalities in Mexico. We use a fully dynamic event-study specification to

9. Figure C.14 shows the geographic coverage of the rollout. It shows that beneficiaries receiving cards through Prospera are distributed over the entire country.

Figure 3.4: Households in the Rollout and Treated Municipalities



(a) Households in the Rollout

(b) Municipalities Across Time

Note: Panel (a) shows the share of households in a municipality that were part of both Progresa and the rollout of debit cards in each of the treated municipalities. Panel (b) shows the cumulative distribution of municipalities treated. The source for both panels is the administrative data of the Prospera program from 2007 to 2015.

compare several outcome variables before and after the rollout of debit cards. Let Y_{lmt} be an outcome variable for locality (or municipality when locality level data is not available) m at time t (e.g. informal workers, local taxes, number of homicides, etc). The specification for our event study is as follows:

$$Y_{mt} = \alpha + \sum_{k=-\infty}^{\infty} \gamma_k \mathbb{1}\{K_{mt} = k\} + \theta_m + \lambda_t + \zeta X_{mt} + \epsilon_{mt} \tag{3.1}$$

where θ_m are locality-fixed effects and λ_t are time effects. K_{lt} denotes the number of periods relative to the rollout of debit cards so that γ_k for $k < 0$ corresponds to pre-trends and $k \geq 0$ corresponds to dynamic effects k periods after the rollout. X_{mt} represent a set of locality-specific time-varying controls such as the number of families in Prospera in locality i at time t. Since all the localities are treated, we require an additional restriction on the pre-trends in order to estimate the time fixed effects. Moreover, since different locations become treated at different times, heterogeneous treatment effects across time could be relevant.

For this reason, we follow the methodology developed by Borusyak et al. (2020) to estimate a robust and efficient estimator that allows the implementation of two-way fixed effects in staggered designs.[10] This methodology is also robust to using locations that have not been treated yet as controls.

Alternatively, in Table 3.3 and Section C.3 we implement a semi-dynamic event study design for several specifications:

$$\ln Y_{mt} = \alpha + \beta \ln \text{CardShock}_{mt} + \theta_m + \lambda_t + \zeta X_{mt} + \epsilon_{mt} \tag{3.2}$$

where CardShock_{mt} is an indicator that equals to one when the municipality is treated. This specification does not have identification issues if the timing of the treatment is random, as described by the Prospera authorities, conditional on the fixed effects and controls. It estimates the average treatment effect, assumed to be homogeneous, for all periods following the event.[11] Since the error term might be both serially and cross-sectionally correlated, we use Driscoll and Kraay standard errors in the semi-dynamic specifications. For some of the outcomes we study, there is no data available at the locality level and bi-monthly frequency. In such cases, data availability determines both the aggregation of the outcome variables and the time frequency we consider, we provide details for each outcome variable below.

We begin by presenting evidence on the adoption of debit and credit cards. We use data from the Mexico's National Banking and Securities Commission (CNBV). Since the

10. The estimator developed by Borusyak et al. (2020) takes an "imputation" form and is constructed in three steps: i) unit and period effects are fitted by regression on untreated observations only, ii) unit and period effects are used to impute the untreated potential outcomes and obtain an estimated treatment effect, and iii) a weighted average of these treatment effect estimates is taken with weights, corresponding to the estimation target. Our results are also robust to using the methodology developed by De Chaisemartin and D'Haultfœuille (020b).

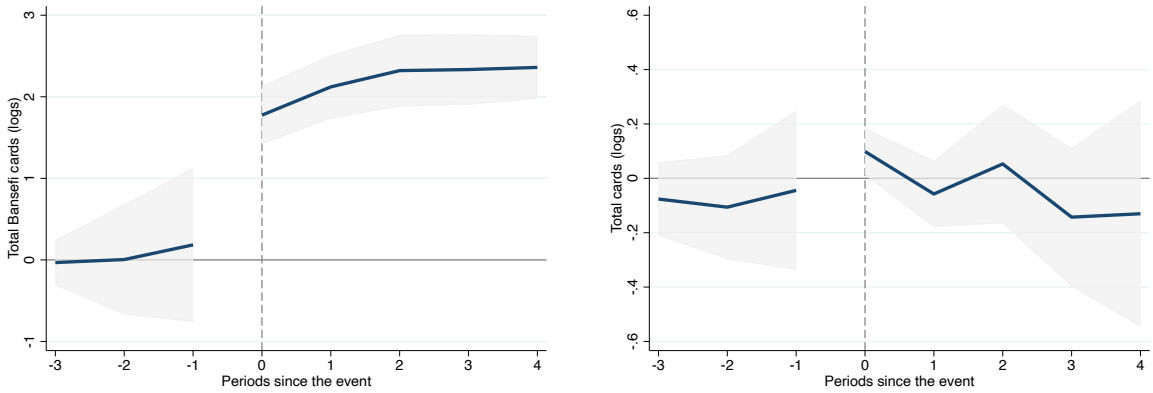
11. Section C.3.1 also presents results where we use, instead of a dummy variable equal to one after the treatment, the share of households in a municipality that are both part of Prospera and the were included in the rollout.

data is quarterly and at the municipality level, and the administrative data from Prospera is bi-monthly and at the locality level, we implement Equation 3.1 at the municipality and bi-annual level since both data sets coincide at this frequency. We first verify that, at this frequency and level of aggregation, our specification is informative of an increase in debit cards provided by Bansefi. Panels (a) of Figure 3.5 shows a substantial increase in Bansefi debit cards after the start of the rollout. Higgins (2019) studies the same specification for non-Bansefi debit cards and documents their prevalence increases after the shock. He argues that this increase is due to indirect network externalities, where other consumers benefit from the increase of debit card users from the Prospera program. Here, we estimate the same specification but including all municipalities treated, instead of only urban municipalities, and extending the administrative data from Prospera to 2015. The graphs show that, conditional on municipality- and time-fixed effects, no pre-trends appear before the shock. This pattern is consistent with the timing of the introduction of cash being randomly assigned conditional on the municipality- and time-fixed effects. The identification assumption of this exercise is precisely that the rollout in these municipalities was not anticipated. Panel (b) shows a smaller and more transient response of total cards, which include all debit cards and credit cards, after the rollout. Columns (1)-(3) of Table 3.3 show the results of the semi-dynamic event study and Table C.3 show other robustness checks.[12] Column (3) shows that, under the assumption of constant treatment effects, the total number of cards increased approximately 11% after including controls such as the income per capita of the municipality, total employment, number of progresa families, and total population of the municipality. The effect of the rollout on debit cards (excluding Prospera cards) is larger than that on all debit cards. If we account only for the intensive margin (e.g. restricting

12. Results presented in Table 3.3 use the inverse hyperbolic sine transformation in the dependent variable to account for the extensive margin at the municipality-period level. In this case, Bellemare and Wichman (2020) show that the percentage change in the dependent variable (for large values of it) due to a discrete change in a dummy variable is approximately $e^{\hat{\beta}} - 1$. In turn, a Taylor expansion of $e^{\hat{\beta}} - 1$ show that it is close to $\hat{\beta}$ for small values of $\hat{\beta}$.

the sample to municipality-periods with positive debit cards) the effect of the rollout on debit cards (excluding Prospera cards) is larger.[13] This evidence, as well as the evidence in Higgins (2019) and Bachas et al. (2017) on the increase in ATM transactions after the rollout, is indicative that the prevalence of cash decreased in the treated municipalities.

Figure 3.5: Event Study: Bansefi Cards and Total Cards



(a) Debit (Prospera Cards)

(b) Total Cards

Note: The graph shows the evolution of Bansefi debit cards and total cards before and after the rollout of cards. The figures plot the coefficients of γ_k after estimating Equation 3.1. The dashed line marks the period that cards were rollout in the municipality. Each period is a 6-months interval. The gray area depicts the 95% confidence interval.

Next, we study the impact of the shock on homicides. We use data of homicide victims based on the vital statics published by the INEGI. The data is collected from public health records filed by coroner's offices and it is based on death certificates identifying the cause of death. The data include the date and place of the homicide; thus the information is available at the locality and monthly level. Since the administrative data of Prospera is at the bi-monthly level, in this case we estimate Equation 3.1 at this frequency and include municipality $\times$ period effects to further control for trends. Panel (a) of Figure 3.6 shows our results. We do not find evidence that the debit card shock decreased homicides. Table C.5 presents the results under the semi-dynamic specification. Consistent with the dynamic

13. This effect is larger regardless of whether we measure the left hand side variable in logs or using the inverse hyperbolic sine transformation. These results are presented in Table C.4.

specification of the event-study, homicides do not decrease after the shock. If anything, in this specification, we find a small increase in homicides that is statistically significant but very small in economic magnitude.

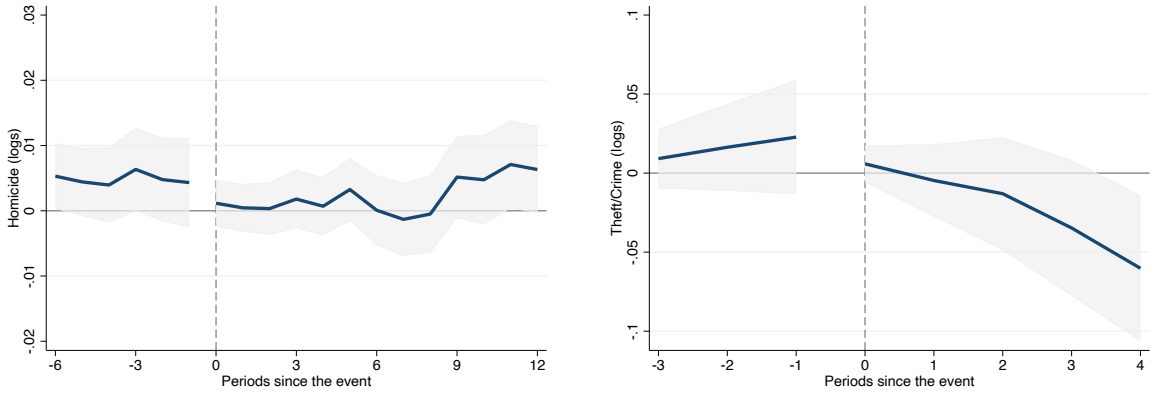
An alternative way to study the patterns of homicides is to combine two different data sets. From 2005-2010, INEGI collected crime statistics for each municipality and made it available at the State and Municipal Databases (SIMBAD). The data set is at the annual level and has information on robberies, damages, injuries, sex crimes, kidnapping, and homicides. Starting in 2011, the main source for information of criminal activity at the municipality level is reported by the Executive Secretariat of the Public Security National System (SESNSP). The data set is based on police investigations and includes the number of victims contained in those investigations.[14] Importantly, since the data is based on cases handled by law enforcement investigations, it often overestimates the number of homicides relative to the vital statistics. Nonetheless, Column (4) of Table 3.3 shows that we do not find any effect on homicides and Table C.6 shows that we find similar patterns when we use data based on death certificates or data from law enforcement cases.

Using the combined data set we are able to study other crimes, in particular theft, which includes burglary and robbery; our crime data before 2011 do not distinguish among them. Panel (b) of Figure 3.6 shows a small decrease in theft after the rollout of cards by approximately a 5% on average. Importantly, in this case, the dependent variable is the logarithm of total thefts divided by total crimes in municipality i and period t. We use this dependent variable in order to further control for potential trends on criminal activity. Column (5) of Table 3.3 shows a negative, but not significant, decline in thefts. Nonetheless, Table C.7 shows that in the semi-dynamic specification the decline in theft is significant in the unweighted specifications, particularly when we use either total thefts over population or total thefts over total crimes as dependent variables. Column (6) of Table 3.3 and Table

14. Although there are differences in methodology across these data sets, Figure C.15 shows that there is a smooth transition in the aggregate series for several crimes including homicides.

C.8 show our findings for total crimes at the municipality level. In this case, we do not find a significant decline. Overall, we find a small decline in theft after the rollout of debit cards but we do not find statistically significant evidence that total crime declined at the municipality level after the shock.

Figure 3.6: Event Study: Homicides and Thefts



(a) Homicides (b) Theft

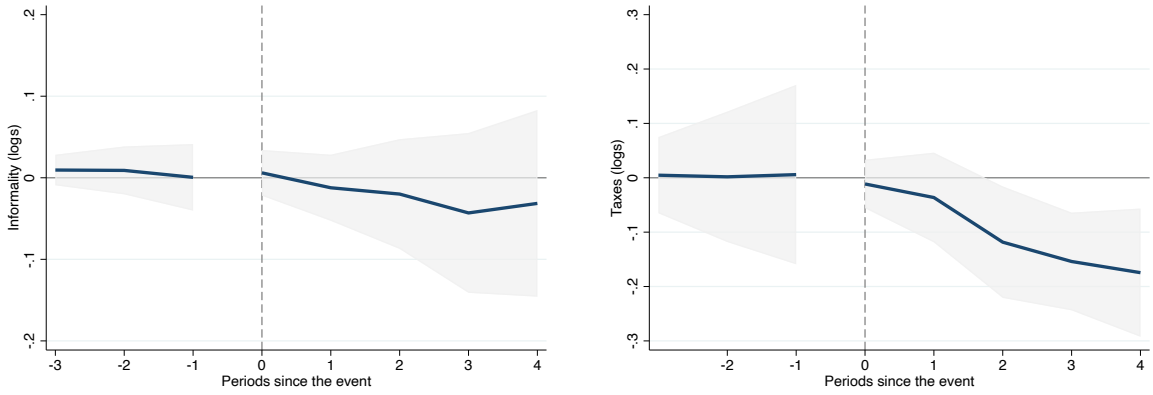
Note: The graph shows the evolution of homicides and thefts before and after the rollout of cards. The figures plot the coefficients of γ_k after estimating Equation 3.1. The dashed line marks the period that cards were rollout in the municipality. In Panel (a) each period is a 2-months interval and in Panel (b) each period is a year. The gray area depicts the 95% confidence interval.

Next, we study the impact of the shock on informality and tax collection. In Mexico, almost 60% of workers are informal. We use the National Employment Survey (ENOE), the main source of labor market statistics in Mexico. ENOE provides quarterly level representative samples of the national labor market.[15] Since the data from Progresa is bi-monthly, we aggregate the data at the bi-annual level. Our measure of informality is constructed using the definitions provided by INEGI. A worker employed in the informal sector is one that is employed, works for a non-agricultural economic unit that operates from the resources of the household, but without forming itself a company, so that the income, materials and equipment used for the business are not independent and/or distinguishable from those of

15. The data is a rotating panel (households are interviewed for five consecutive quarters and then replaced) and has 120,260 households and 420,000 individuals per quarter on average.

the household. Informal workers are employed with no benefits, health benefits only (universally provided to all workers by the government). Our definition of informal worker includes both workers in the informal sector and those outside the informal sector working without benefits.[16] Panel (a) of Figure 3.7 shows our results in the fully dynamic specification of the event study. We do not find an effect of the shock on the logarithm of informal workers in a municipality. Column (7) of Table 3.3 and Table C.9 show similar findings when we use the semi-dynamic specification. We find similar results when we consider the total number of self-employed workers as a dependent variable.

Figure 3.7: Event Study: Informality and Taxes



(a) Informality (b) Taxes

Note: The graph shows the evolution of informal workers and local taxes before and after the rollout of cards. The figures plot the coefficients of γ_k after estimating Equation 3.1. The dashed line marks the period that cards were rollout in the municipality. In Panel (a) each period is a 6-months interval and in Panel (b) each period is a year. The gray area depicts the 95% confidence interval.

To further check the impact of the shock on tax evasion, we use information of local tax collection from the State and Municipal Public Finances (EFIPEM) collected by INEGI. This database is the most detailed available account of public finances for both federal, state-level and local spending at the municipality-level and at annual frequency. The information is obtained from the Ministry of Finance of each state and from the Treasury of each municipality.

16. Our conclusions do not change if we use either of these measures separately.

It includes local tax collection of payroll taxes and real-estate taxes, among others.[17] We use the total taxes collected by each municipality in a calendar year as dependent variable. Panel (b) of Figure 3.7 shows our results. We find that the increase in the prevalence of cards in a municipality decreases the amount of taxes collected by the municipality. Column (8) of Table 3.3 and Table C.10 show similar but not significant results.

Table 3.3: Effect of Card Shock

Note: The table reports the results for the coefficient of β after estimating Equation 3.2. The dependent variable in Column (1) is the logarithm of debit cards. Column (2) use debit cards excluding those given as part of the Prospera program through Bansefi and Column (3) use the sum of debit cards and credit cards as dependent variable. The dependent variable in Column (4) is the logarithm of homicides using data from SESNSP based on criminal cases, in Column (5) the logarithm of total thefts and in Column (6) the logarithm of theft divided by total crimes. The dependent variable in Column (7) is the logarithm of informal workers and Column (8) is the logarithm of local taxes. We use the inverse hyperbolic sine transformation in all cases. The controls we use include income per capita, total employment, and total population, and the total number of families in the Prospera program. The specifications that are weights use the total population in the municipality. We use Driscoll and Kraay standard errors in all specifications.

	(1) Debit	(2) Debit Not Prospera	(3) Total Cards	(4) Homicides	(5) Theft	(6) Crimes	(7) Informal	(8) Taxes
Card Shock	0.1318***	0.1673***	0.1139***	0.0923	-0.0238	0.0164	0.0089	-0.0060
	(0.029)	(0.040)	(0.026)	(0.062)	(0.013)	(0.014)	(0.010)	(0.016)
Obs.	5,212	5,212	5,212	3,149	3,027	3,027	6,224	2,895
Municip.	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Weights	Y	Y	Y	Y	Y	Y	Y	Y

3.2.2 Discussion

Overall, we find evidence that the rollout of Prospera cards, which distributed debit cards to more than 2% of households in Mexico, increased the number of households owning either

17. At the local level, municipalities have their own treasuries and enforce their local tax law, which determines the structure of each tax. Local taxes account for approximately 21.5% of the total income of the municipalities. Our measure of local taxes do not include federal taxes collected by the Tax Administration Service (SAT), part of the Ministry of Finance. The main federal taxes are value added taxes (IVA), income taxes (ISR), and excise taxes (IEPS).

a debit card or a credit card. Despite the size of the shock, we do not find evidence that the shock decreased the total number of homicides, which we can study at the finest level of geographic aggregation and at the highest possible time frequency. If anything, under some specifications, there is a statistically significant increase in homicides although it is of small magnitude. We find a decline in thefts of approximately 2-5%; under some specifications, particularly those weighted by population, the estimates are not very precise. Overall, our results are consistent with Wright et al. (2017) and Gandelman et al. (2019) who show an impact of cash on property crimes but not on violent crimes.

The number of informal and self-employed workers are not statistically different before or after the rollout of cards. We find that the shock decreases tax collection, but a limitation of our analysis is that our measures of local taxes do not include federal taxes, including value added taxes. To the best of our knowledge, this information is not available at the municipality level. Hondroyiannis and Papaoikonomou (2017) show evidence that restrictions on the use of cash could increase VAT revenue. The impact of the card shock on VAT revenue in Mexico is not clear given the evidence presented by Higgins (2019). He shows that, although corner stores were more likely to adopt POS terminals after the shock, the change in consumption at corner stores was driven mostly by customers who already had cards. Furthermore, Bachas et al. (2017) show that Prospera beneficiaries responded to receiving a debit card by decreasing total consumption to finance an increase in overall savings.

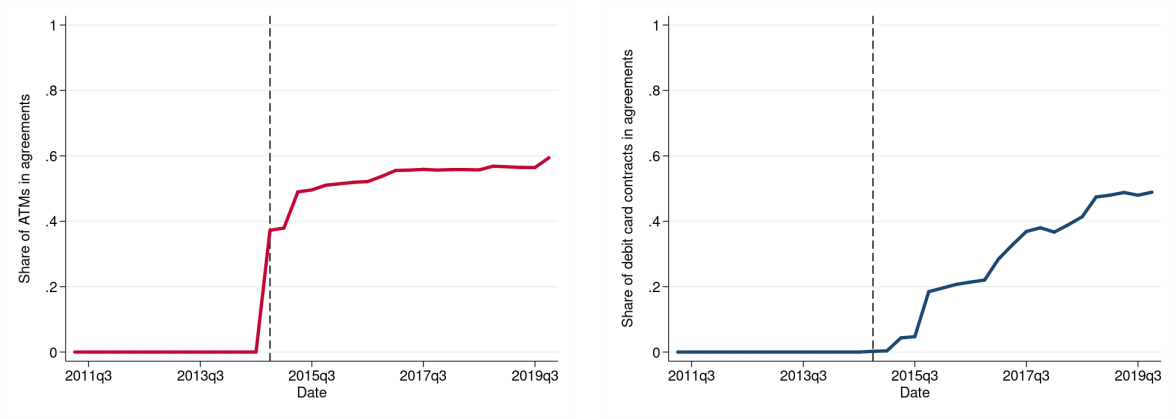
3.3 ATM-Sharing Agreements

Next, we consider a policy change that triggered 24 ATM-sharing agreements between banks throughout 2014-2019 (see the list of agreements in Table C.18).[18] When two banks agree to share their ATM infrastructure, customers of one bank can use the ATMs of the other

18. The 24 agreements happened between 29 different pairs of banks. There are approximately 64 banks in our sample (4,032 pairs of different banks).

at a reduced fee.[19] The Financial Reform of 2014 effectively allowed banks to celebrate these agreements—since the regulatory requirements were too restrictive before. The law was put in place in October 2014, when the Bank of Mexico issued the new requirements for the agreements.[20] The first agreement was implemented one month after the law and more agreements have been implemented every year since then. The reduction in ATM fees ranged from 7.2% to 100%. Figure 3.8 shows the share of ATMs and debit cards that were part of an agreement after the Financial Reform of 2014; more than 50% of ATMs and 50% of cards entered into some agreement between 2014 and 2019.

Figure 3.8: Share of ATMs and Debit Cards in Agreements



(a) ATMs (b) Debit Cards

Note: The figure show the quarterly fraction of ATMSs and the fraction of debit card contracts that belong to banks in agreements. The vertical dashed line indicates when the Bank of Mexico instituted the ATM-sharing agreements policy (October 2014). Note that an ATM or a card might enter in an agreement multiple times; in the figure, we consider the first time each entered in an agreement. The data comes from the CNBV.

19. Agreements are not necessarily symmetrical (i.e., they could benefit only the customers of one of the two banks).

20. The Financial Reform of 2014 modified the Law of Transparency and Financial Services Ordering (LTOSF) to allow banks to enter in agreements to charge lower fees to customers of other banks (for financial services in general). Before the reform, banks were only allowed to charge lower fees to their own customers. Circular 15/2014 of the Bank of Mexico eliminated the requirement of constituting a third party to be able to enter into an ATM-sharing agreement.

3.3.1 Shift-Share Design

Because agreements occur between banks at the national level, a natural empirical strategy is a shift-share design—or "Bartik" instrument—that exploits the differential exposure of municipalities to these common agreements. In particular, municipalities that have a large presence of banks in an agreement will be more exposed relative to municipalities with a small presence of these banks.

Let w_{ijmt} denote the number of withdrawals using the card from bank i on ATM j in municipaliy m in period t (which corresponds to quarters in our data). Let w_{mt} be the total withdrawals in municipality m, so $w_{mt} = \sum_i \sum_j w_{ijmt}$. Start with the "accounting" equality:

$$d \ln w_{mt} = \sum_i \sum_j d \ln w_{ijmt} s^w_{ijmt} \tag{3.3}$$

where $s^w_{ijmt} \equiv w_{ijmt}/w_{mt}$ is the share of withdrawals on bank pair ij. As usual in this type of design, we fix the shares to an initial time period, s^w_{ijm0}. Given that the CNBV financial data set only has ATM withdrawal information at the bank-level and not at the bank-pair level, we replace the share of withdrawals from card i on ATM j with the inner product of the card share of i and the ATM withdrawal share of j. In other words, we approximate s^w_{ijm0} with $z_{ijm0} \equiv s^{card}_{im0} \times s^w_{jm0}$ where $s^{card}_{im0} = cards_{im0}/cards_{m0}$ and $s^w_{jm0} = w_{jm0}/w_{m0}$.[21]

The next step in a shift-share design is to decompose the bank-pair-location-period growth rate into a bank-pair-period growth rate (g_{ijt}) and an idiosyncratic bank-pair-location-period term ($\widetilde{g}_{ijmt}$):

$$d \ln w_{ijmt} = \underbrace{d \ln w_{ijt}}_{\equiv g_{ijt}} + \widetilde{g}_{ijmt}$$

We further instrument the bank-pair-period growth rate with agreement dummies E_{ijt}, that indicate whether there was an agreement in place between banks i, j in the previous

21. The variable $cards_{im0}$ denotes the number of card contracts issued by bank i in municipality m on period t.

period. We weight this variable by the percentage change in fee ($d \ln p_{ijt}$) associated with the agreement (i.e., the "size" of the agreement):[22]

$$d \ln w_{ijt} = \beta E_{ijt} d \ln p_{ijt} + \psi_{ijt}$$

Plugging back in Equation 3.3 we get the "first-stage" Bartik-type equation:

$$d \ln w_{mt} = \gamma^w \underbrace{\sum_i \sum_j E_{ijt} d \ln p_{ijt} z_{ijm0}}_{\text{Bartik instrument} \equiv B_{mt}} + \theta_m^w + \lambda_t^w + \zeta^w X_{mt} + \epsilon_{mt}^w \tag{3.4}$$

The Bartik instrument is the inner product of the agreement shocks, $E_{ijt} d \ln p_{ijt}$, and the exposure to these shocks, z_{ijm0}. θ_m^w and λ_t^w are municipality and time fixed-effects, respectively, and X_{mt} represents a set of municipality-specific time-varying controls.[23] We also estimate reduced-form regressions analogous to Equation 3.4 considering different outcomes y_{mt} such as crime, informality and taxation:

$$d \ln y_{mt} = \gamma B_{mt} + \theta_m + \lambda_t + \zeta X_{mt} + \epsilon_{mt} \tag{3.5}$$

where the coefficient of interest is γ in Equation 3.5 interpreted as the differential change in the growth rate of the outcome variable y_{mt} in municipalities with a higher exposure to the shock, relative to municipalities with a lower exposure. Identification, as in Goldsmith-Pinkham et al. (2020), relies in the exogeneity of the shares z_{ijm0} with respect to the error terms ϵ_{mt} after adding the controls and fixed effects. Because we exploit differential exposure

22. See the change in fees for each agreement in Table C.18. We only observe the change in fees for banks that enter into agreements. Results remain very similar if we do not weight agreements by the fee reductions. To help with interpretability, we use the absolute value of the percentage change in fees; i.e., an agreement with a 50% fee reduction will have $d \ln p_{ijt} = 0.5$.

23. Figure C.16 shows that there is no evidence of pretrends in the first stage. The Bartik instrument does not seem to have an effect on the growth rate of ATM withdrawals on periods before the agreements take place.

of municipalities to national-level agreements, identification in terms of shares is a more adequate assumption than identification coming from exogenous shocks (see, for instance, Borusyak et al., 2018).[24]

Table C.26 and Figure C.17 contain the diagnostics suggested by Goldsmith-Pinkham et al. (2020). They show that the Bartik estimator can be decomposed as a weighted sum of the coefficients of just-identified IV regressions, where each agreement share is used as an instrument.[25] Intuitively, agreements with larger weights (called "Rotemberg" weights) tend to drive the estimates. The top 5 agreements concentrate 34% of the positive weight in the estimator—slightly less than in other applications. Panel D shows that the overidentification test does not reject the null of exogenous instruments or no misspecification. Hence, the assumption of constant effects (across time and municipalities) is reasonable. Nevertheless, Figure C.17 shows substantial heterogeneity in the 2SLS estimates across agreements. As Goldsmith-Pinkham et al. (2020) point out in the context of heterogeneous effects, this suggests that some of the underlying effects could potentially receive negative weight. This affects the LATE-like interpretation of our estimate, suggesting that it need not be robust to heterogeneous effects.[26]

Our implementation considers only those municipalities where there are at least two different banks with ATMs and debit cards (such that they are potential candidates to be exposed to an ATM-sharing agreement). The data of ATM withdrawals is available since March 2011, so we use March, 2011, to December, 2012, as baseline period.[27]

24. In particular, note from Table C.18 that the agreements occur between a small subset of banks and no more than 45% of Banks signed an agreement.

25. We can decompose the Bartik instrument as a sum over ATM agreements (rather than bank pairs), $B_{mt} = \sum_k E_{kt} d \ln p_{kt} z_{km0}$, where k denotes one of the 56 national-level agreements. We can do this because the shocks E_{kt} are always zero for banks that do not enter into an agreement.

26. We thank our discussant Gabriel Chodorow-Reich for pointing this out. To the best of our knowledge, the literature has not yet developed an estimator robust to unrestricted heterogeneity in this context. The closest related work is de Chaisemartin and Lei (2021), who develop an estimator that is robust to location-specific effects in the context of random shocks, as opposed to the random shares approach that we adopt.

27. Note that we are using the lagged agreement dummies, E_{ijt}, to construct the agreement shocks in

Table 3.4 presents the results from our preferred specification, which includes municipality and quarter fixed effects as well as controls for income per capita, total employment, and total population. These regressions are weighted by population and standard errors are clustered at the municipality level. Column (1) shows the effect of the Bartik instrument on the number of ATM withdrawals. It shows that the growth rate of ATM withdrawals is positive and strongly significant. The reduction in ATM-fees resulted in more ATM transactions in municipalities more exposed to the shock. This estimate implies that bank-sharing agreements dropping ATM fees completely in a municipality more than doubles withdrawals. One standard deviation increase in the Bartik instrument, conditional on an agreement taking place in a municipality (mean 0.001, std. 0.007), changes ATM transactions positively by 1.2 percent. Column (2) shows the results when we use the number of debit card contracts as a dependent variable. We observe a positive response of debit cards to the shock, but the coefficient is not statistically significant.

In Columns (3)-(6) we use homicides, thefts/robberies, pedestrian theft, and total crime as dependent variables. We find that pedestrian theft, a street crime where the victim is a pedestrian and the offender attempts the theft of cash or other property, is negative and statistically significant. We do not find the effect of ATM agreements on homicides, total thefts/robberies, or total crime statistically significant. These results are consistent with an interpretation of ATM sharing agreements as arrangements that reduced the average cash holdings per individual, thus decreasing the possibility that cash related crimes take place. The last two columns show results using the total number of informal workers and the local tax collection as dependent variables. Consistent with the evidence presented in Section 3.2, we do not find that the shock affected informality and we find evidence that local tax collection decreased.

Equation 3.4. We consider alternative leads and lags in Figure C.16 in Section C.4 as a robustness check.

Table 3.4: Effects of ATM-Sharing Agreements

Note: The table reports the results for the coefficient γ after estimating Equation 3.5. Observations are at the municipality-quarterly level. The dependent variable in Columns (1) is the quarterly change in the logarithm of the total ATM withdrawal count. Column (2) uses the quarterly change in the logarithm of debit card contracts. Column (3) is the quarterly change in the logarithm of homicides (using data from INEGI based on death certificates). Column (4) is the quarterly change in the logarithm of total thefts/robberies. Column (5) is the quarterly change in the logarithm of total thefts to pedestrians. Column (6) is the quarterly change in the logarithm of total crimes. Column (7) is the quarterly change in the logarithm of informal workers. Column (8) is the quarterly change in the logarithm of local taxes. We use the inverse hyperbolic sine transformation in all cases. The controls we use include income per capita, total employment, and total population. Regressions are weighted by the total population in the municipality on the pre-period. Standard errors are clustered at the municipality level.

	(1) ATM Trans.	(2) Debit	(3) Homicide	(4) Thefts	(5) Ped. Theft	(6) Crimes	(7) Informal	(8) Taxes
Bartik	1.6770***	0.6504	-2.9215	-3.7118	-6.1498*	-3.8172	-1.6562	-2.8440**
	(0.600)	(1.093)	(3.138)	(3.426)	(3.454)	(2.497)	(1.268)	(1.385)
Obs.	20,695	20,695	20,710	20,710	20,710	20,710	20,710	3,822
Mun.	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Weights	Y	Y	Y	Y	Y	Y	Y	Y

3.3.2 Discussion

We interpret the implementation of ATM sharing agreements as a shock that dropped the price of ATM withdrawals and thus is likely to have decreased the average cash holdings of customers. This shock provides exogenous variation to the use of cash similar to the rollout of debit cards explored in Section 3.2. Indeed, we find similar results in these two quasi-natural experiments. First, in both experiments there is an increase in the number of ATM transactions. The number of debit cards increased in both cases. None of the experiments have a significant effect on the number of informal workers or on the total number of homicides. Both experiments have an effect on cash-related crimes such as theft.

Importantly, the fact that after 2011 the Mexican Criminal Incidence data report counts of theft across different categories allows us to explore the impact of ATM sharing agreements on different types of theft. Consistent with a reduction of cash in circulation, we find that pedestrian thefts decline in response to agreement shocks. In both quasi-natural experiments, we estimate a negative effect on local tax collection, which could imply additional costs of policies restricting the use of cash. In Section 3.4.6, in order to present conservative estimates of the social benefits of precluding the use of cash, we focus only on quantifying the benefits of reducing the prevalence of cash-related criminal activities. Furthermore, given that the empirical evidence in Section 3.2 and Section 3.3 is reduced form in the sense that the analysis does not measure directly the impact of the card rollout or ATM agreements on the level of cash holdings. To the best of our knowledge, there are no public data sets measuring cash holdings at the individual or regional level.[28] For this reason, in Section 3.4.6 we take a conservative approach and assume that restrictions on the use of cash eradicate all thefts and robberies.

3.4 Simple Cash-Credit Model for Welfare Analysis

In this section we present a simple model where utility comes from differentiated goods, which themselves are aggregates of the same good/service paid in cash or by other means of payment, which we denote by card. This is a reduced form, or indirect utility, which should capture how agents' choice of means of payment depend on the relative price. We first present a representative agent version of the model, and then a version with both "banked" and "unbanked" households, which is important for matching the intensive/extensive margins decisions observed in the data. Finally, we return to the evidence on cash and credit usage described above to calibrate the model and to quantify the costs of either a large tax on cash

28. The closest data set that exists is the account-level data from Bansefi used by Bachas, Gertler, Higgins, and Seira (2017), which shows that there is an increase in withdrawals after the rollout.

or an outright ban.

Our choice of a cash-credit model, along the lines of Lucas and Stokey (1987), provides a tractable framework that allows for a simple welfare analysis of the restrictions on cash usage. Related efforts to give an explicit account of the fundamental transactions choices, and hence of the fundamental nature of the welfare costs, such as Gomis-Porqueras et al. (2014); Alvarez and Lippi (2017); Wang et al. (2020); Deviatov and Wallace (2014), provide an interesting avenue for future research towards a deeper understanding of the costs and benefits of cash usage.

3.4.1 A Representative Household Model of Cash-Credit Choice

We assume an agent's utility over a set of $\mathcal{A}$ varieties of goods, indexed by α is given by

$$u\left(\{x_\alpha\}_{\alpha\in\mathcal{A}}\right) \equiv \left(\sum_{\alpha\in\mathcal{A}} \phi_\alpha^{\frac{1}{\sigma}} x_\alpha^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}} \tag{3.6}$$

where ϕ_α is a preference weight parameter and σ is the constant substitution elasticity between goods.

Following Lucas and Stokey (1987) we assume that each good variety α can be purchased using either cash or an alternative means of payment, which we refer to as credit. The quantity of the goods are denoted by a if paid by cash and c if paid by credit. Thus the quantity x_α is itself a composite of cash and credit purchases for that variety according to

$$x_\alpha = \left(\alpha^{\frac{1}{\eta}} c_\alpha^{\frac{\eta-1}{\eta}} + (1-\alpha)^{\frac{1}{\eta}} a_\alpha^{\frac{\eta-1}{\eta}}\right)^{\frac{\eta}{\eta-1}} \tag{3.7}$$

where η is the substitution elasticity between cash and credit goods and α is the name of the good, and also the preference weight parameter for the "good" with credit share α.

We view as a reasonable hypothesis to consider parameter values such that the substitu-

tion elasticity between physically goods/services is smaller than the substitution elasticity between cash and credit, i.e. $\sigma \leq \eta$, but of course the model could use any parameter. For instance, below we discuss a simple expression for a lower bound on the cost of a ban of using cash as a means of payment that holds when $\eta = \sigma$.

We select the units of the goods/services so that we can normalize each good's price in terms of the numeraire and consider the agent's budget constraint in the baseline prices as:

$$\sum_{\alpha \in \mathcal{A}} \Big(c_\alpha + a_\alpha (1+\tau) \Big) = y + \varrho \tag{3.8}$$

where y is the agent's income, τ represents a tax on cash purchases, and ϱ a transfer, that is used to rebate the taxes levied on cash purchases.

The ideal price index p_α for the cash-credit bundle of type α given the tax rate on cash purchases τ is the usual one implied by CES utility.[29] This is the minimum cost, in units of the numeraire (goods paid with credit), which yields a utility $x_\alpha = 1$. It is given by

$$p(\alpha;\tau) = \Big(\alpha + (1-\alpha)(1+\tau)^{1-\eta} \Big)^{\frac{1}{1-\eta}} \tag{3.9}$$

Aggregating across all goods varieties α yields the ideal price index $\mathcal{P}(\tau)$ for the agent's aggregate consumption as a function of τ:

$$\mathcal{P}(\tau) = \left(\sum_{\alpha \in \mathcal{A}} \phi_\alpha \, p(\alpha;\tau)^{1-\sigma} \right)^{\frac{1}{1-\sigma}} = \left(\sum_{\alpha \in \mathcal{A}} \phi_\alpha \Big(\alpha + (1-\alpha)(1+\tau)^{1-\eta} \Big)^{\frac{1-\sigma}{1-\eta}} \right)^{\frac{1}{1-\sigma}} \tag{3.10}$$

Let A denote the quantity of cash goods bought per unit of income, i.e. $A = \sum_{\alpha \in \mathcal{A}} a_\alpha / (y + \varrho)$. We can use A to compute the rebate to the agent of the taxes levied on cash payments

29. To see this just solve the dual problem of choosing a_α and c_α optimally to minimize the cost yielding one unit of utility.

as $\varrho = \tau(y+\varrho)A$, so that $\varrho = y\tau A/(1-\tau A)$. The quantity $A(1+\tau)$ is the share of expenditure paid with cash of the total income $y + \varrho$. Of course in equilibrium A is itself a function of τ, so that when useful we will write $A(\tau)$.

The parameters ϕ_α have the interpretation of the expenditure share on the goods with credit share α at baseline prices, i.e. when $\tau = 0$. Recall that in this baseline case units are chosen so that $p(\alpha; 0) = 1$ for all goods. Note that in this case the share of goods paid with cash is $A(0) = \sum_{\alpha \in \mathcal{A}} \phi_\alpha (1 - \alpha)$. We can, in principle, measure ϕ_α and α for different categories of goods and services using expenditure surveys, such as in Table 3.1.

We assume the tax on cash goods is rebated to the agents, so that the welfare cost of a tax on cash is measured by

$$W(\tau) \equiv \frac{y + \varrho}{\mathcal{P}(\tau)} = \frac{y}{\mathcal{P}(\tau)} \frac{1}{(1 - \tau A(\tau))} \tag{3.11}$$

where we use expression for ϱ derived above. The term $y + \varrho$ denotes the sources of income, given by income y and transfers $\varrho = y\tau A/(1 - \tau A)$. These resources are used by the agent to buy goods and pay taxes. If $\tau = 0$ then $\mathcal{P} = 1$ and $W(0)/y = 1$, which is the baseline level of welfare absent a tax on cash. If cash is taxed $\tau > 0$ then $\mathcal{P} > 1$, so that any given level of welfare is more expensive, and welfare decreases. The welfare cost must take into account the fact that the tax on cash is rebated to the households, an effect measured by the $A\tau$ term in the numerator of Equation 3.11. Taxing cash gives rise to a welfare cost as it distorts the agent's optimal choices. We will use Equation 3.11 to assess the cost of restrictions to cash usage, or a finite tax on cash ($\tau < \infty$), as well as a ban on cash modelled as an infinite tax on cash goods ($\tau \to \infty$).

To compute A we use the following equations that are readily derived from the agent's first order conditions for a CES utility function:

$$\frac{a_\alpha}{x_\alpha} = (1 - \alpha) \left(\frac{1 + \tau}{p_\alpha} \right)^{-\eta} \quad \text{and} \quad \frac{x_\alpha}{(y + \varrho)/\mathcal{P}} = \phi_\alpha \left(\frac{p_\alpha}{\mathcal{P}} \right)^{-\sigma} \tag{3.12}$$

where above we omit that the ideal price $\mathcal{P}$ and p depend on τ. We can then write the share of expenditures paid in cash:

$$(1+\tau)A(\tau) = (1+\tau)\frac{\sum_{\alpha\in\mathcal{A}} a_\alpha}{y+\varrho} = \left(\frac{1+\tau}{\mathcal{P}(\tau)}\right)^{1-\eta} \sum_{\alpha\in\mathcal{A}} (1-\alpha)\phi_\alpha \ \left(\frac{p_\alpha(\tau)}{\mathcal{P}(\tau)}\right)^{\eta-\sigma} \tag{3.13}$$

so that welfare as a function of τ is given by

$$W(\tau) \equiv \frac{y}{\mathcal{P}(\tau)} \left(\frac{1}{1 - \frac{\tau}{1+\tau}\left(\frac{1+\tau}{\mathcal{P}(\tau)}\right)^{1-\eta} \sum_{\alpha\in\mathcal{A}} (1-\alpha)\phi_\alpha \ \left(\frac{p_\alpha(\tau)}{\mathcal{P}(\tau)}\right)^{\eta-\sigma}} \right) \tag{3.14}$$

We refer to $-\log W(\tau)$ as the (private) welfare cost of taxing cash, expressed in log points. We call it private because it abstract from external effects such as crime, tax avoidance, etc.

3.4.2 Welfare Cost of Taxing Cash: Summary

This subsection summarizes both the theory and the calibration exercises developed below. We will argue below through a mixture of analytical results and numerical evaluation for calibrated parameter values that, while the model presented above allows for heterogeneity in the cash share across goods, welfare cost of a tax on cash $-\log W(\tau)$ is mainly determined by $\bar{\alpha}$ and η where $\bar{\alpha}$ is the average expenditure in credit

$$\bar{\alpha} = \sum_{\alpha\in\mathcal{A}} \phi_\alpha \alpha. \tag{3.15}$$

In other words, the level of σ and the distribution of expenditure shares across goods are quantitatively less relevant. For simplicity, we could then consider the case where $\mathcal{A}$ has only one good. In particular, $\phi_{\bar{\alpha}} = 1$, $u = x_{\bar{\alpha}}$, the budget constraints simplifies to $c_{\bar{\alpha}} + a_{\bar{\alpha}}(1+\tau) = y + \varrho$ and likewise the expressions for $A(\tau)$ and $W(\tau)$ shown in Equation 3.13 and Equation 3.14. Moreover, it can be shown that the welfare cost of taxing cash in the model with only

intensive margin of cash-credit choice is an lower bound to the welfare cost in a model that also includes an extensive margin choice of adopting credit for the unbanked.

We reach these conclusions through the analysis in the following subsections. In Section 3.4.3 we develop a series of analytical results for a ban on cash (e.g. $\tau \to \infty$), where we obtain an upper bound for $-\log W(\infty)$. We also develop a second order expansion around $\tau = 0$ for the case where there is either no heterogeneity or the case where $\sigma = \eta$. Section 3.4.4 explains how the model developed in this section can be used as a lower bound to evaluate the welfare cost of taxing cash including both an intensive margin choice across payment methods and also an extensive margin choice for unbanked agents. Section 3.4.5 reviews the estimates of the relevant parameters to quantify the welfare cost of a tax on cash.

Lastly, the reader interested in our main findings can jump to Figure 3.9 and Figure 3.10 at the end of Section 3.4.5. The right panel of Figure 3.9 shows the welfare cost of taxing cash for a large range of tax rates. The figure displays the results for different distributions of cash shares in the empirically plausible range and for different values of σ. For reasons detailed in Section 3.4.5, we focus our analysis in tax rates of approximately 40%. It is clear from these figures that the value of σ and the distribution of cash shares across goods are not quantitatively important relative to the other features. As a consequence, Figure 3.10 shows the welfare cost of taxing cash without heterogeneity. The left hand side panel explores a tax of 40% and the right hand side panel a ban on cash (e.g. $\tau \to \infty$). They display the welfare cost as a function of the two key parameters: the elasticity of substitution, η, and the average cash share, $\bar{\alpha}$. Our preferred estimates are $\eta = 5$ and $\bar{\alpha} = 0.40$ as discussed in Section 3.4.5, which are chosen to be both empirically relevant and conservative e.g. meaning to make the welfare cost of taxing cash small. Even for these conservative estimates, the welfare cost of taxing cash is large: 6.5% of GDP for a 40% tax and 10% of GDP for a ban on cash. A tax of 40% is equivalent to a budget-neutral policy of a tax rate to cash transactions of about

28% and a subsidy to credit transactions of about 8% (see Section C.1).

3.4.3 Analytical Results: Intensive Choice of Cash-Credit

In this subsection we derive results for both the case of a infinite tax on cash and the case for a finite tax on cash.

A ban on cash. We can obtain the private welfare cost of a ban on cash, by letting $\tau \to \infty$. Assume that $\eta > 1$ so that $A \to 0$ and hence cash is not used, then we have

$$\lim_{\tau\to\infty} W(\tau) = \lim_{\tau\to\infty} \frac{y}{\mathcal{P}(\tau)} = \frac{y}{\left(\sum_{\alpha\in\mathcal{A}} \phi_\alpha \alpha^{\frac{1-\sigma}{1-\eta}}\right)^{\frac{1}{1-\sigma}}}$$

Three remarks are in order.

1. If all goods have the same credit share, i.e. if $\phi_{\bar{\alpha}} = 1$, then we have

$$\mathcal{P}(\infty, \sigma, \eta) \equiv \lim_{\tau\to\infty} \mathcal{P}(\tau, \sigma, \eta) = \bar{\alpha}^{\frac{1}{1-\eta}}$$

   which is, trivially, independent of σ.

2. If $\sigma = \eta$ then,

$$\mathcal{P}(\infty, \sigma, \eta) = \bar{\alpha}^{\frac{1}{1-\eta}}$$

   where $\bar{\alpha} = \sum_{\alpha\in\mathcal{A}} \phi_\alpha \alpha$, which is the aggregate share of expenditure on credit at baseline prices.

3. The cost of a ban is decreasing in the elasticity of substitution between goods, i.e.

$$\mathcal{P}(\infty, \sigma', \eta) < \mathcal{P}(\infty, \sigma, \eta) \text{ for any two } \sigma' > \sigma \geq 1 \quad .$$

This is quite intuitive, as higher elasticity σ makes it easier for the agent to substitute to goods with smaller cash share. To see why this must hold, note that $\mathcal{P}(\infty)$ solves the same equation as the consumption equivalent for an agent with risky consumption $x = \alpha^{1/(1-\eta)}$, and a CRRA utility function i.e. $\mathcal{P}(\infty)^{1-\sigma}/(1-\sigma) = E[\frac{x^{1-\sigma}}{1-\sigma}]$. Thus, using the Arrow-Pratt Theorem, we obtain the desired results.

Under the assumption that $\sigma \leq \eta$ we can obtain a simple lower bound for the cost of a ban, namely

$$\mathcal{P}(\infty, \sigma, \eta) \geq \mathcal{P}(\infty, \eta, \eta) = \bar{\alpha}^{1/(1-\eta)}$$

where again $\bar{\alpha} = \sum_{\alpha \in \mathcal{A}} \phi_\alpha \alpha$ is the aggregate share of expenditure on credit at baseline prices. Note that in this case we have that the welfare cost gives:

$$-\log W(\infty) \equiv \log \mathcal{P}(\infty, \eta, \eta) = \frac{1}{1-\eta} \log \bar{\alpha}$$

This shows that the cost of the ban of cash is inversely proportional to the elasticity η, and that it is decreasing and convex in the share of credit $\bar{\alpha}$.

No heterogeneity or $\eta = \sigma$. Under the assumption that the substitution elasticity across varieties equals the elasticity across means of payments or equivalently the case where there is no heterogeneity (e.g. $\mathcal{A}$ has only one element), the formulas simplify and we can write:

$$\mathcal{P}(\tau) = \left(\sum_{\alpha \in \mathcal{A}} \phi_\alpha \left(\alpha_\alpha + (1 - \alpha_\alpha)(1+\tau)^{1-\eta} \right) \right)^{\frac{1}{1-\eta}} = \left(\bar{\alpha} + (1-\bar{\alpha})(1+\tau)^{1-\eta} \right)^{\frac{1}{1-\eta}}$$

where $\bar{\alpha} = \sum_{\alpha \in \mathcal{A}} \phi_\alpha \alpha$ is the baseline aggregate share of payments in credit. Computing welfare gives

$$W(\tau) = \frac{y}{\mathcal{P}(\tau)} \frac{1}{1 - \tau A(\tau)} = \frac{y}{\left(\bar{\alpha} + (1-\bar{\alpha})(1+\tau)^{1-\eta}\right)^{\frac{1}{1-\eta}}} \left(\frac{1}{1 - \frac{\tau}{1+\tau} \frac{(1+\tau)^{1-\eta}(1-\bar{\alpha})}{(\bar{\alpha}+(1-\bar{\alpha})(1+\tau)^{1-\eta})}} \right)$$

A few remarks are in order:

1. By the same argument used above, for any fixed tax τ, the ideal price index $\mathcal{P}(\tau)$ decreases with η.

2. It is clear from the expression above, that for $\bar{\alpha} < 1$ and a fixed $0 < \tau < \infty$, the share of cash $(1+\tau)A$ decreases with η. Hence $1/(1-\tau A) = 1/(1 - \frac{\tau}{1+\tau}(1+\tau)A)$ also decreases with η.

3. For small τ the welfare cost is increasing in η for $0 < \bar{\alpha} < 1$, i.e. a second order approximation around $\tau = 0$ gives the Harberger's triangle type expression:

$$-\log W(\tau) = \frac{1}{2}(1-\bar{\alpha})\,\bar{\alpha}\,\eta\,\tau^2 + o(\tau^2) \tag{3.16}$$

Hence, for a fixed $0 < \tau < \infty$ the welfare cost defined as $-\log W(\tau)$, is a non-monotone function of the elasticity of substitution η. Instead, as shown above, as $\tau \to \infty$ we have that the welfare cost is $-\log W(\infty) = \frac{-\log(\bar{\alpha})}{\eta - 1}$, and thus at very large τ the welfare cost is decreasing in η. This is to be compared with the second order approximation derived in Equation 3.16 which shows that for small τ, the welfare cost is increasing in η.

3.4.4 Intensive-Extensive Choice of Cash-Credit

Next we modify the representative agent setup described above to model the agent's choice to be unbanked, and thus be a cash only user, or to pay a fixed cost and access both cash

and credit services. The extension is motivated by the empirical observations that several households are unbanked and thus do not have a cash-credit choice at the moment. We derive a lower bound for the cost of a ban on cash, or a large tax, that has an expression identical to the ones derived for the representative agent model above. The key difference is in the interpretation, and hence calibration. We show that in this case one should use the fraction of credit purchases $\bar{\alpha}$ that corresponds to the currently banked households, i.e. what we identify as the mixed users in Mexico. This has the effect of reducing substantially the cost of a ban (or a large tax) on cash.

The outline of the model is as follows. We assume that in the baseline case there is a fraction $\beta \in [0, 1]$ of the population that have access to both payment methods (the "banked" population), and the remaining fraction $1 - \beta$, which we refer to as the "unbanked", whose only means of payment is cash. We assume that the utility function of both types is the same, and given by Equation 3.6 and Equation 3.7. The problem of the "banked" households is the one described above in the representative agent version. Instead, for the unbanked households we assume that if they pay a fixed cost $\psi > 0$, measured in the same units as utility w, then they gain access to both means of payment, and hence will face the same problem as the currently banked households. Below we describe in more detail the problem of the unbanked households. Finally, we give a formula for a lower bound on the cost of a (large) tax on cash τ that applies to the case where there is a fraction β of banked and $1 - \beta$ of unbanked households. The expressions for this lower bound are identical to the ones derived in the representative agent model. The difference, as mentioned above, is in the interpretation. Since, in the baseline situation, only the banked households use both means of payments, then we need to calibrate the model to their share of credit payments $\bar{\alpha}$, which is larger –inversely proportional to their share of expenditure in the population. This larger share of payments in credit, by using the expressions derived above, reduces the cost of a ban on cash since everyone is more predisposed to use credit.

We let $U(\tau)$ be the utility for the unbanked facing a tax on cash τ.

$$U(\tau) = \max\left\{\hat{W}(\tau), W(\tau) - \psi\right\} \tag{3.17}$$

where $\hat{W}(\tau)$ is the utility for the unbanked conditional on not adopting credit, while $W(\tau)$ is the utility of the household conditional on adopting credit analyzed in the previous section. We use ψ for the flow equivalent of the fixed cost that an unbanked agent has to pay to become banked and have access to credit as a means of payment. The utility $\hat{W}(\tau)$ solves

$$\hat{W}(\tau) = \max_{a_\alpha} \sum_{\alpha\in\mathcal{A}} \left[\hat{\phi}_\alpha^{1/\sigma}\, a_\alpha^{1-1/\sigma}\right]^{\frac{\sigma}{\sigma-1}} \text{ s. t. : } \sum_{\alpha\in\mathcal{A}} a_\alpha(1+\tau) = y + \varrho \tag{3.18}$$

$$\text{and where } \hat{\phi}_\alpha \equiv \phi_\alpha\,(1-\alpha)^{\frac{\sigma-1}{\eta-1}} \tag{3.19}$$

since the unbanked household problem is equivalent to one where we set $c_\alpha = 0$ for all α and hence

$$\hat{p}(\alpha;\tau) = (1+\tau)\,(1-\alpha)^{\frac{1}{1-\eta}} \text{ and } \hat{x}_\alpha = a_\alpha\,(1-\alpha)^{\frac{1}{\eta-1}} \tag{3.20}$$

In this section we assume that the banked households have already paid the cost of using credit, so we ignore these costs since they are sunk. For future reference we make two remarks.

1. For $\tau = \infty$ the unbanked households will choose to pay the cost ψ and have access to both means of payments. Likewise, for τ large, but finite, the unbanked households will still pay the cost.

2. Since in the baseline case with zero tax on cash and no rebate (i.e. $\tau = \varrho = 0$), the unbanked have chosen not to pay the fixed cost ψ, so that $U(0) = \hat{W}(0)$, then we can

obtain the following lower bound for its value:

$$\psi \geq \underline{\psi} \equiv W(0) - \hat{W}(0) \tag{3.21}$$

In words, $\underline{\psi}$ is the minimum fixed cost that will make the unbanked indifferent between using both means of payments or just cash at baseline prices (when $\tau = 0$).

Let's assume that τ is large enough so that the unbanked agents will pay the cost ψ and use both means of payments when facing the tax on cash τ. The difference in utility after and before the tax on cash $U(\tau) - U(0)$ for the unbanked is:

$$U(\tau) - U(0) = W(\tau) - \psi - U(0) \leq W(\tau) - \underline{\psi} - U(0) = W(\tau) - W(0) \tag{3.22}$$

where the first equality follows from the assumption on τ, the second inequality follows from $\psi \geq \underline{\psi}$, and the last equality from the definition of $\underline{\psi}$. Importantly, the last difference is exactly the welfare cost of a tax τ for the banked households, i.e for large enough τ we have:

$$\text{Welfare cost for unbanked} \equiv U(0) - U(\tau) \geq W(0) - W(\tau) \equiv \text{Welfare cost for banked} \tag{3.23}$$

This equation arises because the cost of being unbanked rises with τ, so for high enough τ the unbanked will pay the fixed cost. At that point, their utility cost has two components: (i) the fixed cost ψ that moves their net-of-fixed-cost welfare to be the same as of a banked agent with no tax, and (ii) the loss in welfare for a banked agent when the tax rises. Absent evidence on the fixed cost, Equation 3.23 justifies applying the welfare change for ex ante banked agents to all agents as a lower bound on the welfare change.

We briefly discuss the hypothesis we use to derive the inequality in Equation 3.23. To accomplish this analytically, consider the simple model where either α has a degenerate distribution, or where $\sigma = \eta$. Under these assumptions one can verify that for each triple

$(\eta, \bar{\alpha}, \tilde{\tau})$, where $\eta > 2$, $\bar{\alpha} \in (0,1)$, and $\tilde{\tau} > 0$, there is a non-empty interval of values ψ given by $[\underline{\psi}, \overline{\psi}]$, for which simultaneously: (i) it is optimal for the (unbanked) not to pay the fixed cost and remains using only cash when $\tau = 0$, and (ii) it is optimal to pay the fixed cost and used both means of payments when $\tau = \tilde{\tau}$. Using our functional forms, we show below that for a fixed $\bar{\alpha}$ and η, $\overline{\psi}$ is strictly increasing in $\tilde{\tau}$ provided that $\eta > 2$, a reasonable assumption given the empirical estimates of this parameter. The fact that $\overline{\psi}$ is increasing in $\tilde{\tau}$ means that a higher tax rate on cash gives larger incentives to adopt the banking technology.

To prove the existence of the interval of ψ and its characterization as a function of $\tilde{\tau}$, we consider the following inequality:

$$\underline{\psi} \equiv W(0) - \hat{W}(0) \leq \psi \leq W(\tilde{\tau}) - \hat{W}(\tilde{\tau}) \equiv \overline{\psi}$$

Verifying this inequality is simplified by using that the rebate ϱ for the case of $\tau > 0$ cancels in the right hand side, so, one can compare the (the reciprocal) of the corresponding ideal prices. Define $D(\tau) \equiv W(\tau) - \hat{W}(\tau)$. Then, under the stated assumptions, this simplifies to the following inequality:

$$D(0) = \frac{y}{1} - \frac{y}{(1-\bar{\alpha})^{\frac{1}{1-\eta}}} \leq D(\tilde{\tau}) = \frac{y}{\left[\bar{\alpha} + (1-\bar{\alpha})(1+\tilde{\tau})^{1-\eta}\right]^{\frac{1}{1-\eta}}} - \frac{y}{(1-\bar{\alpha})^{\frac{1}{1-\eta}}(1+\tilde{\tau})}$$

That $D(0) \leq D(\tilde{\tau})$ is immediate from the fact that $W(\tilde{\tau}) \geq \hat{W}(\tilde{\tau})$ since the latter is welfare obtained under the constraint that $c = 0$, an outcome that can always be replicated by the unconstrained agent. Moreover, using our functional forms and differentiating $D(\tilde{\tau})$ and using that $\bar{\alpha} \in (0,1)$, shows that the function $D(\tilde{\tau})$ is increasing in $\tilde{\tau}$ provided $\eta > 2$.[30]

30. Differentiating $D(\tau)$ gives, after simple algebra,

$$\frac{\partial D(\tau)}{\partial \tau} = y(1-\alpha)^{\frac{1}{\eta-1}}(1+\tau)^{-2}\left[1 - \left(1 + \frac{\alpha}{1-\alpha}(1+\tau)^{\eta-1}\right)^{\frac{2-\eta}{\eta-1}}\right]$$

Inspecting the expression reveals that the derivative is positive for every τ provided that $\eta > 2$.

3.4.5 Quantifying the Cost of Taxing Cash

Next we use the model sketched in Section 3.4.1 and Section 3.4.4 to gauge the welfare cost of taxing cash. The quantification is based on the Equation 3.14, where welfare is measured as a function of the tax on cash goods, τ, the distribution of credit shares $\{\phi_\alpha\}$, and the elasticities σ and η. We used the results of Section 3.4.4 to obtain a lower bound for the cost in the case of having both banked and unbanked agents. Our preferred estimate of a 40% tax on cash is a cost of approximately 6% of GDP or higher. We remark that the welfare cost is expressed in units of GDP, which makes it convenient to quantify the magnitude of the costs in terms of a compensating variation of GDP itself. Thus, the metric does not imply that taxing cash will lead to a change in the GDP level. As a matter of fact, conventionally measured GDP in our simple model is constant.[31]

The objective is to parameterize the model to replicate behavior observed in Mexico circa 2016, and use an elasticity of substitution between cash and credit expenditures, η, estimated in Alvarez and Argente (2020a) and Alvarez and Argente (2020b), to analyze several counterfactuals where cash expenditures are subject to a tax τ per unit of cash expenditure. Alvarez and Argente (2020a) estimates $\eta = 3$ using a large field experiment where riders where faced with different prices for Uber trips depending on whether the trips were paid with cash of with cards. Importantly, they find that a CES function summarizes well preferences between paying in cash or cards for price variation in the range of 40%. For this reason, below we consider a tax on cash $\tau = 0.40$ besides an outright ban of cash. Alvarez and Argente (2020b) study the ban on cash for Uber payments in the city of Puebla, Mexico. The changes in trips after the ban on cash for riders that before the ban have used cash and card with different intensity imply a long-run elasticity of substitution between 3 and 5. Thus, in this paper, we use the latter as our benchmark value and apply it to all

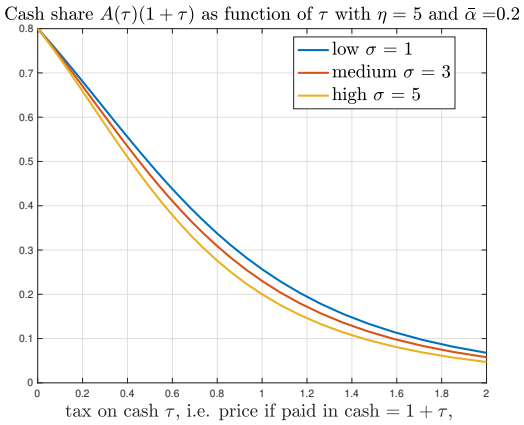
31. We thank our discussant Gabriel Chodorow-Reich for suggesting to us to clarify this potentially confusing point. Equation 3.14 gives the welfare-equivalent output reduction in the absence of any distortion in relative prices.

the goods in the economy. Several alternative parameterizations will be used to discuss the robustness of the findings.

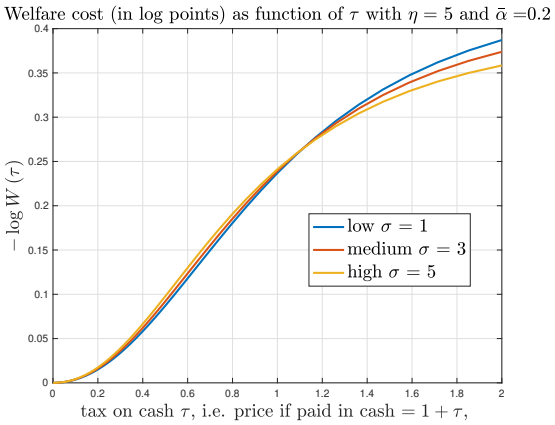
Figure 3.9: The private welfare cost as a function of the tax τ

Low credit share: $\bar{\alpha} = 0.20$

Cash share

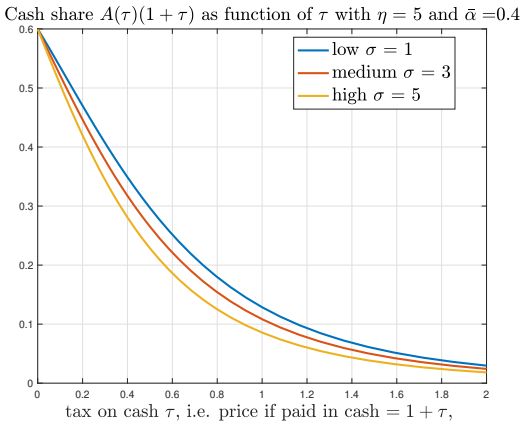


Welfare cost

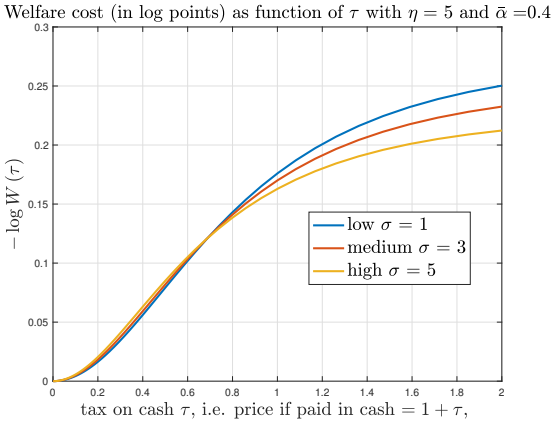


High credit share: $\bar{\alpha} = 0.40$

Cash share



Welfare cost



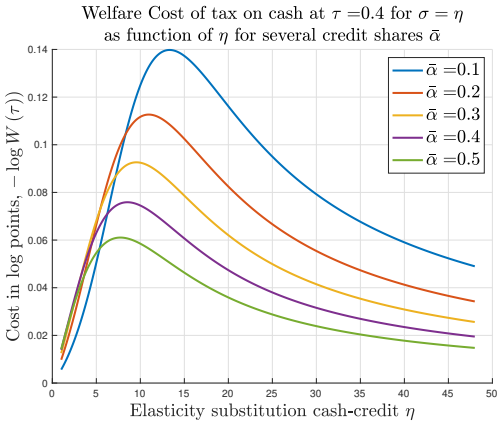
We start with a discussion of the distribution of the share of payments made in cash vs those made with other methods, which we refer to as credit. As mentioned in Section 3.1 cash is used extensively in Mexico. In Table 3.1 we use a consumption survey and for each tercile of the entire population ordered by expenditure we display the fraction of expenditure

for each broad category of goods, and for each category the fraction that is paid with card. Overall this gives a fraction of expenditures paid with card, denoted by $\bar{\alpha}$ in the model, just above 5%. In that section we also discussed that half of the population own a credit or debit card. Finally, in Figure 3.2 and Table 3.2, we display the share of consumption paid in card for those that have completed at least one purchase using cards during the survey. The expenditure of those households amount to about 21% of the total expenditure, which will correspond to $\beta = 0.21$ in the model of Section 3.4.4. The fraction of consumption paid in card for these mixed users is approximately $\bar{\alpha} \approx 0.25$. Based upon the results in Section 3.4.4, we can estimate a lower bound on the cost of a tax on cash τ for the entire population by using the expression in Equation 3.14 for the distribution with $\bar{\alpha} = 0.25$ or higher. We include results for larger values of $\bar{\alpha}$ since in the National Survey of Firms' Financing (ENAFIN) (Figure C.9) the fraction of revenues paid with card is as high as 18.5% – obtained as a weighted average of 25% for formal firms (75% of GDP) and assumed to be zero for informal firms. Some of the payments for firms are intra-firm transactions, hence this is an upper bound for households. To understand the robustness of the estimates, we report results using values of $\bar{\alpha}$ as high as 0.50, and to be conservative we use $\bar{\alpha} = 0.40$ as our benchmark.

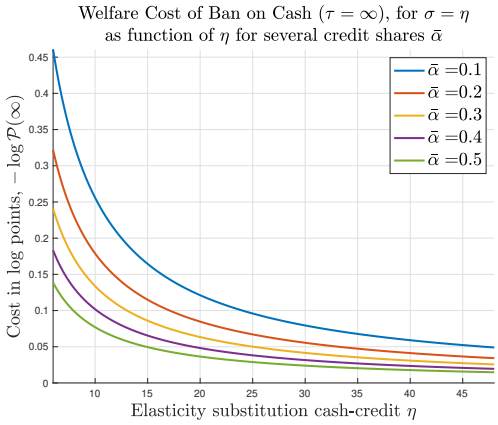
Figure 3.9 plots the cash share of expenditures $A(\tau)(1+\tau)$ and the welfare cost $-\log W(\tau)$ both as function of the tax on cash τ. In all the plots on this figure we use the distribution of the share of cash purchases for mixed users scaled up so that the average credit share $\bar{\alpha}$ is either 20% in the top panel or 40% in the bottom panel. In each plot we include three lines, corresponding to three values of σ. In each plot we let the tax on cash vary between 0 and 200%, or $\tau \in [0, 2]$. Figure 3.9 uses $\eta = 5$ in all plots. The welfare costs are quite insensitive to σ compared with the effect of other parameters. As explained in Section 3.4.1, the welfare costs are decreasing in the average credit share $\bar{\alpha}$, but they are still very high for $\bar{\alpha} = 0.4$. In view of the insensitivity with respect to σ, in Figure 3.10 we

Figure 3.10: The private welfare cost as a function of the elasticity η

A 40% tax on cash



A ban on cash ($\tau \to \infty$)



consider the case of $\sigma = \eta$, which gives a lower bound for the cost, and vary η over a large range of values. Figure 3.10 has two panels, one corresponding to $\tau = 0.4$, which is in the upper range of the experimental evidence, and the other panel for $\tau = \infty$, i.e. a ban on cash. In each case we plot the welfare cost for five different values of $\bar{\alpha}$. As explained in Section 3.4.1, we expect the welfare cost to be non-monotone for small τ. The non-monotonicity is clearly present in the left panel of Figure 3.10. For instance, for $\tau = 0.40$ and $\bar{\alpha} = 0.40$ the welfare cost for $\eta = 5$ is similar to the one for $\eta = 12$, and in both cases about 6.5%. For $\bar{\alpha} = 0.20$ the welfare cost corresponding to $\eta = 5$ is similar to the one for $\eta = 20$, and both of them are close to 8%. Instead the welfare cost of a ban on cash, i.e. the welfare cost of $\tau = \infty$ is about 10% for $\bar{\alpha} = 0.40$ and $\eta = 10$. Using much larger elasticities, say around $\eta = 30$ for $\bar{\alpha} = 0.40$, the welfare cost is smaller but still sizeable, say about 3% for both $\tau = 0.40$ and for $\bar{\tau} = \infty$. Summarizing, our preferred estimate of a 40% tax on cash is a cost of approximately 6% of GDP or higher.

3.4.6 Social Benefits of Curbing Cash Related Crimes

This section discusses some evidence on the social cost of crime with the goal to quantify the benefits that follow a reduction, or even the eradication, of criminal activities related to cash. As reported in Section 3.2 and Section 3.3, a reduction in the use of cash caused a statistically significant reduction in theft and robberies, while no significant change was detected in other categories such as violent crime or tax avoidance. For this reason, we focus on theft and robberies to quantify the benefits associated to a reduction in the use of cash. Furthermore, in order to provide an upper bound of the social benefits of policies restricting the use of cash, in what follows, we assume that all thefts and robberies related to cash are eradicated as a result of these policies. Measuring the incidence of these two crimes in Mexico between 2014 and 2016, and assessing the deadweight losses of such crimes, give an upper bound for the social benefits of eradicating both crimes between 0.48% and 1.28% of GDP.

Next, we illustrate the details of this computation, that is made of two steps. First we use aggregate statistics from National Survey on Victimization and Perception of Public Safety (ENVIPE) to measure the prevalence of cash-related theft and robberies in Mexico and quantify their magnitude.[32] This step yields an estimate of the direct cost of cash crimes, measured as the fraction of GDP that is stolen. We refer to this magnitude as the direct cost, because such a measure does not include the indirect costs triggered by crime, such as the preventive police cost, the judiciary costs, as well as the intangible costs associated to the crimes (psychological costs for the victims and other costs). Second, we quantify the indirect costs, i.e. the deadweight losses caused by theft and robberies, drawing from estimates of the cost of crime developed in economics of crime literature, such as Price (2000); Albertson and Fox (2008); Heeks et al. (2018), and the summary of the main estimates for the tangible

32. The years are ENVIPE 2017, ENVIPE 2016, and ENVIPE 2015, all from INEGI. Each survey reports data for the year before the one indicated in the title.

and intangible indirect costs of crime collected in the meta study by Wickramasekera et al. (2015). According to the Beckerian logic, the welfare loss of cash crimes is measured by the deadweight losses of such crimes, since the direct cost represents a transfer from one individual to another. We assume that the deadweight losses are proportional to the direct cost, therefore, both steps are needed to quantify the benefits of eradicating cash crimes.

Quantifying the direct cost of cash theft and robberies in Mexico. We quantify the direct cost due to robbery and theft associated with incidents where cash is stolen. Alternative measures can make this figure as high as 0.8% of GDP, or as low as 0.29% of GDP. This range informs us on the order of magnitude of the direct costs of crime. In short, using the number of incidents of cash related theft and robbery (about 13% per year) times the currency in circulation (about 6% of GDP) yields a loss of 0.8% of GDP per year. Alternatively, using the reported direct economic cost from the victimization survey, we get 0.29% of GDP per year.

We estimate these values in two different ways. The number of incidents comes from the ENVIPE victimization surveys (various years).[33] In particular, the first figure comes from averaging the rates from 2014 to 2016 of all crimes reported which are (42+35+37)/3 = 38 per 100 inhabitants. We take the fraction of those events that correspond to theft and robbery in the street or public transportation, plus robbery in other forms, plus other crimes, which include "express kidnapping". Note that we are excluding robbery at home, while in other categories we include crimes that may not be cash related. For 2016 these fractions are 25.9+5.1+3.4 = 34.4% (ENVIPE 2017), for 2015 they are 28.2+3.7+2.9=34.8% and for 2014 they are 28.6+3.5+3.0=35.1, so that the average fraction of incidents where cash is taken is about 35%. The product of the total crime incidence (38%) and the fraction of cash related crimes (35%) yields an average probability of cash theft of 13% per year. We apply this crime incidence to the stock of currency in the hands of the public between 2014 and

33. http://en.www.inegi.org.mx/programas/envipe/2019/

2016, which is about 6% of GDP, yielding a total cost of $0.13 \times 0.06 = 0.0079$ of GDP per person per year, or 0.8% of GDP per person per year.

The second estimate uses the victimization survey, which reports for all crimes an average loss of 1.27% of GDP for 2014, 1.25% for 2015 and 1.10% for 2016, or an average loss of 1.20% of GDP overall. Of these costs, the following fraction corresponds to the "economic losses as a consequences of these crimes" 68.3% for 2014, 62.9% for 2015 and 60.6% for 2016 or 63.9% on average. Multiplying these two averages and using that 38% of the crimes are thefts and robberies, which we associate with cash being stolen, we get $1.20 \times 0.639 \times 0.38$ or 0.29% of GDP per year in direct economic losses.

Quantifying the deadweight loss of cash related crimes. Most crimes involve tangible and intangible indirect costs. We rely on estimates of these deadweight losses developed in the economics of crime literature, such as the contributions surveyed in Wickramasekera et al. (2015). A synopsis of the various costs estimated by 14 different studies on the issue is given in Table 3 by Wickramasekera et al. (2015).[34] Several studies estimate the direct cost of the crime, and of the associated indirect tangible (police, medical assistance and judiciary) as well as of the indirect intangible costs (reflecting the fear, pain, suffering, and lost quality of life).[35]

There are 7 studies reporting the costs of robberies in the survey. These allow us to compute the deadweight loss per dollar stolen, measured by the ratio of the total social cost of the crime (including both tangible and intangible indirect costs) relative to the direct cost

34. These estimates are available mostly for developed countries (Australia, New Zealand, UK, USA).

35. As the authors explain "Indirect costs refer to the economic value of consequences of crime that do not involve a direct monetary exchange. These include lost productivity of both offenders and/or victims, and the value of volunteer time. Often lost productivity is estimated by calculating the forgone productivity as a result of the offence. For example, lost productivity can be determined by multiplying hourly average income with the number of hours a victim has spent out of work as a consequence of a crime. Intangible costs are costs incurred by victims, potential victims and society which include fear, pain, suffering, and lost quality of life. These costs are the most difficult to quantify as there is no market value or monetary exchange. As a result, intangible costs are usually inferred by revealed or stated preference-based methods such as willingness-to-pay or contingent valuation."

of the crime. For robberies, the average value of the deadweight loss per dollar stolen is 3.1, suggesting that every dollar robbed causes an additional 3.1 dollars of deadweight loss.[36]

There are also 7 studies concerning the cost of theft in the survey. For theft, the average value of the deadweight loss per dollar stolen is 1.1, suggesting that every dollar stolen causes an additional 1.1 dollars of deadweight loss. The deadweight loss is much smaller than the one for robberies because the lack of violence in thefts significantly reduces the indirect costs.[37]

We discussed above, in reporting the ENVIPE results, that the relative frequency of theft is about 25% while the frequency of robbery is about 8.5% of all crimes, so that their relative weight is 75% and 25% respectively. Using these weigths to combine the deadweight losses for theft (1.1 per dollar) and robbery (3.1 per dollar) with their relative frequency we obtain an average deadweight loss of about $0.75 \times 1.1 + 0.25 \times 3.1 = 1.6$ per dollar crime committed. Multiplying this average deadweight loss with the average cost of crime discussed above gives a lower bound of $1.6 \times 0.29 = 0.48\%$ of GDP and an upper bound of $1.6 \times 0.8 = 1.28\%$ of GDP.

3.5 Conclusion

Policies restricting the use of cash have recently received great interest and their possibility has been debated both by policymakers and academics. However, there are almost no

36. A recent report of the UK Home Office indicates substantively larger costs associated to robberies, with a deadweight loss close to a factor of 10 (of which 3.6 is due to physical and emotional harm and another 3.6 is due to judiciary costs), see Table 1 in Heeks et al. (2018). Including this study raises the average deadweight loss of robberies to a value of about 4 times the dollars stolen.

37. ENVIPE also includes a measure of the indirect costs of crime. They include health costs and the costs of preventive activities such as (e.g. changing doors and windows, changing doors' locks, installing fences, organizing joint activities with neighbors, acquiring a guard dog). The total of these indirect costs is in the order of 64% of the direct costs, much lower than our benchmark estimates. These estimates, however, do not include utility enhancing activities that were prevented by crime (such as going out at night, wearing personal valuable items, etc). Indeed, in ENVIPE, a large fraction of households report making adjustments to their daily activities because of crime. Unfortunately, we do not have an estimate of the monetary value of these costs.

attempts to quantify the welfare consequences of such policies accounting for both social benefits and private losses. In this paper, we attempt such a calculation for the case of Mexico. The social benefits of restricting the use of cash rely on estimates of the elasticity of crime and informality obtained from two quasi-natural experiments in Mexico that encouraged the use of debit cards. The private costs are estimated using a reduced form model and expenditure shares obtained from the Mexican expenditure survey. We find that the private costs of restricting the use of cash are at least twice as large as the social benefits.

Our calculation naturally relies on several assumptions that are necessary to make progress on the matter. We see the exploration of the robustness of these assumptions as a fertile ground for future research. First, our estimates for the private costs of restricting the use of cash heavily rely on the available estimates of the elasticity of substitution between cash and cards available in the literature. These estimates are calculated using experimental and observational data for Uber services in Mexico. More work is needed to compute an elasticity applicable to the entire economy.

Second, we do not consider the effect of the policy on tax evasion given that we do not find significant effects of the two policies we study on informality. Interestingly, no major effects on tax compliance are visible even in Lahiri (2020) analysis of the great demonetization that occurred in India in 2016.[38] It is possible, however, that there is an impact of cash on tax evasion and other crimes, especially if the size of the intervention is larger. A full ban on cash, for instance, could have an impact on crimes such as extortion.

Third, our calculation does not consider the general equilibrium effects of the policy. One is the restriction that the storability of cash has on the level of nominal interest rates, i.e. the zero lower bound. Another is the supply side response of alternative payment methods. These calculations could change the welfare effects of the policy estimated in this paper.

38. Cite from page 64 "The general picture that emerges from Figure 2 is that there has been some improvement in public finances in India since 2016, but it is difficult to attribute this to demonetization because the changes appear to be consistent with a prior trend. Hence, the indirect effect of demonetization on seizing undeclared income seems muted at best."

Lastly, our results could differ for countries with different levels of financial development. We believe that checking the robustness of our findings for different countries, as well as our estimated elasticities, are important areas for future research.

REFERENCES

Abadie, A. (2002). Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American statistical Association 97*(457), 284–292.

Acemoglu, D., A. Ozdaglar, and J. Siderius (2021). Misinformation: Strategic sharing, homophily, and endogenous echo chambers. Technical report, National Bureau of Economic Research.

Akerlof, G. and J. L. Yellen (1994). *Gang behavior, law enforcement, and community values*. Canadian Institute for Advanced Research Washington, DC.

Albert, A. B., H. E. Jacobs, and G. N. Siperstein (2016). Sticks, stones, and stigma: Student bystander behavior in response to hearing the word "retard". *Intellectual and developmental disabilities 54*(6), 391–401.

Albertson, K. and C. Fox (2008). Estimating the costs of crime in england and wales. *Safer Communities 7*(4), 25–33.

Algaba, A., D. Ardia, K. Bluteau, S. Borms, and K. Boudt (2020). Econometrics meets sentiment: An overview of methodology and applications. *Journal of Economic Surveys 34*(3), 512–547.

Ali, S., M. H. Saeed, E. Aldreabi, J. Blackburn, E. De Cristofaro, S. Zannettou, and G. Stringhini (2021). Understanding the effect of deplatforming on social networks. In *13th ACM Web Science Conference 2021*, pp. 187–195.

Allcott, H., L. Braghieri, S. Eichmeyer, and M. Gentzkow (2020). The welfare effects of social media. *American Economic Review 110*(3), 629–76.

Allcott, H. and J. B. Kessler (2019). The welfare effects of nudges: A case study of energy use social comparisons. *American Economic Journal: Applied Economics 11*(1), 236–76.

Alvarez, F. and D. Argente (2020a). Consumer surplus of alternative payment methods: Paying uber with cash.

Alvarez, F., D. Argente, R. Jimenez, and F. Lippi (2022). Cash: A blessing or a curse? *Journal of Monetary Economics 125*, 85–128.

Alvarez, F. and F. Lippi (2017). Cash burns: An inventory model with a cash-credit choice. *Journal of Monetary Economics 90*(C), 99–112.

Alvarez, F. E. and D. O. Argente (2020b). On the effects of the availability of means of payments: The case of uber.

Álvarez-Benjumea, A. and F. Winter (2018). Normative change and culture of hate: An experiment in online environments. *European Sociological Review 34*(3), 223–237.

Amazon (2020). Price gouging has no place in our stores.

Ambuehl, S., M. Niederle, and A. E. Roth (2015). More money, more problems? can high pay be coercive and repugnant? *American Economic Review 105*(5), 357–60.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström (2006). Elicitation using multiple price list formats. *Experimental Economics 9*(4), 383–405.

Anderson, E. T. and D. I. Simester (2010). Price stickiness and customer antagonism. *The quarterly journal of economics 125*(2), 729–765.

Anti-Defamation League (2021). Online hate and harassment. the american experience 2021. *Center for Technology and Society*. Accessed: 2021-10-23.

Arango, A., J. Pérez, and B. Poblete (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pp. 45–54.

Ba, B. (2018). Going the extra mile: The cost of complaint filing, accountability, and law enforcement outcomes in chicago. Technical report, Working paper.

Bachas, P., P. Gertler, S. Higgins, and E. Seira (2017, March). How debit cards enable the poor to save more. Working Paper 23252, National Bureau of Economic Research.

Balafoutas, L. and N. Nikiforakis (2012). Norm enforcement in the city: A natural field experiment. *European Economic Review 56*(8), 1773–1785.

Balafoutas, L., N. Nikiforakis, and B. Rockenbach (2016). Altruistic punishment does not increase with the severity of norm violations in the field. *Nature communications 7*(1), 1–6.

Barzel, Y. (1974). A theory of rationing by waiting. *The Journal of Law and Economics 17*(1), 73–95.

Beatty, T., G. Lade, and J. Shimshack (2020). Hurricanes and gasoline price gouging.

Becker, G. S. (1965). A theory of the allocation of time. *The economic journal*, 493–517.

Becker, G. S. and K. M. Murphy (1993). A simple theory of advertising as a good or bad. *The Quarterly Journal of Economics 108*(4), 941–964.

Bellemare, M. F. and C. J. Wichman (2020). Elasticities and the inverse hyperbolic sine transformation. *Oxford Bulletin of Economics and Statistics 82*(1), 50–61.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

Berry, J., G. Fischer, and R. Guiteras (2020). Eliciting and utilizing willingness to pay: Evidence from field trials in northern ghana. *Journal of Political Economy 128*(4), 1436–1473.

Berry, S., A. Gandhi, and P. Haile (2013). Connected substitutes and invertibility of demand. *Econometrica 81*(5), 2087–2111.

Bianchi, C. (2014). Slurs and appropriation: An echoic account. *Journal of Pragmatics 66*, 35–44.

Borusyak, K., P. Hull, and X. Jaravel (2018). Quasi-experimental shift-share research designs. Technical report, National Bureau of Economic Research.

Borusyak, K., X. Jaravel, and J. Spiess (2020). Revisiting event study designs: Robust and efficient estimation. Technical report, Working Paper.

Bottan, N. L. and R. Perez-Truglia (2017). Choosing your pond: Revealed-preference estimates of relative income concerns. *Available at SSRN 2944427*.

Bradford, B., F. Grisel, T. L. Meares, E. Owens, B. L. Pineda, J. N. Shapiro, T. R. Tyler, and D. E. Peterman (2019). Report of the facebook data transparency advisory group. *Yale Justice Collaboratory*.

Braghieri, L., R. Levy, and A. Makarin (2021). Social media and mental health. *Available at SSRN*.

Briglevics, T. and S. Schuh (2020). This is"what's in your wallet"... and here's how you use it. Disscussion Paper, West Virginia university.

Bronfenbrenner, M. (1947). Price control under imperfect competition. *The American Economic Review 37*(1), 107–120.

Bundesamt für Justiz (2019). Federal office of justice issues fine against facebook. `https://www.bundesjustizamt.de/DE/Presse/Archiv/2019/20190702_EN.html`. Accessed: 2021-09-30.

Bundesbank, D. (2017). War on cash: Is there a future for cash? various authors, International cash conference.

Bursztyn, L., G. Egorov, R. Enikolopov, and M. Petrova (2019). Social media and xenophobia: evidence from russia. Technical report, National Bureau of Economic Research.

Bursztyn, L., A. L. González, and D. Yanagizawa-Drott (2020). Misperceived social norms: Women working outside the home in saudi arabia. *American economic review 110*(10), 2997–3029.

Bursztyn, L., I. K. Haaland, A. Rao, and C. P. Roth (2020). Disguising prejudice: Popular rationales as excuses for intolerant expression. Technical report, National Bureau of Economic Research.

Bursztyn, L. and D. Y. Yang (2021). Misperceptions about others. Technical report, National Bureau of Economic Research.

Cabral, L. and L. Xu (2020). Seller reputation and price gouging: Evidence from the covid-19 pandemic. *Mimeo, April*.

Carlsmith, K. M., J. M. Darley, and P. H. Robinson (2002). Why do we punish? deterrence and just deserts as motives for punishment. *Journal of personality and social psychology 83*(2), 284.

Carlson, C. R. (2021). *Hate Speech*. MIT Press.

Carlson, C. R. and H. Rousselle (2020). Report and repeat: Investigating facebook's hate speech removal process. *First Monday*.

Cavallo, A., E. Cavallo, and R. Rigobon (2014). Prices and supply disruptions during natural disasters. *Review of Income and Wealth 60*, S449–S471.

Center for Countering Digital Hate (2021). Failure to protect: how tech giants fail to act on user reports of antisemitism. Technical report.

Chakraborti, R. and G. Roberts (2020a). Anti-price gouging laws, shortages, and covid-19: Big data insights from consumer searches. *SSRN Working Paper*.

Chakraborti, R. and G. Roberts (2020b). Learning to hoard: the effects of preexisting and surprise price-gouging regulation during the covid-19 pandemic. *SSRN Working Paper*.

Chandler, J., C. Rosenzweig, A. J. Moss, J. Robinson, and L. Litman (2019). Online panels in social science research: Expanding sampling methods beyond mechanical turk. *Behavior research methods 51*(5), 2022–2038.

Chandrasekharan, E., U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert (2017). You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction 1*(CSCW), 1–22.

Chen, Y. and D. Y. Yang (2019). The impact of media censorship: 1984 or brave new world? *American Economic Review 109*(6), 2294–2332.

Chodorow-Reich, G., G. Gopinath, P. Mishra, and A. Narayanan (2018, December). Cash and the economy: Evidence from india's demonetization. Working Paper 25370, National Bureau of Economic Research.

Chowdhury, F. A., L. Allen, M. Yousuf, and A. Mueen (2020). On twitter purge: A retrospective analysis of suspended users. In *Companion Proceedings of the Web Conference 2020*, pp. 371–378.

Clemens, M. A. (2018). Testing for repugnance in economic transactions: Evidence from guest work in the gulf. *The Journal of Legal Studies 47*(S1), S5–S44.

Coppock, A. and O. A. McClellan (2019). Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics 6*(1), 2053168018822174.

Correia-da Silva, J., B. Jullien, Y. Lefouili, and J. Pinho (2019). Horizontal mergers between multisided platforms: Insights from cournot competition. *Journal of Economics & Management Strategy 28*(1), 109–124.

Cowgill, B. and C. E. Tucker (2019). Economics, fairness and algorithmic bias. *preparation for: Journal of Economic Perspectives*.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition 108*(2), 353–380.

Davidson, T., D. Warmsley, M. Macy, and I. Weber (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, Volume 11.

Dávila Lárraga, L. G. (2016). How does prospera work?: Best practices in the implementation of conditional cash transfer programs in latin america and the caribbean. *Inter-American Development Bank*.

De Chaisemartin, C. and X. D'Haultfœuille (2020b). Difference-in-differences estimators of intertemporal treatment effects. *Available at SSRN 3731856*.

de Chaisemartin, C. and Z. Lei (2021). Are bartik regressions always robust to heterogeneous treatment effects? *Available at SSRN 3802200*.

de Quidt, J., L. Vesterlund, and A. Wilson (2019). Experimenter demand effects. In A. Ule and A. Schram (Eds.), *Handbook of Research Methods and Applications in Experimental Economics*, Chapter 20, pp. 384–400. Cheltenham, UK: Edward Elgar Publishing.

Deviatov, A. and N. Wallace (2014, April). Optimal inflation in a model of inside money. *Review of Economic Dynamics 17*(2), 287–293.

Dhakal, V., A. M. Feit, P. O. Kristensson, and A. Oulasvirta (2018). Observations on typing from 136 million keystrokes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–12.

Di Tella, R. and J. Dubra (2014). Anger and regulation. *The Scandinavian Journal of Economics 116*(3), 734–765.

Dixon, L., J. Li, J. Sorensen, N. Thain, and L. Vasserman (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73.

Dworczak, P., S. D. Kominers, and M. Akbarpour (2019). Redistribution through markets. *Becker Friedman Institute for Research in Economics Working Paper* (2018-16).

Dwyer, C. and A. Aubrey (2020, Apr). Cdc now recommends americans consider wearing cloth face coverings in public.

Elliott, L. J., M. Ljubijanac, and D. Wieczorek (2019). The effect of screen size on reading speed: A comparison of three screens to print. In *International Conference on Applied Human Factors and Ergonomics*, pp. 103–109. Springer.

ElSherief, M., V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Volume 12.

Elías, J. J., N. Lacetera, and M. Macis (2019, August). Paying for Kidneys? A Randomized Survey and Choice Experiment. *American Economic Review 109*(8), 2855–88.

Enikolopov, R., A. Makarin, and M. Petrova (2020). Social media and protest participation: Evidence from russia. *Econometrica 88*(4), 1479–1514.

Eyster, E., K. Madarász, and P. Michaillat (2021). Pricing under fairness concerns. *Journal of the European Economic Association*.

Facebook (2021). Community standards enforcement report, second quarter 2021. `https://about.fb.com/news/2021/08/community-standards-enforcement-report-q2-2021/`. Accessed: 2021-10-23.

Fehr, E. and U. Fischbacher (2004). Third-party punishment and social norms. *Evolution and human behavior 25*(2), 63–87.

Fehr, E., U. Fischbacher, and S. Gächter (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human nature 13*(1), 1–25.

Fehr, E. and S. Gächter (2002). Altruistic punishment in humans. *Nature 415*(6868), 137–140.

Fortuna, P. and S. Nunes (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR) 51*(4), 1–30.

Fortuna, P., J. Soler-Company, and L. Wanner (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management 58*(3), 102524.

Founta, A. M., C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Fujiwara, T., K. Müller, and C. Schwarz (2021). The effect of social media on elections: Evidence from the united states. Technical report, National Bureau of Economic Research.

Gadde, V. and K. Beykpour (2018). Setting the record straight on shadow banning. https://blog.twitter.com/en_us/topics/company/2018/Setting-the-record-straight-on-shadow-banning. Accessed: 2021-10-13.

Gandelman, N., I. Munyo, and E. Schertz (2019). Cash and crime. Mimeo, Universidad ORT Uruguay.

Gentzkow, M. (2007). Valuing new goods in a model with complementarity: Online newspapers. *American Economic Review 97*(3), 713–744.

Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as data. *Journal of Economic Literature 57*(3), 535–74.

Giansiracusa, N. (2021). Facebook uses deceptive math to hide its hate speech problem. *Wired*.

Gillespie, T. (2018). *Custodians of the Internet*. Yale University Press.

Ginther, M. R., R. J. Bonnie, M. B. Hoffman, F. X. Shen, K. W. Simons, O. D. Jones, and R. Marois (2016). Parsing the behavioral and brain mechanisms of third-party punishment. *Journal of Neuroscience 36*(36), 9420–9434.

Giosa, P. (2020). Exploitative pricing in the time of coronavirus—the response of eu competition law and the prospect of price regulation. *Journal of European Competition Law & Practice 11*(9), 499–508.

Goldsmith-Pinkham, P., I. Sorkin, and H. Swift (2020). Bartik instruments: What, when, why, and how. *American Economic Review 110*(8), 2586–2624.

Gomis-Porqueras, P., A. Peralta-Alva, and C. Waller (2014). The shadow economy as an equilibrium outcome. *Journal of Economic Dynamics and Control 41*(C), 1–19.

Gonzalez, R. and G. Wu (1999). On the shape of the probability weighting function. *Cognitive Psychology 38*(1), 129 – 166.

Guhl, J. and J. Davey (2020). Hosting the 'holohoax': A snapshot of holocaust denial across social media. *The Institute for Strategic Dialogue*.

Haaland, I., C. Roth, and J. Wohlfart (2020). Designing information provision experiments.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis 20*(1), 25–46.

Han, X. and Y. Tsvetkov (2020). Fortifying toxic speech detectors against veiled toxicity. *arXiv preprint arXiv:2010.03154*.

Harrison, G. (1992). Theory and misbehavior of first-price auctions: Reply. *American Economic Review 82*(5), 1426–43.

Harrison, G. W., M. I. Lau, E. E. Rutström, and M. B. Sullivan (2005). Eliciting risk and time preferences using field experiments: Some methodological issues. In *Field experiments in economics*. Emerald Group Publishing Limited.

Harrison, G. W. and J. A. List (2004). Field experiments. *Journal of Economic literature 42*(4), 1009–1055.

He, H. and L. Harris (2020). The impact of covid-19 pandemic on corporate social responsibility and marketing philosophy. *Journal of Business Research 116*, 176–182.

He, K., X. Zhang, S. Ren, and J. Sun (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Heeks, M., S. Reed, M. Tafsiri, and S. Prince (2018). The economic and social costs of crime. Research Report 99, Home Office.

Henrich, J., R. McElreath, A. Barr, J. Ensminger, C. Barrett, A. Bolyanatz, J. C. Cardenas, M. Gurven, E. Gwako, N. Henrich, et al. (2006). Costly punishment across human societies. *Science 312*(5781), 1767–1770.

Herrmann, B., C. Thöni, and S. Gächter (2008). Antisocial punishment across societies. *Science 319*(5868), 1362–1367.

Higgins, S. (2019). Financial technology adoption. *JMP Berkeley*.

Hobbs, W. R. and M. E. Roberts (2018). How sudden censorship can increase access to information. *American Political Science Review 112*(3), 621–636.

Holshue, M. L., C. DeBolt, S. Lindquist, K. H. Lofy, J. Wiesman, H. Bruce, C. Spitters, K. Ericson, S. Wilkerson, A. Tural, G. Diaz, A. Cohn, L. Fox, A. Patel, S. I. Gerber, L. Kim, S. Tong, X. Lu, S. Lindstrom, M. A. Pallansch, W. C. Weldon, H. M. Biggs, T. M. Uyeki, and S. K. Pillai (2020). First case of 2019 novel coronavirus in the united states. *New England Journal of Medicine 382*(10), 929–936. PMID: 32004427.

Hondroyiannis, G. and D. Papaoikonomou (2017). The effect of card payments on vat revenue: New evidence from greece. *Economics Letters 157*, 17–20.

Hotz, V. J., G. W. Imbens, and J. H. Mortimer (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics 125*(1-2), 241–270.

Huang, J., D. Reiley, and N. Riabov (2018). Measuring consumer sensitivity to audio advertising: A field experiment on Pandora internet radio. *Available at SSRN 3166676*.

Hunt, M. G., R. Marx, C. Lipson, and J. Young (2018). No more fomo: Limiting social media decreases loneliness and depression. *Journal of Social and Clinical Psychology 37*(10), 751–768.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jhaver, S., C. Boylston, D. Yang, and A. Bruckman (2021). Evaluating the effectiveness of deplatforming as a moderation strategy on twitter.

Jourová, V. (2016). Code of conduct on countering illegal hate speech online: First results on implementation. Technical report, European Commission, Directorate-General for Justice and Consumers.

Kahneman, D., J. L. Knetsch, and R. Thaler (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *The American economic review*, 728–741.

Kamerow, D. (2020). Covid-19: the crisis of personal protective equipment in the us. *BMJ 369*.

Kaye, D. (2019). *Speech Police: The Global Struggle to Govern the Internet*. Columbia Global Reports.

Kennedy, C. J., G. Bacon, A. Sahn, and C. von Vacano (2020). Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Kuziemko, I., M. I. Norton, E. Saez, and S. Stantcheva (2015, April). How elastic are preferences for redistribution? evidence from randomized survey experiments. *American Economic Review 105*(4), 1478–1508.

Lahiri, A. (2020, February). The great indian demonetization. *Journal of Economic Perspectives 34*(1), 55–74.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review 111*(3), 831–70.

List, J. A. (2020). Non est disputandum de generalizability? a glimpse into the external validity trial. Technical report, National Bureau of Economic Research.

Liu, Y., P. Yildirim, and Z. J. Zhang (2021). Implications of revenue models and technology for content moderation strategies. *Implications of Revenue Models and Technology for Content Moderation Strategies (November 23, 2021)*.

Lucas, Robert E, J. and N. L. Stokey (1987). Money and interest in a cash-in-advance economy. *Econometrica 55*(3), 491–513.

MacIntyre, C. R., H. Seale, T. C. Dung, N. T. Hien, P. T. Nga, A. A. Chughtai, B. Rahman, D. E. Dwyer, and Q. Wang (2015). A cluster randomised trial of cloth masks compared with medical masks in healthcare workers. *BMJ open 5*(4), e006577.

Manski, C. F. (1989). Schooling as experimentation: a reappraisal of the postsecondary dropout phenomenon. *Economics of Education review 8*(4), 305–312.

Marcelo, D., A. Raina, and S. Rawat (2020, 01). Private sector participation in disaster recovery and mitigation [disaster recovery guidance series by the global facility for disaster reduction and recovery].

Matias, J., A. Johnson, W. E. Boesel, B. Keegan, J. Friedman, and C. DeTar (2015). Reporting, reviewing, and responding to harassment on twitter. *Available at SSRN 2602018*.

Melendez, S. (2020). Twitter automatically flags more than half of all tweets that violate its rules. *Fast Company*. Accessed: 2021-10-13.

Merrer, E. L., B. Morgan, and G. Trédan (2020). Setting the record straighter on shadow banning. *arXiv preprint arXiv:2012.05101*.

Montiel Olea, J. L. and M. Plagborg-Møller (2019). Simultaneous confidence bands: Theory, implementation, and an application to svars. *Journal of Applied Econometrics 34*(1), 1–17.

Mosquera, R., M. Odunowo, T. McNamara, X. Guo, and R. Petrie (2020). The economic effects of facebook. *Experimental Economics 23*(2), 575–602.

Moss, A. J., C. Rosenzweig, J. Robinson, and L. Litman (2020). Is it ethical to use mechanical turk for behavioral research? relevant data from a representative survey of mturk participants and wages.

Müller, K. and C. Schwarz (2020a). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*.

Müller, K. and C. Schwarz (2020b). From hashtag to hate crime: Twitter and anti-minority sentiment. *Available at SSRN 3149103*.

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior 39*(3), 629–649.

Munger, K. (2021). Don't@ me: Experimentally reducing partisan incivility on twitter. *Journal of Experimental Political Science 8*(2), 102–116.

Naldi, M. (2019). A review of sentiment computation methods with r packages. *arXiv preprint arXiv:1901.08319*.

Palin, K., A. M. Feit, S. Kim, P. O. Kristensson, and A. Oulasvirta (2019). How do people type on mobile devices? observations from a study with 37,000 volunteers. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, pp. 1–12.

Pedersen, E. J., W. H. McAuliffe, and M. E. McCullough (2018). The unresponsive avenger: More evidence that disinterested third parties do not punish altruistically. *Journal of Experimental Psychology: General 147*(4), 514.

Pew Research Center (2021). Social media use in 2021. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/. Accessed: 2021-10-24.

Pigou, A. C. (1920). *The economics of welfare / by A. C. Pigou*. Macmillan London.

Price, R. (2000). The Economic and Social Costs of Crime. Research Study 217, Home Office.

Rauchfleisch, A. and J. Kaiser (2021). Deplatforming the far-right: An analysis of YouTube and BitChute. *Available at SSRN*.

Relia, K., Z. Li, S. H. Cook, and R. Chunara (2019). Race, ethnicity and national origin-based discrimination in social media and hate crimes across 100 us cities. In *Proceedings of the International AAAI Conference on Web and Social Media*, Volume 13, pp. 417–427.

Rey, J. D. (2020, Apr). Amazon is banning the sale of n95 and surgical masks to the general public. *Vox.com*.

Reynders, D. (2020). Countering illegal hate speech online: 5th evaluation of the code of conduct. Technical report, European Commission, Directorate-General for Justice and Consumers.

Roberts, M. and M. E. Roberts (2018). *Censored*. Princeton University Press.

Rotemberg, J. J. (2005). Customer anger at price increases, changes in the frequency of price adjustment and monetary policy. *Journal of monetary economics 52*(4), 829–852.

Rotemberg, J. J. (2008). Behavioral aspects of price setting, and their policy implications. Technical report, National Bureau of Economic Research.

Rotemberg, J. J. (2011). Fair pricing. *Journal of the European Economic Association 9*(5), 952–981.

Roth, A. E. (2007). Repugnance as a constraint on markets. *Journal of Economic perspectives 21*(3), 37–58.

Roth, A. E. (2015). *Who gets what—and why: The new economics of matchmaking and market design*. Houghton Mifflin Harcourt.

Roth, J. and P. H. Sant'Anna (2021). Efficient estimation for staggered rollout designs. *arXiv preprint arXiv:2102.01291*.

Sands, P. (2017). The dark side of cash – facilitating crime and impeding monetary policy. In I. C. Conference (Ed.), *War on Cash: Is there a Future for Cash?*, pp. 22–43. Bundesbank.

Schneider, F. (2017). Restricting or abolishing cash: an effective instrument for fighting the shadow economy, crime and terrorism? In I. C. Conference (Ed.), *War on Cash: Is there a Future for Cash?*, pp. 44–91. Bundesbank.

Seyler, D., S. Tan, D. Li, J. Zhang, and P. Li (2021). Textual analysis and timely detection of suspended social media accounts.

Sherry, M. (2019). *Disability Hate Speech: Social, Cultural and Political Contexts*, Chapter Disablist hate speech online. Routledge.

Siegel, A. A. and V. Badaan (2020). # no2sectarianism: Experimental approaches to reducing sectarian hate speech online. *American Political Science Review 114*(3), 837–855.

Spence, A. M. (1975). Monopoly, quality, and regulation. *The Bell Journal of Economics*, 417–429.

Stantcheva, S. (2020). Understanding economic policies: What do people know and how can they learn? Technical report, mimeo, presentation slides online at https://scholar. harvard. edu ....

Strossen, N. (2018). *Hate: Why we should resist it with free speech, not censorship*. Oxford University Press.

Stüber, R. (2021). Why high incentives cause repugnance: A framed field experiment. *Available at SSRN 3850618*.

Sullivan, C. D. (2021). Eliciting preferences over life and death: Experimental evidence from organ transplantation.

Tan, G. and J. Zhou (2021). The effects of competition and entry in multi-sided markets. *The Review of Economic Studies 88*(2), 1002–1030.

Tobin, J. (1970). On limiting the domain of inequality. *The Journal of Law and Economics 13*(2), 263–277.

Twitter (2018). Serving healthy conversation. `https://blog.twitter.com/official/en_us/topics/product/2018/Serving_Healthy_Conversation.html`. Accessed: 2021-10-19.

Twitter (2020a). Q4 2020 letter to shareholders. `https://s22.q4cdn.com/826641620/files/doc_financials/2020/q4/FINAL-Q4'20-TWTR-Shareholder-Letter.pdf`. Accessed: 2021-10-13.

Twitter (2020b). Rules enforcement report. `https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jul-dec`. Accessed: 2021-10-13.

Twitter (2021a). Debunking twitter myths. `https://help.twitter.com/en/using-twitter/debunking-twitter-myths`. Accessed: 2021-10-13.

Twitter (2021b). Hateful conduct policy. `https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy`. Accessed: 2021-06-20.

Twitter (2021c). Notices on twitter and what they mean. https://help.twitter.com/en/rules-and-policies/notices-on-twitter. Accessed: 2021-10-13.

Twitter (2021d). Our range of enforcement options. https://help.twitter.com/en/rules-and-policies/enforcement-options. Accessed: 2021-06-21.

Twitter (2021e). Report abusive behavior. https://help.twitter.com/en/safety-and-security/report-abusive-behavior. Accessed: 2021-06-21.

Twitter (2021f). Response codes. https://developer.twitter.com/ja/docs/basics/response-codes. Accessed: 2021-06-21.

Uber (2016). Uber y la cdmx acuerdan tarifas durante contingencias. https://www.uber.com/es-MX/blog/mexico-city/uber-y-la-cdmx-acuerdan-tarifas-durante-contingencias/.

Urminsky, O., C. Hansen, and V. Chernozhukov (2016). Using double-lasso regression for principled variable selection. *Available at SSRN 2733374*.

Vidgen, B., S. Hale, S. Staton, T. Melham, H. Margetts, O. Kammar, and M. Szymczak (2020). Recalibrating classifiers for interpretable abusive content detection. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pp. 132–138.

Waldron, J. (2012). *The harm in hate speech*. Harvard University Press.

Wang, L., R. Wright, and L. Q. Liu (2020). Sticky prices and costly credit. *International Economic Review 61*(1), 37–70.

Weitzman, M. L. (1991). Price distortion and shortage deformation, or what happened to the soap? *The American Economic Review*, 401–414.

Weyl, E. G. (2010). A price theory of multi-sided platforms. *American Economic Review 100*(4), 1642–72.

Whalen, A. (2020). What Did Twitter Do to James Woods? The Story Behind the Trend. https://www.newsweek.com/james-woods-twitter-feed-locked-freejameswoods-return-andrew-gillum\protect\discretionary{\char\hyphenchar\font}{}{}1494188. Accessed: 2021-10-13.

White, A. and E. G. Weyl (2016). Insulated platform competition. *Available at SSRN 1694317*.

Wickramasekera, N., J. Wright, H. Elsey, J. Murray, and S. Tubeuf (2015). Cost of crime: A systematic review. *Journal of Criminal Justice 43*(3), 218–228.

Wojcik, S. and A. Hughes (2019). Sizing up twitter users. *Pew Research Center 24*.

Wright, R., E. Tekin, V. Topalli, C. McClellan, T. Dickinson, and R. Rosenfeld (2017). Less cash, less crime: Evidence from the electronic benefit transfer program. *Journal of Law and Economics 60*(2), 361 – 383.

Wulczyn, E., N. Thain, and L. Dixon (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pp. 1391–1399.

Yang, K.-C., O. Varol, P.-M. Hui, and F. Menczer (2020). Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 34, pp. 1096–1103.

Yildirim, M. M., J. Nagler, R. Bonneau, and J. A. Tucker (2021). Short of suspension: How suspension warnings can reduce hate speech on twitter. *Perspectives on Politics*, 1–13.

Zannettou, S. (2021). " i won the election!": An empirical analysis of soft moderation interventions on twitter. *arXiv preprint arXiv:2101.07183*.

APPENDIX A

APPENDIX TO CHAPTER 1

A.1 Formal Propositions and Model Extensions

A.1.1 Overprovision or underprovision of moderation

The following proposition reproduces Spence's result for the model of Section 1.2.[1]

Proposition 1. *For fixed quantities T^θ and assuming that second-order conditions hold, the platform can overprovide or underprovide moderation relative to a surplus-maximizing planner. A sufficient condition for underprovision is that $P^\theta_{T^\theta} < 0$, for overprovision is that $P^\theta_{T^\theta} > 0$, and for efficient provision is that $P^\theta_{T^\theta} = 0$.*

Proof. A social planner chooses $\mathbf{T}$ and c to maximize total surplus W, which equals:

$$\begin{aligned} W(\mathbf{T}, c) =& \underbrace{w\left(\int_0^{t^A} p^A(t, T^H, c)dt + \int_0^{t^H} p^H(T^A, t, c)dt - p^A T^A - p^H T^H\right)}_{\text{Consumer surplus}} \\ &+ \underbrace{a(p^A(\mathbf{T}, c)T^A + p^H(\mathbf{T}, c)T^H) - \phi(\mathbf{T}, c)}_{\text{Producer surplus}} \\ =& w\left(\int_0^{T^A} p^A(t, T^H, c)dt + \int_0^{T^H} p^H(T^A, t, c)dt\right) \\ &+ (a - w)(p^A(\mathbf{T}, c)T^A + p^H(\mathbf{T}, c)T^H) - \phi(\mathbf{T}, c). \end{aligned}$$

The first two terms in the second equality are the areas below the inverse demand curves and the last term is the cost function, but the third term is new. This new term appears because the platform collects time with an opportunity cost w and sells it to advertisers for

1. The model differs from Spence's framework in two ways. First, the monopolist sells two "products" instead of one. Second, there is a gap between the opportunity cost of time (w) and the value of time spent watching ads (a).

a price a. To the best of my knowledge there are no analyses that compare the price of advertisements of social media to the opportunity cost of time, so the magnitude of $a - w$ is unknown.[2]

The first-order condition with respect to c from this problem is:

$$w\left(\int_0^{T^A}\frac{\partial p^A}{\partial c}dt+\int_0^{T^H}\frac{\partial p^H}{\partial c}dt\right)+(a-w)\left(\frac{\partial p^A}{\partial c}T^A+\frac{\partial p^H}{\partial c}T^H\right)=\frac{\partial\phi}{\partial c} \tag{A.1}$$

Suppose that $p^{\theta}_{t^{\theta}} < 0$.[3] Then $\partial p^A(t, T^H, c)/\partial c > \partial p^A(\mathbf{T}, c)/\partial c$ for all $t < T^A$ and likewise for H. Then, the left-hand side of equation (A.1) satisfies:

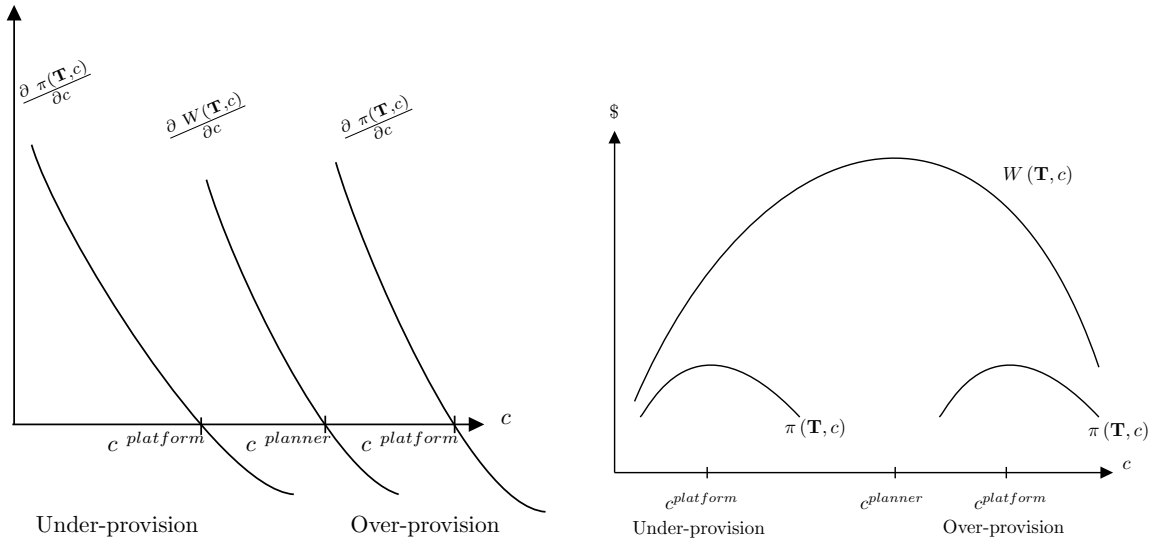
$$\begin{aligned}
&w\left(\int_0^{T^A}\frac{\partial p^A(t,T^H,c)}{\partial c}dt+\int_0^{T^H}\frac{\partial p^H(T^A,t,c)}{\partial c}dt\right)\\
&+(a-w)\left(\frac{\partial p^A(\mathbf{T},c)}{\partial c}T^A+\frac{\partial p^H(\mathbf{T},c)}{\partial c}T^H\right)\\
&>w\left(\frac{\partial p^A(\mathbf{T},c)}{\partial c}T^A+\frac{\partial p^H(\mathbf{T},c)}{\partial c}T^H\right)\\
&+(a-w)\left(\frac{\partial p^A(\mathbf{T},c)}{\partial c}T^A+\frac{\partial p^H(\mathbf{T},c)}{\partial c}T^H\right)\\
&=a\left(\frac{\partial p^A(\mathbf{T},c)}{\partial c}T^A+\frac{\partial p^H(\mathbf{T},c)}{\partial c}T^H\right),
\end{aligned}$$

which is identical to the left-hand side of equation (1.4). Since equations (1.4) and (A.1) both have $\partial\phi/\partial c$ on the right-hand side, this means that the planner's first-order condition is above the monopolist's one for fixed t^{θ} and all c: $\partial W(\mathbf{T}, c)/\partial c < \partial\pi(\mathbf{T}, c)/\partial c$. Assuming that

2. A no-arbitrage argument suggests that $a \approx w$. Suppose that ad prices were higher than the opportunity cost of time. This creates incentives for companies to pay users to watch advertisements. While Becker and Murphy (1993) argue that this might not be profitable, since consumers would "buy" a large number of ads and ignore as many as possible, current technology might facilitate this. Indeed, websites like `adwallet.com` reward consumers for watching ads. On the other hand, if $w > a$, platforms would find it more profitable to have consumers complete tasks rather than show them ads; e.g., "Fill out this survey in order to proceed to your feed".

3. The proof is analogous for the opposite case.

second-order conditions hold, this means that the root of the planner's first-order condition, $c^{planner}$, is higher than the root of the monopolist's condition, $c^{platform}$, so there is under-provision of moderation. Figure A.1 illustrates the proof.



(a) First-order conditions **(b)** Surplus and profits

Figure A.1: Illustration of moderation overprovision and underprovision

□

A.1.2 Generalization to Multiple Platforms

Assume without loss of generality that there are two platforms $j \in \{1, 2\}$. The solution concept of the model is a Cournot equilibrium as in Correia-da Silva et al. (2019).[4] First, platforms simultaneously set the amount of content T_j^θ on each side of the market and the moderation rates c_j. Then, given the quantities and moderation rates, prices adjust to equate demand and supply on each platform.

A fraction μ^θ of users are of type $\theta \in \{A, H\}$. Consumers are now characterized by the parameter vectors $\boldsymbol{\gamma}^\theta = (\gamma_1^\theta, \gamma_2^\theta, \delta_1^\theta, \delta_2^\theta)$. The γ's govern how utility responds to spillovers

4. Alternative solution concepts are flat pricing (Tan and Zhou, 2021) and insulating equilibrium (White and Weyl, 2016). See Correia-da Silva et al. (2019) for more discussion.

and the δ's govern membership benefits. As in (Weyl, 2010), the conditional density of membership benefits has full support. Users decide whether to join one of the platforms or neither. Below I discuss an extension to a multi-homing case. Once consumers join a platform, they decide how much time to spend on it. If they join platform j, they obtain membership benefits δ_j^θ and indirect utility:

$$v_j^\theta(\mathbf{T}_j, c_j, p_j^\theta, \gamma_j^\theta) = \max_{t\in[0,T]} u^\theta(t, \mathbf{T}_j, c_j; \gamma_j^\theta) - t \times w(1 + p_j^\theta),$$

where $\mathbf{T}_j = (T_j^A, T_j^H)$

Define the vectors $\mathbf{T} = (\mathbf{T}_1, \mathbf{T}_2)$, $\mathbf{T}^\theta = (T_1^\theta, T_2^\theta)$, $\boldsymbol{p}^\theta = (p_1^\theta, p_2^\theta)$, and $\boldsymbol{c} = (c_1, c_2)$, and the set of types that decide to use platform j as:

$$\overline{\boldsymbol{\gamma}}_j^\theta(\mathbf{T}, \boldsymbol{c}, \boldsymbol{p}^\theta) \equiv \left\{\boldsymbol{\gamma}^\theta : v_j^\theta(\mathbf{T}_j, c_j, p_j^\theta, \gamma_j^\theta) + \delta_j^\theta \geq \max\{v_{-j}^\theta(\mathbf{T}_{-j}, c_{-j}, p_{-j}^\theta, \gamma_{-j}^\theta) + \delta_{-j}^\theta, 0\}\right\},$$

where $-j$ denotes the other platform. Let $t_j(\mathbf{T}_j, c_j, p_j^\theta, \gamma_j^\theta)$ be the optimal time spent on platform j. Aggregate demands are:

$$T_j^\theta(\mathbf{T}, \boldsymbol{c}, \boldsymbol{p}^\theta) = \mu^\theta \int_{\overline{\boldsymbol{\gamma}}^\theta(\mathbf{T}, \boldsymbol{c}, \boldsymbol{p}^\theta)} t_j(\mathbf{T}_j, c_j, p_j^\theta, \gamma_j^\theta) f^\theta(\boldsymbol{\gamma}^\theta) d\boldsymbol{\gamma}^\theta$$

The consumer equilibrium constraints are, for all j and θ:

$$t_j^\theta = T_j^\theta(\mathbf{T}, \boldsymbol{c}, \boldsymbol{p}^\theta)$$

Inverting the demand curves is not as straightforward as in Weyl (2010), since demands now depend on the other platform's prices. We can, however, use the global inverse function theorem from Berry et al. (2013) to obtain the twice-continuously differentiable inverse demands $P_j^\theta(\mathbf{T}, \boldsymbol{c})$.[5]

5. Note that demands T_j^θ are twice-continuously differentiable and strictly decreasing in prices p_{-j}^θ and

The problem of platform j is now:

$$\max_{T_j^A, T_j^H, c_j} \pi_j(\mathbf{T}, \boldsymbol{c}) \equiv a \left(P_j^A(\mathbf{T}, \boldsymbol{c})\, T_j^A + P_j^H(\mathbf{T}, \boldsymbol{c})\, T_j^H \right) - \phi_j(\mathbf{T}_j, c_j).$$

The first-order condition with respect to the moderation rate is identical to equation (1.4), but using residual inverse demands instead of the market inverse demand curve:

$$a \left(\frac{\partial P_j^A}{\partial c_j}\, T_j^A + \frac{\partial P_j^H}{\partial c_j}\, T_j^H \right) = \frac{\partial C_j}{\partial c_j}$$

Hence, the same intuition of the platform's moderation decision holds in a model with two platforms. Moderation is a quality decision that allows platforms to increase their advertising revenue. The increase in ad revenue is the weighted change in willingness to pay of both types of users.

The following proposition shows that it is sufficient to measure the change in surplus on a sample of existing consumers; one can ignore the change in marginal users since they get zero surplus by definition.

Proposition 2. *The derivative of consumer surplus with respect to the moderation rate of platform j equals the average derivative of consumer surplus among users of that platform:*

$$\frac{\partial CS(\mathbf{T}, \boldsymbol{c}, \boldsymbol{p})}{\partial c_j} = \sum_\theta \mu^\theta \int_{\overline{\boldsymbol{\gamma}}_j^\theta(\mathbf{T}, \boldsymbol{c}, \boldsymbol{p}^\theta)} \frac{\partial v_j^\theta(\mathbf{T}_j, c_j, p_j^\theta, \gamma_j^\theta)}{\partial c_j} f^\theta(\boldsymbol{\gamma}^\theta) d\boldsymbol{\gamma}^\theta.$$

Proof. Define the membership benefit from joining platform $-j$ relative to the membership benefit from j as $\widetilde{\delta}_{-j}^\theta \equiv \delta_{-j}^\theta - \delta_j^\theta$. Define also the vector of network parameters of both platforms $\gamma^\theta \equiv (\gamma_1^\theta, \gamma_2^\theta)$, the vector of parameters $\widetilde{\boldsymbol{\gamma}}^\theta \equiv (\gamma^\theta, \delta_j^\theta, \widetilde{\delta}_{-j}^\theta)$ and the distribution of types $\widetilde{f}^\theta(\widetilde{\boldsymbol{\gamma}}^\theta) \equiv f^\theta(\gamma^\theta, \delta_j^\theta, \delta_{-j}^\theta + \delta_j^\theta)$. The membership benefits of those users who join

weakly decreasing in prices $p_{-j}^{-\mathcal{I}}$ and $p_j^{-\mathcal{I}}$, where $-\mathcal{I}$ denotes the other side. Moreover, the demand of the outside option is stricly increasing in all prices. Hence, this model satisfies all the conditions of Corollary 2 from Berry et al. (2013).

platform j are bounded as follows:

$$\begin{aligned}\delta_j^\theta &\geq -v_j^\theta(\mathbf{T}_j, c_j, p_j^\theta, \gamma_j^\theta),\\ \widetilde{\delta}_{-j} &\leq v_j^\theta(\mathbf{T}_j, c_j, p_j^\theta, \gamma_j^\theta) - v_{-j}^\theta(\mathbf{T}_{-j}, c_{-j}, p_{-j}^\theta, \gamma_j^\theta)\end{aligned}$$

Likewise, the bounds of the membership benefits of those users who join platform $-j$ are:

$$\begin{aligned}\delta_j^\theta &\geq -v_{-j}^\theta(\mathbf{T}_{-j}, c_{-j}, p_{-j}^\theta, \gamma_j^\theta) - \widetilde{\delta}_{-j},\\ \widetilde{\delta}_{-j} &\geq v_j^\theta(\mathbf{T}_j, c_j, p_j^\theta, \gamma_j^\theta) - v_{-j}^\theta(\mathbf{T}_{-j}, c_{-j}, p_{-j}^\theta, \gamma_j^\theta)\end{aligned}$$

Omitting the arguments of v_j^θ and v_{-j}^θ for brevity, the consumer surplus is:

$$\begin{aligned}CS(\mathbf{T}, \boldsymbol{c}, \boldsymbol{p}) = \sum_\theta \mu^\theta \Bigg(&\underbrace{\int \int_{-\infty}^{v_j^\theta - v_{-j}^\theta} \int_{-v_j^\theta}^{\infty} \left(v_j^\theta + \delta_j^\theta\right) \widetilde{f}^\theta(\widetilde{\boldsymbol{\gamma}}^\theta) d\delta_j^\theta d\widetilde{\delta}_{-j}^\theta d\gamma^\theta}_{\text{Surplus of } j\text{'s consumers}}\\ &+ \underbrace{\int \int_{v_j^\theta - v_{-j}^\theta}^{\infty} \int_{-v_{-j}^\theta - \widetilde{\delta}_{-j}^\theta}^{\infty} \left(v_{-j}^\theta + \widetilde{\delta}_{-j}^\theta + \delta_j^\theta\right) \widetilde{f}^\theta(\widetilde{\boldsymbol{\gamma}}^\theta) d\delta_j^\theta d\widetilde{\delta}_{-j}^\theta d\gamma^\theta}_{\text{Surplus of } -j\text{'s consumers}} \Bigg)\end{aligned} \tag{A.2}$$

Use the Leibniz integral rule to differentiate the first row after the equality sign from the

previous expression with respect to c_j:

$$\int \frac{\partial v_j^\theta}{\partial c_j} \int_{-v_j^\theta}^{\infty} \left(v_j^\theta + \delta_j^\theta\right) \widetilde{f}^\theta(\gamma^\theta, \delta_j^\theta, v_j^\theta - v_{-j}^\theta) d\delta_j^\theta d\gamma^\theta$$
$$+ \int \int_{-\infty}^{v_j^\theta - v_{-j}^\theta} \frac{\partial v_j^\theta}{\partial c_j} \left(\underbrace{v_j^\theta - v_j^\theta}_{=0} \right) \widetilde{f}^\theta(\gamma^\theta, -v_j^\theta, \widetilde{\delta}_{-j}^\theta) d\widetilde{\delta}_{-j}^\theta d\gamma^\theta$$
$$+ \int \int_{-\infty}^{v_j^\theta - v_{-j}^\theta} \int_{-v_j^\theta}^{\infty} \frac{\partial v_j^\theta}{\partial c_j} \widetilde{f}^\theta(\widetilde{\boldsymbol{\gamma}}^\theta) d\delta_j^\theta d\widetilde{\delta}_{-j}^\theta d\gamma^\theta \tag{A.3}$$

Likewise, differentiating the second row of equation (A.2):

$$\int -\frac{\partial v_j^\theta}{\partial c_j} \int_{-v_j^\theta}^{\infty} \left(v_j^\theta + \delta_j^\theta\right) \widetilde{f}^\theta(\gamma^\theta, \delta_j^\theta, v_j^\theta - v_{-j}^\theta) d\delta_j^\theta d\gamma^\theta \tag{A.4}$$

The first row from equation (A.3) cancels with equation (A.4), so adding these two expressions together gives

$$\int \int_{-\infty}^{v_j^\theta - v_{-j}^\theta} \int_{-v_j^\theta}^{\infty} \frac{\partial v_j^\theta}{\partial c_j} \widetilde{f}^\theta(\widetilde{\boldsymbol{\gamma}}^\theta) d\delta_j^\theta d\widetilde{\delta}_{-j}^\theta d\gamma^\theta,$$

which equals the average derivative of consumer surplus among users of j, $\boldsymbol{\gamma}^\theta \in \overline{\boldsymbol{\gamma}}_j^\theta(\mathbf{T}, \boldsymbol{c}, \boldsymbol{p}^\theta)$.

□

The previous single-homing model can be extended to allow multi-homing and flexible substitution or complementarity patterns by considering bundles of platforms as different choices, by using the approach of Gentzkow (2007) as Berry et al. (2013) suggest. Proposition 2 still holds in such a model, but the indirect utility v_j^θ would represent the utility from using any bundle of platforms that includes j, and v_{-j}^θ would correspond to bundles without j. Hence, in that setting, $\partial v_j^\theta / \partial c_j$ would be the derivative of the surplus of using social media, not just platform j, for users who join any bundle of platforms that includes j.

In a multi-platform model, the steps to test whether a platform under-provides or over-provides moderation, for fixed quantities, are as follows. First, introduce a small change

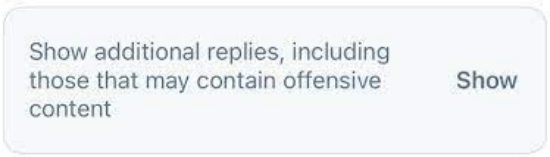
in the moderation rate of the desired platform. Second, compute the average change in consumer surplus from using social media—not just the surplus from using the platform—among representative consumers of the platform.

A.2 Experimental Design

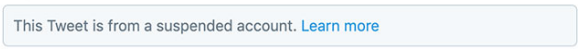
Table A.1: Query list

Holocaust	holohoax, h0lohoax, hol0hoax, holoh0ax, holoho4x, holoho@x, holoho*x, h0l0h0ax, h0l0h04x, holo hoax, h0lo hoax, hol0 hoax, holo h0ax, holo ho4x, holo ho@x, holo ho*x, h0l0 h0ax, h0l0 h04x, holo-hoax, h0lo-hoax, hol0-hoax, holo-h0ax, holo-ho4x, holo-ho@x, holo-ho*x, h0l0-h0ax, h0l0-h04x
Disability	youre a retard, youre a r3tard, youre a r*tard, youre a ret*rd, youre a returd, youre a ret@rd, youre a ret4rd, youre a r3t4rd, youre a r3t@rd, your a retard, your a r3tard, your a r*tard, your a ret*rd, your a returd, your a ret@rd, your a ret4rd, your a r3t4rd, your a r3t@rd, you're a retard, you're a r3tard, you're a r*tard, you're a ret*rd, you're a returd, you're a ret@rd, you're a ret4rd, you're a r3t4rd, you're a r3t@rd, ure a retard, ure a r3tard, ure a r*tard, ure a ret*rd, ure a returd, ure a ret@rd, ure a ret4rd, ure a r3t4rd, ure a r3t@rd, ur a retard, ur a r3tard, ur a r*tard, ur a ret*rd, ur a returd, ur a ret@rd, ur a ret4rd, ur a r3t4rd, ur a r3t@rd, u're a retard, u're a r3tard, u're a r*tard, u're a ret*rd, u're a returd, u're a ret@rd, u're a ret4rd, u're a r3t4rd, u're a r3t@rd youre retarded, youre r3tarded, youre r*tarded, youre ret*rded, youre returded, youre ret@rded, youre ret4rded, youre r3t4rded, youre r3t@rded, you're retarded, you're r3tarded, you're r*tarded, you're ret*rded, you're returded, you're ret@rded, you're ret4rded, you're r3t4rded, you're r3t@rded, ure retarded, ure r3tarded, ure r*tarded, ure ret*rded, ure returded, ure ret@rded, ure ret4rded, ure r3t4rded, ure r3t@rded, u're retarded, u're r3tarded, u're r*tarded, u're ret*rded, u're returded, u're ret@rded, u're ret4rded, u're r3t4rded, u're r3t@rded youre a retarded, youre a r3tarded, youre a r*tarded, youre a ret*rded, youre a returded, youre a ret@rded, youre a ret4rded, youre a r3t4rded, youre a r3t@rded, your a retarded, your a r3tarded, your a r*tarded, your a ret*rded, your a returded, your a ret@rded, your a ret4rded, your a r3t4rded, your a r3t@rded, you're a retarded, you're a r3tarded, you're a r*tarded, you're a ret*rded, you're a returded, you're a ret@rded, you're a ret4rded, you're a r3t4rded, you're a r3t@rded, ure a retarded, ure a r3tarded, ure a r*tarded, ure a ret*rded, ure a returded, ure a ret@rded, ure a ret4rded, ure a r3t4rded, ure a r3t@rded, ur a retarded, ur a r3tarded, ur a r*tarded, ur a ret*rded, ur a returded, ur a ret@rded, ur a ret4rded, ur a r3t4rded, ur a r3t@rded, u're a retarded, u're a r3tarded, u're a r*tarded, u're a ret*rded, u're a returded, u're a ret@rded, u're a ret4rded, u're a r3t4rded, u're a r3t@rded

(a) Public notice of Tweet removal

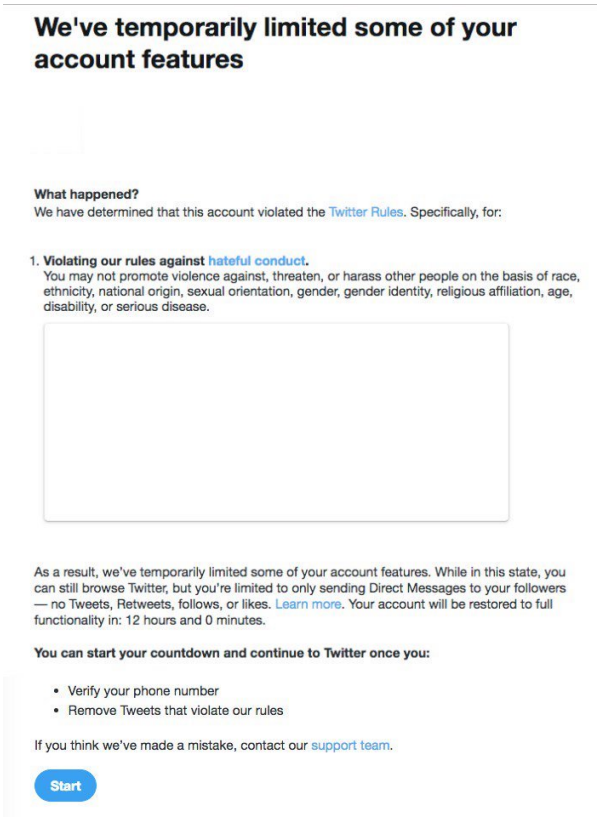


(b) Public notice of shadowban

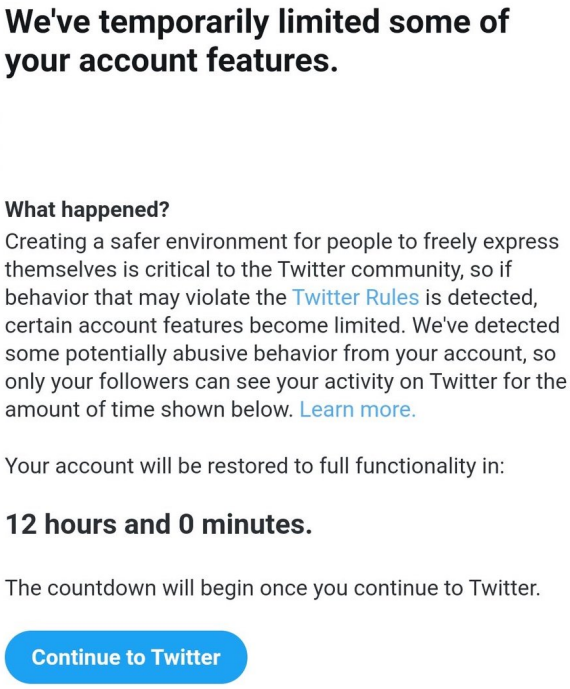


(c) Public notice of suspension (1)

(d) Public notice of suspension (2)



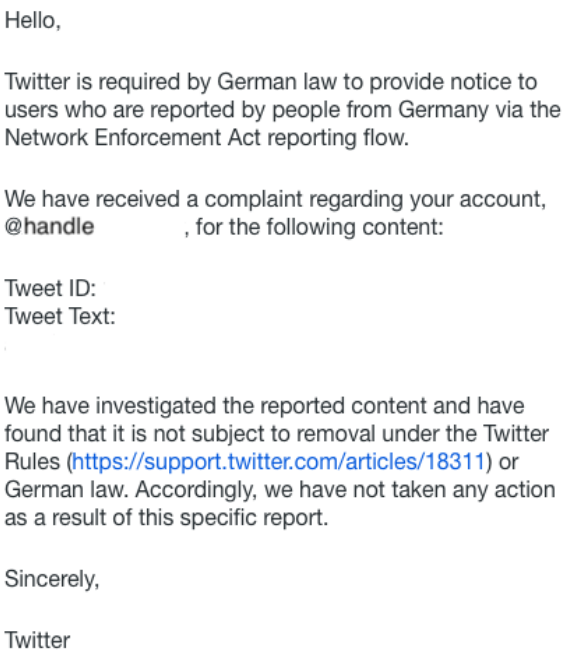
(e) Notification of a Tweet removal



(f) Notification of a locked account

Figure A.2: Public notices and notifications

Notes: This figure includes images of public notices of different sanctions and an example of a notification that a users receive when their account is locked.



Hello,

Twitter is required by German law to provide notice to users who are reported by people from Germany via the Network Enforcement Act reporting flow.

We have received a complaint regarding your account, @handle , for the following content:

Tweet ID:
Tweet Text:

We have investigated the reported content and have found that it is not subject to removal under the Twitter Rules (https://support.twitter.com/articles/18311) or German law. Accordingly, we have not taken any action as a result of this specific report.

Sincerely,

Twitter

Figure A.3: Screenshot of a notification of a user report

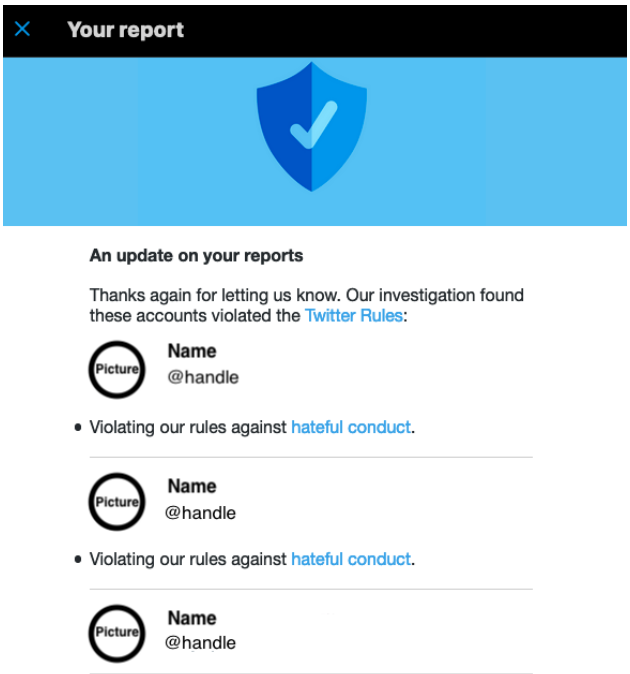


Figure A.4: Screenshot of an update on reports

A.3 Data Appendix

A.3.1 Measurement of sanctions

The "lookup statuses" or "lookup users" endpoints of the Twitter API indicate when a tweet or account go missing. Among missing accounts and statuses, the "show users" or "show

Table A.2: Variable definition

Variable	Definition
Account years	Years from account creation date until measurement date
Tweets per day	Statuses count divided by days since account creation
Likes per day	Likes count divided by days since account creation
Followers	Number of accounts that follow a user
Followed	Number of accounts that the user follows
Bot score	Probability of being a bot, from Botometer API
Is bot	Indicates whether bot score ≥ 0.5
Initial shadow ban	Whether an account is shadow banned at the time of sampling
Word count	Number of words in a tweet
Is toxic	Indicates whether toxicity ≥ 0.8
Is hate (MTurk)	Indicates whether a majority of MTurkers label the post as hate
Is reply	Indicates whether the tweet is a reply to another user
Is attack (MTurk)	Indicates whether the majority of MTurkers consider the post to be an attack on another user
Is quote	Indicates whether the tweet is a quote to another user
Is mention	Indicates whether the tweet mentions another user
Has media	Indicates whether the tweet contains a video or picture
Disability key word	Indicates whether the tweet contains the expression "r*t*rd"
Holocaust key word	Indicates whether the tweet contains the expressions "h*l*h**x", "h*l*c**st", "jew"
Tweet from phone	Indicates whether the source of the tweet is Twitter for iPhone or Twitter for Android
Has description	Indicates whether a profile has a description
Has location	Indicates whether a profile has a location
Default picture	Indicates whether a profile has a default profile picture
Is verified	Indicates whether an account is verified
Has Instagram	Indicates whether a profile description, location or URL contains an Instagram handle
Has backup	Indicates whether a profile description, location or URL contains an alternative or backup Twitter handle
Has pronouns	Indicates whether a profile description or location contains pronouns or a `carrd.co` link
Under 18	Indicates whether a profile description or location contains numbers 13 to 17 (in number or word), years 2003 to 2008 or words like "minor" or "teen"
Previous toxicity	Indicates whether any of a user's most recent 50 tweets has toxicity ≥ 0.8
Previous disability	Indicates whether any of a user's most recent 50 tweets has the expression "r*t*rd"
Previous Holocaust	Indicates whether any of a user's most recent 50 tweets has the expressions "h*l*h**x", "h*l*c**st", "jew'

statuses" endpoints of the API return an error code that details why they were missing (see Twitter (2021f) for a full list of error codes). With the error code information one can measure the following events:

- Twitter required the removal of a post, but it has not been removed by the user. This

Table A.3: Balance in the reporting experiment

Characteristic	Control		Treatment		Difference	
	Mean	SD	Mean	SD	Normalized	p-value
Observations	3,074		3,074			
Accounts						
Account years	3.21	3.4	3.23	3.5	-0.01	0.77
Tweets per day	11.19	24.2	12.05	25.2	-0.03	0.20
Likes per day	23.39	50.2	24.95	51.6	-0.03	0.22
Followers	517.07	3,439.5	752.64	6,476.9	-0.05	0.08
Followed	426.84	751.4	440.65	946.0	-0.02	0.55
Bot score	0.24	0.1	0.24	0.1	0.01	0.54
Initial shadow ban	0.71	0.5	0.71	0.5	0.01	0.78
Tweets						
Word count	16.02	13.2	15.93	13.4	0.01	0.80
Is toxic	0.81	0.4	0.80	0.4	0.04	0.19
Is hate (MTurk)	0.31	0.5	0.30	0.5	0.00	0.81
Is reply	0.84	0.4	0.84	0.4	0.00	0.95
Is attack (MTurk)	0.78	0.4	0.78	0.4	-0.02	0.42
Is quote	0.07	0.3	0.07	0.3	0.02	0.53
Is mention	0.85	0.4	0.85	0.4	0.01	0.76
Has media	0.04	0.2	0.04	0.2	-0.04	0.13
Tweet from phone	0.80	0.4	0.78	0.4	0.03	0.25
Profiles						
Has description	0.82	0.4	0.82	0.4	-0.01	0.70
Has location	0.51	0.5	0.51	0.5	-0.01	0.53
Default picture	0.04	0.2	0.03	0.2	0.02	0.42
Is verified	0.00	0.0	0.00	0.0	-0.02	0.51
Has Instagram	0.01	0.1	0.02	0.1	-0.02	0.46
Has backup	0.01	0.1	0.01	0.1	0.04	0.15
Timelines						
Previous toxicity	0.94	0.2	0.93	0.2	0.01	0.83
Previous disability	0.39	0.5	0.39	0.5	0.01	0.74
Previous Holocaust	0.10	0.3	0.10	0.3	-0.01	0.98
Joint tests/differences						
F-test (p-value)						0.70
Multivariate normalized difference					0.12	

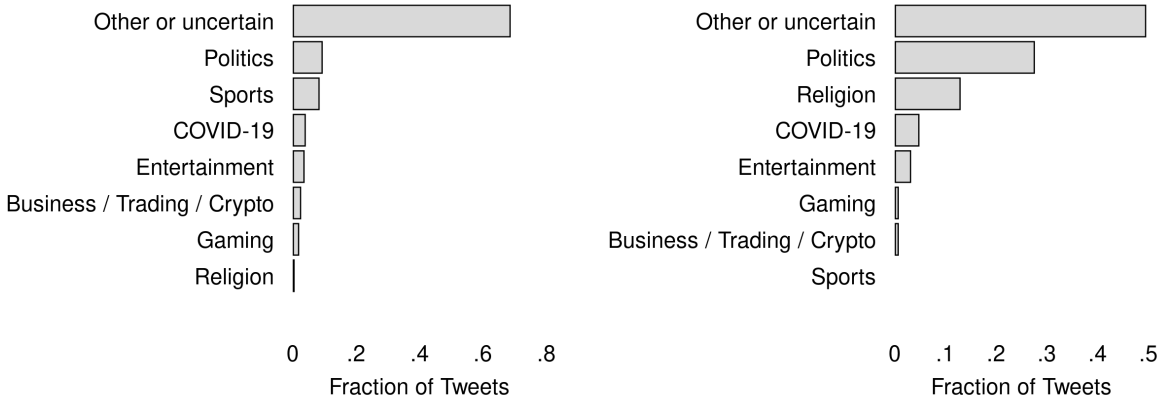
Notes: Columns 2 to 5 display means and standard deviations (SD). Column 6 displays normalized differences (Imbens and Rubin, 2015); all variables have differences below the recommended 0.25. Column 7 has p-values from regressions of characteristics on a treatment dummy and strata fixed-effects. F-tests are from regressions of a treatment indicator on pre-treatment variables.

is reflected in a missing status with error code 421.

- Twitter required the removal of a post, and it has been removed by the user. This is

Figure A.5: Screenshot of a reply

Notes: Some Tweets in my sample are replies or comments to other users' Tweets.



(a) Disability **(b)** Holocaust

Figure A.6: Topic classification by slur

Notes: Each figure presents the distribution of Tweets by their main topic. Three MTurk workers read each Tweet and decided its most relevant topic among the eight options in the figures. The main topic is the one that two or three workers agreed upon. If there was no agreement, the topic of the Tweet is set to "Other or uncertain".

reflected in a missing status with error code 422. After some days, the status transitions to code 144 (deleted status). Twitter claims that the notice will be available 14 days

Table A.4: Reporting accounts summary statistics

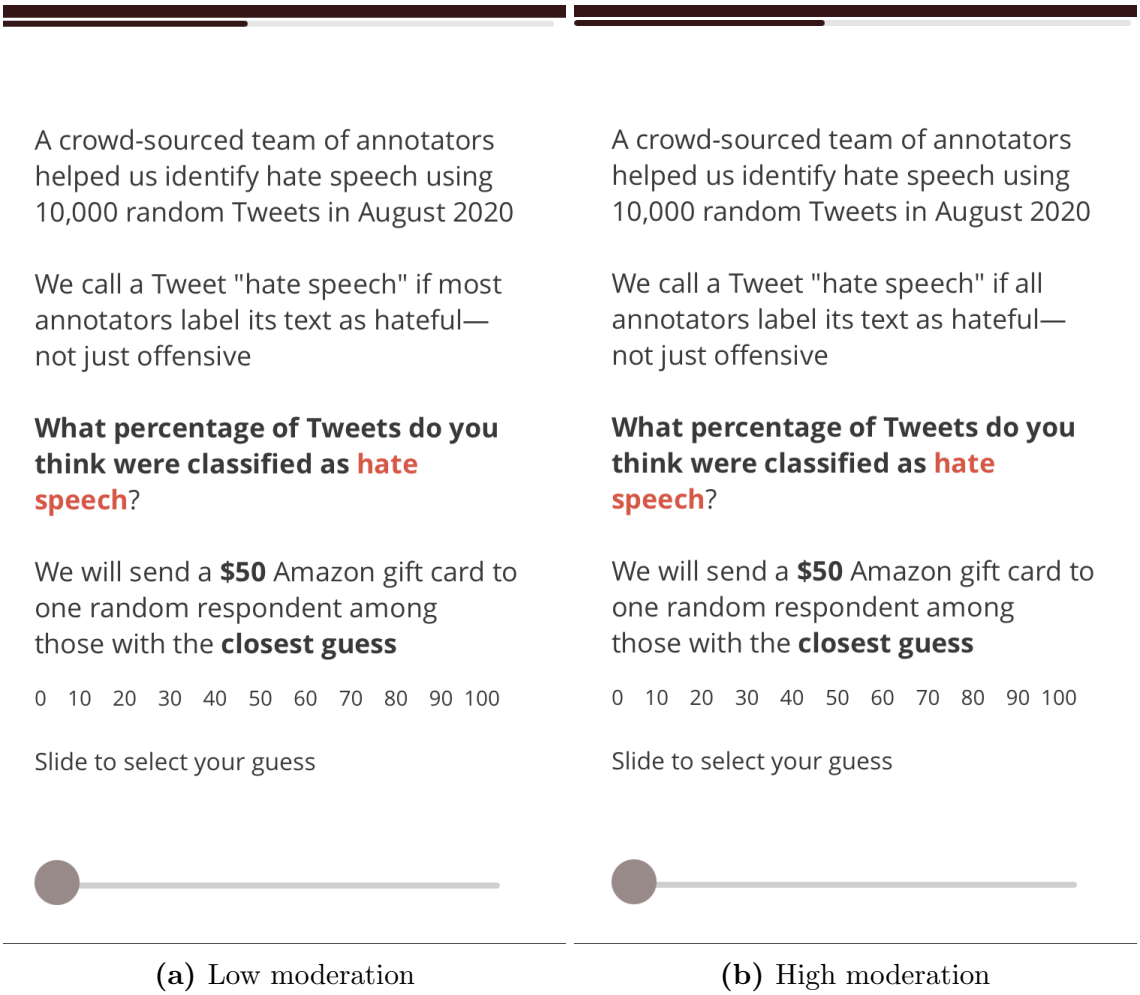
	Account										
	1	2	3	4	5	6	7	8	9	10	11
Accounts											
City	CHI	CHI	NYC	MIA	LA	LA	DAL	SF	ATL	CHI	DC
Email	Yes	Yes	Yes	No	Yes	No	Yes	No	No	Yes	No
Phone	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mobile	Yes	No	No	No	No	Yes	No	No	Yes	No	Yes
App	Yes	No	No	No	No	No	Yes	No	No	Yes	No
Shadow ban	No	No	No	No	Yes	No	No	No	Yes	No	No
Account yrs	2.7	2.5	2.5	2.3	2.3	2.3	2.3	2.2	0.3	9.1	0.1
Tweets/mth	0	1.1	0.3	2.4	2.2	0.1	0.2	0.3	3.5	0.5	4.1
Likes/mth	0	1.0	0.6	1.8	1.8	0.8	1.1	0.5	4.7	1.2	8.2
Followers	0	18	3	2	1	0	0	0	0	168	1
Followed	6	22	28	48	43	25	19	16	14	134	17
Bot score	.	0.4	0.5	0.2	0.3	0.4	0.4	0.5	0.4	0.2	0.5
Profiles											
Description	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Location	Yes	Yes	No	Yes	Yes	No	No	Yes	No	Yes	Yes
Default pic	No	No	No	No	No	No	No	No	No	No	No
Verified	No	No	No	No	No	No	No	No	No	No	No
Protected	No	No	No	No	No	No	No	No	No	No	No

Notes: Each column corresponds to one of the 11 Twitter accounts used for the reporting treatment. City is the location of the virtual private network used for reporting. Email and Phone indicate whether the account had an associated email and phone number, respectively. Mobile indicates whether reporting was done using a phone or a computer. App indicates whether the account was accessed using the official Twitter app or a browser. Data gathered in August, 2021.

after the tweet is removed (Twitter, 2021d) but empirically it seems like this period varies.

- A post is missing because the user deleted it. This is reflected in a missing status with error code 144.

- A post is missing because the user protected their account or because the user blocked my developer account. This is reflected in a missing status with error code 179 or 136, respectively. Protected accounts are also detected with the lookup users endpoint. It is rare to encounter a user that blocks my developer account. Most likely is due to users mass-blocking all the followers of some famous account.



(a) Low moderation **(b)** High moderation

Figure A.7: Instructions and elicitation of beliefs about prevalence

- A post and the account are missing because the user is suspended. This is reflected in a missing user and potentially missing status with code 63 (Chowdhury et al., 2020).
- A post and the account are missing because the user deleted their account. This is reflected in a missing user with code 50.

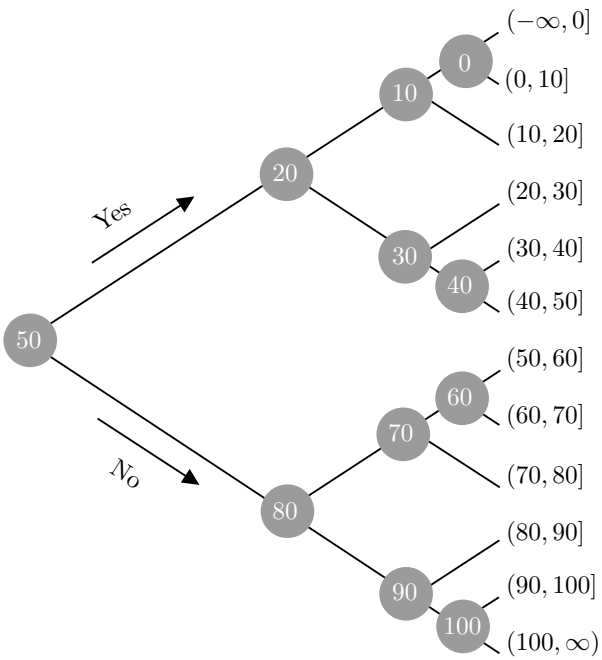


Figure A.8: Iterative multiple price list

Notes: The circles denote compensation (Amazon gift card) offers to deactivate social media. The intervals correspond to the willingness to accept.

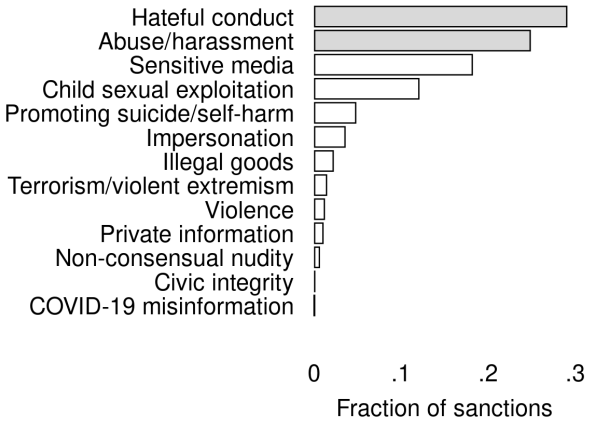


Figure A.9: Histogram of sanctions by rule violation

Notes: This figure plots the fraction of sanctions (actioned accounts) by the type of rule violation. It uses data from the second half of 2020 from Twitter's Transparency Rules Enforcement Report (Twitter, 2020b).

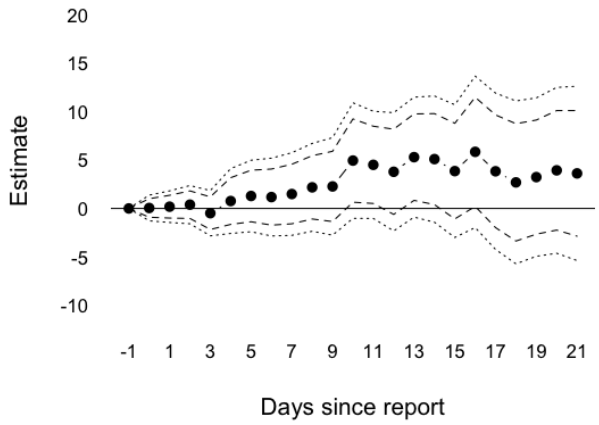
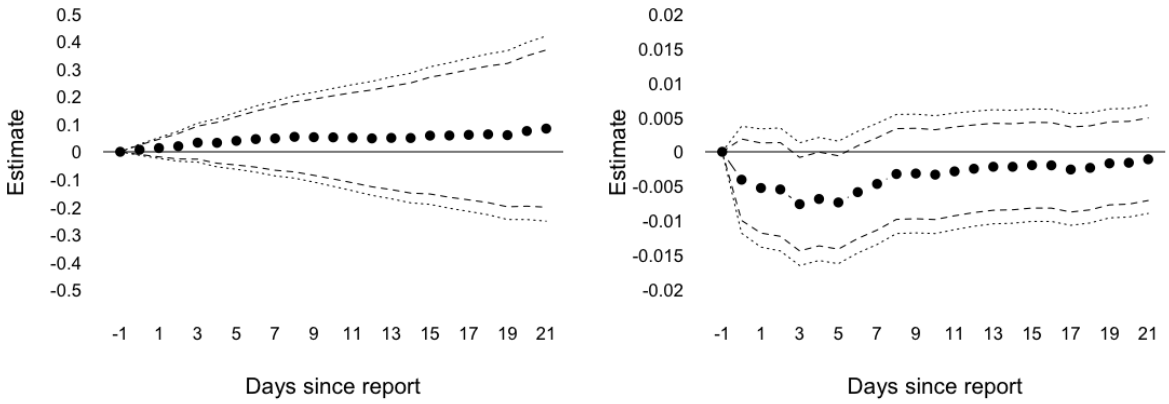


Figure A.10: Effect on hours since last post

Notes: This figure presents dynamic treatment effects on the number of hours since the last post at midnight of every day after reporting. Pointwise confidence intervals are dashed and sup-t confidence bands are dotted.



(a) Hours spent on Twitter **(b)** Fraction of hateful posts

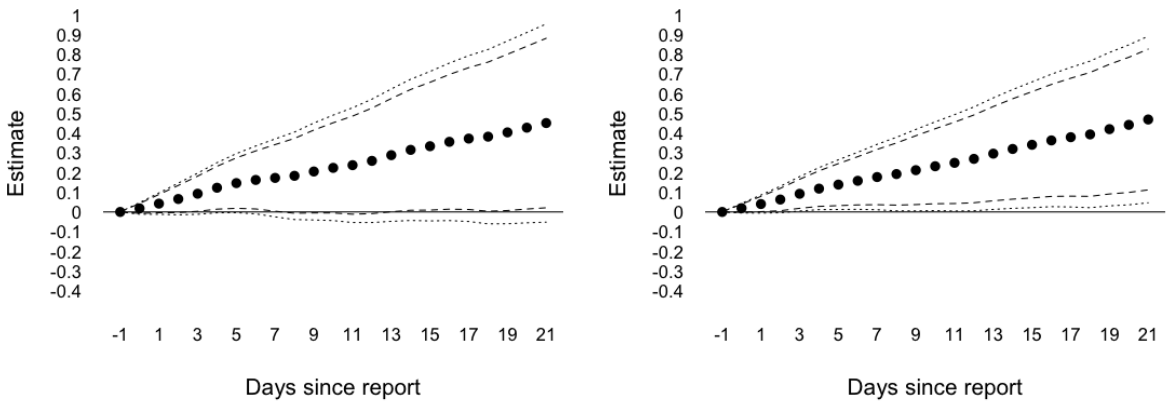
Figure A.11: Cumulative dynamic treatment effects on activity and hatefulness

Notes: This figure presents cumulative dynamic treatment effects, pointwise confidence intervals (dashed), and sup-t simultaneous confidence bands (dotted).

A.3.2 Figures and Tables

A.4 Survey Instruments

A.4.1 Classification of random posts



(a) All replies **(b)** Attacks

Figure A.12: Cumulative dynamic treatment effect on replies activity

Notes: This figure presents cumulative dynamic treatment effects, pointwise confidence intervals (dashed), and sup-t simultaneous confidence bands (dotted). The outcome variable is a measure of time spent of the users that the posts in the sample reply to. It is a linear combination of Tweets and Likes.

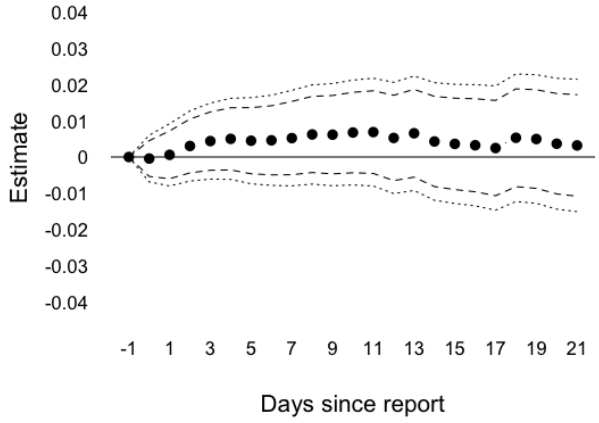


Figure A.13: Cumulative dynamic treatment effect on attrition

Notes: This figure presents dynamic treatment effects on an indicator of whether users drop from the sample at or before every day after reporting. Pointwise confidence intervals are dashed and sup-t confidence bands are dotted.

Table A.5: Balance in the welfare experiment

Characteristic	Control		Treatment		N. Dif.	p-value
	Mean	SD	Mean	SD		
Observations	1,515		1,512			
Demographics						
Age	38.05	12.8	38.10	12.3	-0.00	0.92
Female	0.45	0.5	0.45	0.5	0.01	0.44
College graduate +	0.32	0.5	0.31	0.5	0.03	0.35
Some college	0.33	0.5	0.33	0.5	-0.01	0.73
White non-Hispanic	0.67	0.5	0.69	0.5	-0.05	0.06
Black non-Hispanic	0.15	0.4	0.14	0.4	0.01	0.82
Hispanic	0.09	0.3	0.08	0.3	0.04	0.27
Asian non-Hispanic	0.03	0.2	0.02	0.2	0.06	0.10
Northeast	0.22	0.4	0.25	0.4	-0.07	0.07
Midwest	0.18	0.4	0.18	0.4	0.01	0.84
South	0.39	0.5	0.36	0.5	0.06	0.12
Republican	0.23	0.4	0.22	0.4	0.03	0.47
Democrat	0.52	0.5	0.54	0.5	-0.04	0.33
Christian	0.62	0.5	0.61	0.5	0.02	0.54
Jewish	0.02	0.1	0.02	0.1	-0.00	0.91
Muslim	0.04	0.2	0.04	0.2	-0.01	0.78
Buddhist or Hindu	0.01	0.1	0.01	0.1	0.02	0.65
Income	64.09	32.9	64.22	33.4	-0.00	0.90
Minority	0.48	0.5	0.48	0.5	-0.00	1.00
Twitter / Social media						
Daily hours on Twitter	1.52	2.3	1.52	2.2	-0.00	0.99
Provided handle	0.64	0.5	0.64	0.5	-0.00	1.00
User exists	0.47	0.5	0.47	0.5	-0.01	0.69
Tweets per day	1.29	6.4	1.79	10.7	-0.06	0.29
Likes per day	1.82	6.3	2.86	18.1	-0.08	0.14
Account years	8.06	4.3	7.80	4.2	0.06	0.25
Platforms other than Twitter	5.12	2.1	5.10	2.0	0.01	0.74
Has been harassed online	0.28	0.5	0.29	0.5	-0.01	0.85
Prevalence of hate in feed	20.06	23.1	20.85	24.5	-0.03	0.31
Moderation						
Has been sanctioned	0.23	0.4	0.23	0.4	0.00	1.00
Has reported	0.36	0.5	0.37	0.5	-0.00	0.88
Has been reported	0.12	0.3	0.12	0.3	0.03	0.36
Beliefs						
Prevalence of hate	36.73	25.7	36.12	26.2	0.02	0.51
Likelihood of moderation	39.73	28.6	40.91	28.7	-0.04	0.25
Joint tests/differences						
F-test (p-value)						0.33
Multivariate normalized difference					0.25	

Notes: Columns 2 to 5 display means and standard deviations (SD). Column 6 displays normalized differences (Imbens and Rubin, 2015). Column 7 has p-values from a regression of characteristics on treatment and strata fixed-effects. F-tests are from regressions of a treatment indicator on characteristics.

Table A.6: Effects of reporting on other observable sanctions and self-censorship

Panel A: other Twitter sanctions

	Suspensions			Shadow-bans			Missing Other Tweets		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	-0.000	-0.000	-0.000	0.001	0.001	-0.002	0.004	0.004	0.004
	(0.006)	(0.006)	(0.006)	(0.012)	(0.012)	(0.011)	(0.005)	(0.005)	(0.005)
y Mean	0.05	0.05	0.05	0.26	0.26	0.26	0.05	0.05	0.05
y SD	0.22	0.22	0.22	0.44	0.44	0.44	0.18	0.18	0.18
R^2	0.00	0.03	0.03	0.00	0.02	0.10	0.00	0.02	0.03
Obs.	6,148	6,134	6,134	5,692	5,675	5,675	5,381	5,360	5,360

Panel B: self-censorship

	Tweet deletion			Account deletion			Protecting account		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	0.006	0.005	0.005	0.001	0.001	0.001	0.000	0.000	-0.000
	(0.005)	(0.005)	(0.005)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)
y Mean	0.03	0.03	0.03	0.02	0.02	0.02	0.03	0.03	0.03
y SD	0.18	0.18	0.18	0.13	0.13	0.13	0.17	0.17	0.17
R^2	0.00	0.01	0.01	0.00	0.02	0.02	0.00	0.02	0.04
Obs.	6,148	6,134	6,134	6,148	6,134	6,134	6,148	6,134	6,134
Strata FE	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions on treatment assignment. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,**, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.7: Effects of reporting on other measures of activity

Panel A: Tweets and Likes

	Tweets			Likes		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	35.882*	35.643*	22.946	26.416	27.580	4.863
	(20.694)	(20.753)	(15.419)	(41.243)	(41.467)	(29.651)
y Mean	405.47	405.89	405.89	846.49	847.86	847.86
y SD	782.50	783.58	783.58	1559.14	1560.95	1560.95
R^2	0.00	0.02	0.45	0.00	0.01	0.49
Obs.	5,717	5,697	5,697	5,717	5,697	5,697

Panel B: other activity measures

	Winsorized time			Fraction of active days		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.192	0.192	0.126	-0.001	-0.001	-0.001
	(0.133)	(0.134)	(0.109)	(0.001)	(0.001)	(0.001)
y Mean	3.36	3.37	3.37	1.09	1.09	1.09
y SD	5.05	5.05	5.05	0.04	0.04	0.04
R^2	0.00	0.02	0.34	0.00	0.01	0.01
Obs.	5,717	5,697	5,697	5,727	5,708	5,708
Strata FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions on treatment assignment. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.8: Effects of reporting on other measures of hatefulness

Panel A: extensive margin						
	Posting toxicity ≥ 0.8			Repeating the slur		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.004	0.004	0.003	0.001	0.001	0.001
	(0.008)	(0.008)	(0.008)	(0.013)	(0.013)	(0.011)
y Mean	0.90	0.90	0.90	0.62	0.61	0.61
y SD	0.30	0.30	0.30	0.49	0.49	0.49
R^2	0.00	0.01	0.03	0.00	0.02	0.34
Obs.	5,727	5,708	5,708	5,727	5,708	5,708
Panel B: average scores						
	Average toxicity			Average severe toxicity		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	-0.001	-0.001	0.000	-0.001	-0.001	-0.000
	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
y Mean	0.30	0.30	0.30	0.18	0.18	0.18
y SD	0.11	0.11	0.11	0.09	0.09	0.09
R^2	0.00	0.01	0.09	0.00	0.01	0.08
Obs.	5,631	5,616	5,616	5,631	5,616	5,616
Strata FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions on treatment assignment. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.9: Effects of reporting on other measures of replied users' activity

Panel A: Tweets and Likes

	Tweets			Likes		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	59.739*	58.073*	50.849*	152.538***	148.888***	141.722**
	(30.756)	(30.903)	(30.824)	(56.932)	(57.615)	(57.611)
y Mean	656.20	657.39	657.39	1151.51	1151.96	1151.96
y SD	1060.33	1062.04	1062.04	1963.42	1964.77	1964.77
R^2	0.00	0.02	0.03	0.00	0.02	0.03
Obs.	4,752	4,733	4,733	4,752	4,733	4,733

Panel B: other activity measures

	Winsorized time			Fraction of active days		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.420**	0.408*	0.362*	-0.019	-0.032	-0.033
	(0.210)	(0.211)	(0.210)	(0.113)	(0.113)	(0.113)
y Mean	5.21	5.21	5.21	20.42	20.42	20.42
y SD	7.23	7.24	7.24	3.91	3.91	3.91
R^2	0.00	0.02	0.04	0.00	0.02	0.02
Obs.	4,752	4,733	4,733	4,761	4,742	4,742
Strata FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions on treatment assignment. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.10: Effects of reporting on other measures of replied users' activity, sample of attacks

Panel A: Tweets and Likes

	Tweets			Likes		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	77.136**	74.865**	71.894**	157.203**	156.095**	152.089**
	(32.018)	(32.165)	(32.037)	(61.304)	(62.240)	(62.025)
y Mean	635.01	635.59	635.59	1140.39	1142.13	1142.13
y SD	1035.93	1037.27	1037.27	1983.49	1986.15	1986.15
R^2	0.00	0.02	0.03	0.00	0.02	0.03
Obs.	4,171	4,155	4,155	4,171	4,155	4,155

Panel B: other activity measures

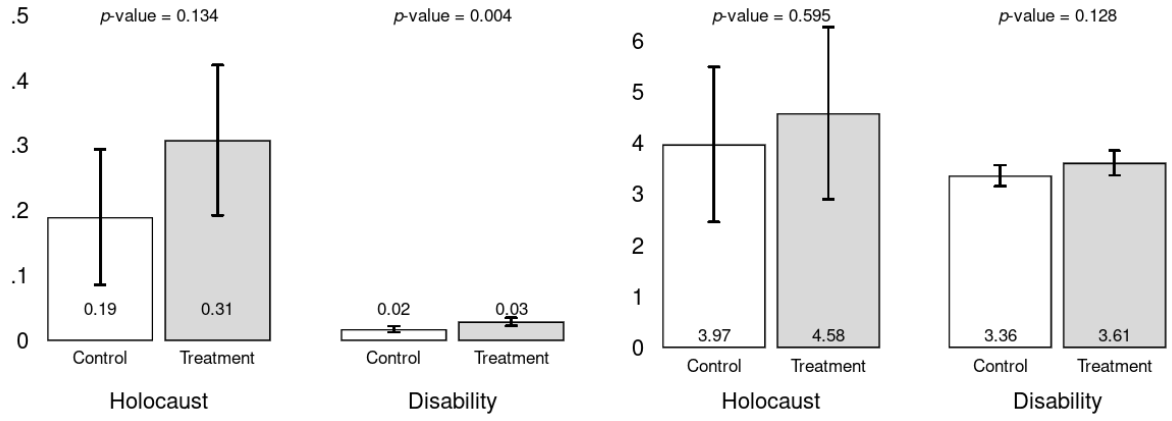
	Winsorized time			Fraction of active days		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.512**	0.499**	0.478**	-0.012	-0.028	-0.028
	(0.221)	(0.223)	(0.222)	(0.123)	(0.123)	(0.123)
y Mean	5.07	5.08	5.08	20.35	20.35	20.35
y SD	7.14	7.15	7.15	3.97	3.97	3.97
R^2	0.00	0.02	0.03	0.00	0.02	0.02
Obs.	4,171	4,155	4,155	4,178	4,162	4,162
Strata FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions on treatment assignment. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,**, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.11: Effects of reporting on attrition

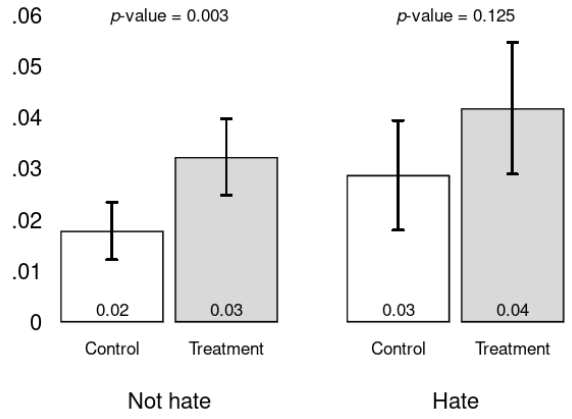
	Attrition on day 21			Attrition on day $\leq$ 21		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.002	0.001	0.001	0.002	0.002	0.001
	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)
y Mean	0.07	0.07	0.07	0.08	0.08	0.08
y SD	0.25	0.25	0.25	0.28	0.28	0.28
R^2	0.00	0.02	0.03	0.00	0.02	0.04
Obs.	6,148	6,134	6,134	6,148	6,134	6,134
Strata FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions on treatment assignment. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

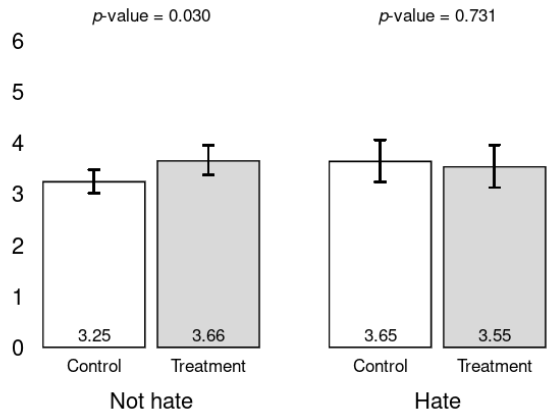


(a) Tweet removals by slur

(b) Time spent by slur



(c) Tweet removals by annotation



(d) Time spent by annotation

Figure A.14: Heterogeneity by slur and hate annotation

Notes: This figure reports estimates of reporting on Tweet removals and users' time spent posting and liking by slur and hate annotation.

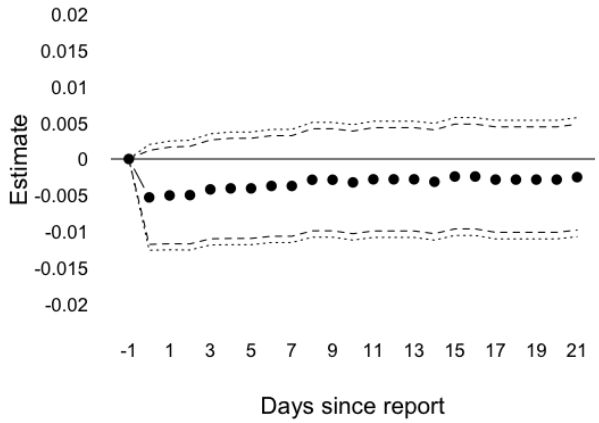


Figure A.15: Cumulative effect on the likelihood of mentioning the replied user

Notes: This figure presents dynamic treatment effects on an indicator of whether the users in the sample mention the replied users again. Pointwise confidence intervals are dashed and sup-t confidence bands are dotted.

Table A.12: Effects on sanctions among Tweets with replied and attacked users

Panel A: Sample of replies

	Tweet removals			Suspensions			Shadow-bans		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	0.008*	0.008*	0.008*	0.003	0.002	0.002	-0.003	-0.001	-0.004
	(0.004)	(0.004)	(0.004)	(0.007)	(0.006)	(0.006)	(0.013)	(0.013)	(0.013)
y Mean	0.02	0.02	0.02	0.05	0.05	0.05	0.24	0.24	0.24
y SD	0.15	0.15	0.15	0.22	0.22	0.22	0.43	0.43	0.43
R^2	0.00	0.08	0.08	0.00	0.03	0.03	0.00	0.02	0.08
Obs.	4,752	4,734	4,734	4,752	4,734	4,734	4,404	4,388	4,388

Panel B: Sample of attacks

	Tweet removals			Suspensions			Shadow-bans		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	0.006	0.004	0.004	0.002	0.001	0.001	-0.010	-0.008	-0.010
	(0.005)	(0.005)	(0.005)	(0.007)	(0.007)	(0.007)	(0.013)	(0.013)	(0.013)
y Mean	0.02	0.02	0.02	0.05	0.05	0.05	0.22	0.22	0.22
y SD	0.15	0.15	0.15	0.22	0.22	0.22	0.42	0.42	0.42
R^2	0.00	0.04	0.04	0.00	0.02	0.02	0.00	0.02	0.06
Obs.	4,165	4,149	4,149	4,165	4,149	4,149	3,860	3,845	3,845
Strata FE	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions on treatment assignment. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

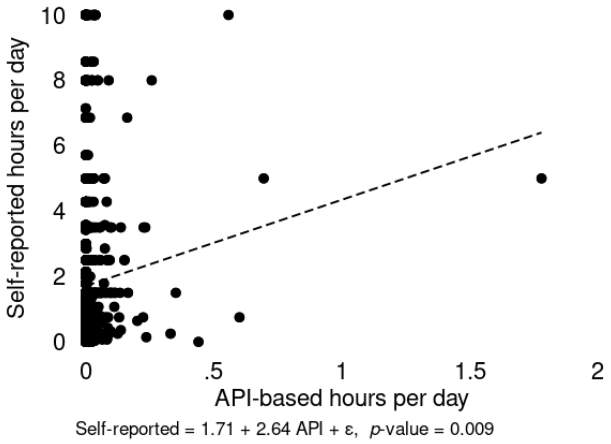


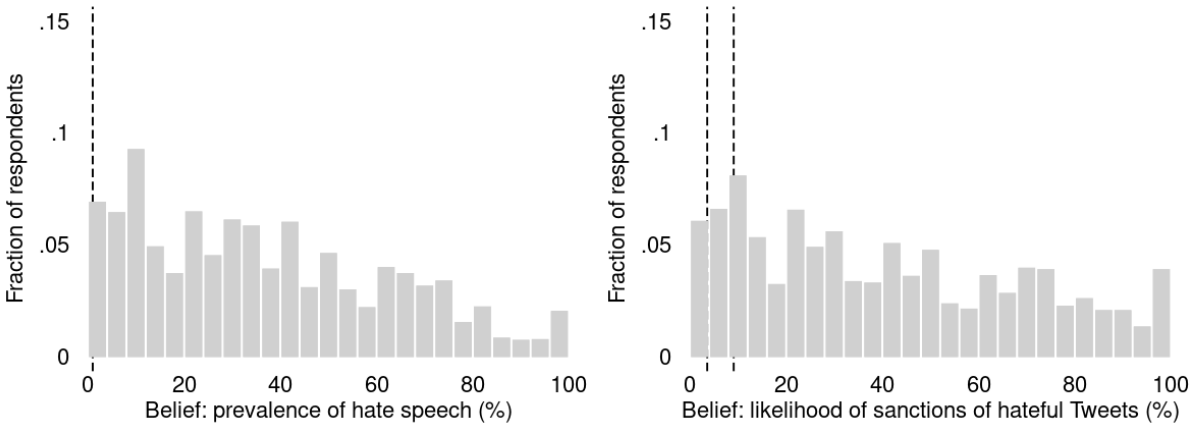
Figure A.16: Self-reported and API-based time spent on Twitter

Notes: This figure presents a comparison between the self-reported hours that participants spend on Twitter with the hours implied by their statuses and likes per day obtained through Twitter's API. The dashed line comes from a linear regression of self-reported hours on API-based hours.

Table A.13: Harassment and moderation experience by subsample

	Means			Difference t-stat.
	Survey	Minority	Not minority	Min.-Not min.
Observations	3,027	1,440	1,587	
Has been harassed	25.2	28.8	20.8	4.07
Prevalence of hate in feed	18.1	20.5	15.1	5.52
Has reported content	32.2	35.7	27.8	3.62
Has been sanctioned or reported	18.5	19.9	16.6	1.98
Tweet removal	9.6	10.4	8.8	1.33
Suspension	5.0	6.0	3.7	2.65
Shadow-ban	6.3	6.2	6.4	-0.16
Account locked	9.8	10.9	8.5	1.86
Has been reported	9.0	9.5	8.3	1.01

Notes: This table presents mean values of variables across different subsamples. It also presents t-statistics from tests of difference in means between minorities and not minorities. Observations are weighted to match representative Twitter users. Minority status Minority status depends on religion, sexual preference, gender, and race.



(a) Prevalence **(b)** Likelihood of moderation

Figure A.17: Beliefs about prevalence and moderation of hate speech

Notes: These figures present histograms of beliefs about prevalence and moderation of hate speech among survey respondents. Prevalence is the fraction of Tweets that are classified as hate speech. Likelihood of moderation is the fraction of hate speech Tweets or users that get removed or de-platformed after 1 month of posting. The dashed lines indicate the observed values of prevalence and moderation in my sample of Tweets. One line in panel (b) corresponds to the majority rule and one to the consensus rule for classifying hate speech.

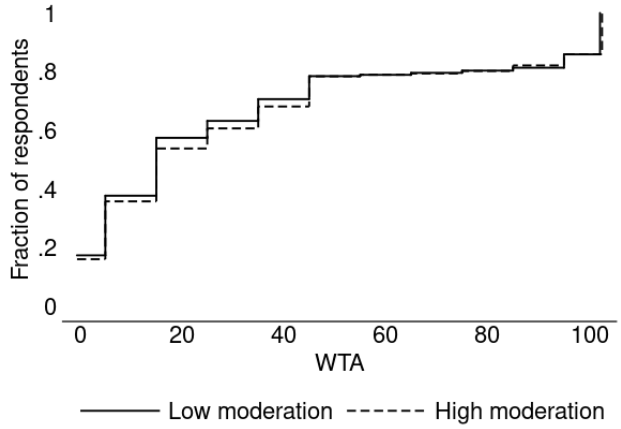


Figure A.18: CDF of the WTA to stop using social media

Notes: This figure displays the CDF of the WTA to stop using social media during one week, by treatment arm. Observations are reweighted to match Twitter users from the ATP on observables.

Table A.14: Effects of information on other measures of socia-media valuation

Panel A						
	WTA uniform distribution			WTA upper endpoint		
	(1)	(2)	(3)	(4)	(5)	(6)
High moderation	-0.150	0.024	0.024	-0.123	0.066	0.030
	(1.802)	(1.778)	(1.778)	(1.879)	(1.853)	(1.853)
y Mean	33.59	33.59	33.59	38.36	38.36	38.36
y SD	36.33	36.33	36.33	37.91	37.91	37.91
R^2	0.00	0.02	0.02	0.00	0.02	0.02
Obs.	2,998	2,998	2,998	2,998	2,998	2,998
Panel B						
	WTA heuristic			TIOLI		
	(1)	(2)	(3)	(4)	(5)	(6)
High moderation	0.276	0.630	0.382	0.011	0.009	0.010
	(2.902)	(2.855)	(2.857)	(0.020)	(0.020)	(0.020)
y Mean	36.40	36.40	36.40	0.78	0.78	0.78
y SD	58.81	58.81	58.81	0.42	0.42	0.42
R^2	0.00	0.02	0.03	0.00	0.01	0.02
Obs.	2,998	2,998	2,998	2,998	2,998	2,998
Strata FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions on treatment assignment. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,**, and *** denote significance at the 10%, 5%, and 1% levels, respectively. Observations are reweighted to match Twitter users from the ATP on observables.

Table A.15: Effects of information on other measures of activity

Panel A: Tweets and Likes

	Tweets			Likes		
	(1)	(2)	(3)	(4)	(5)	(6)
High moderation	3.488	3.359	1.567	13.480*	13.752*	8.927
	(3.733)	(3.621)	(2.623)	(7.636)	(7.356)	(6.140)
y Mean	9.64	9.64	9.64	27.97	27.97	27.97
y SD	70.91	70.91	70.91	140.25	140.25	140.25
R^2	0.00	0.02	0.67	0.00	0.04	0.40
Obs.	1,427	1,427	1,427	1,427	1,427	1,427

Panel B: other activity measures

	Winsorized time			Fraction of active days		
	(1)	(2)	(3)	(4)	(5)	(6)
High moderation	0.022	0.022	0.014	0.017	0.015	0.014
	(0.017)	(0.016)	(0.014)	(0.027)	(0.026)	(0.023)
y Mean	0.07	0.07	0.07	0.28	0.28	0.28
y SD	0.27	0.27	0.27	0.37	0.37	0.37
R^2	0.00	0.05	0.34	0.00	0.04	0.21
Obs.	1,427	1,427	1,427	1,427	1,427	1,427
Strata FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions on treatment assignment. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively. Observations are reweighted to match Twitter users from the ATP on observables.

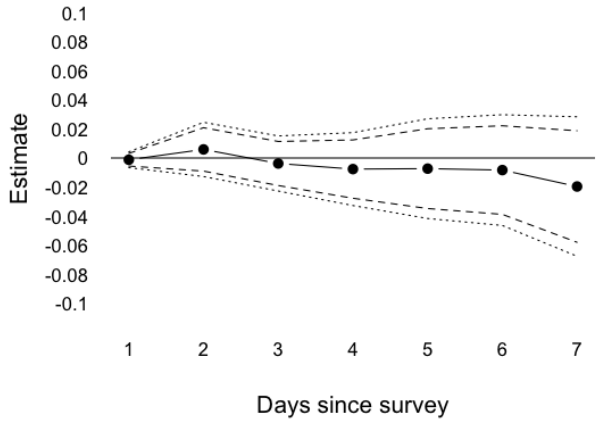


Figure A.19: Cumulative dynamic treatment effects on hours spent on Twitter

Notes: This figure presents dynamic treatment effects of hours spent one week after the survey, pointwise confidence intervals (dashed), and sup-t simultaneous confidence bands (dotted).

Table A.16: Effects of information on inattention and attrition

Panel A: Tweets and Likes

	Inattention: |recollection-info.|			Attrition		
	(1)	(2)	(3)	(4)	(5)	(6)
High moderation	1.252	1.013	1.127	0.004	0.005	0.005
	(0.898)	(0.856)	(0.820)	(0.004)	(0.004)	(0.004)
y Mean	8.90	8.90	8.90	0.01	0.01	0.01
y SD	19.82	19.82	19.82	0.10	0.10	0.10
R^2	0.00	0.12	0.24	0.00	0.01	0.01
Obs.	2,997	2,997	2,997	3,027	3,027	3,027
Strata FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions on treatment assignment. Attrition indicates whether participants who finished the prescreening questions did not finish the survey. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,**, and *** denote significance at the 10%, 5%, and 1% levels, respectively. Observations are reweighted to match Twitter users from the ATP on observables.

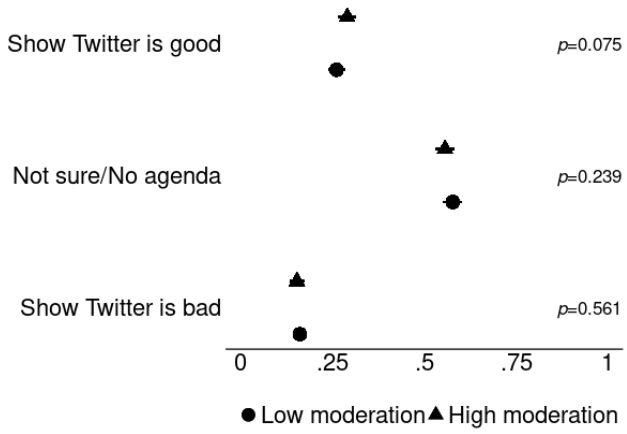
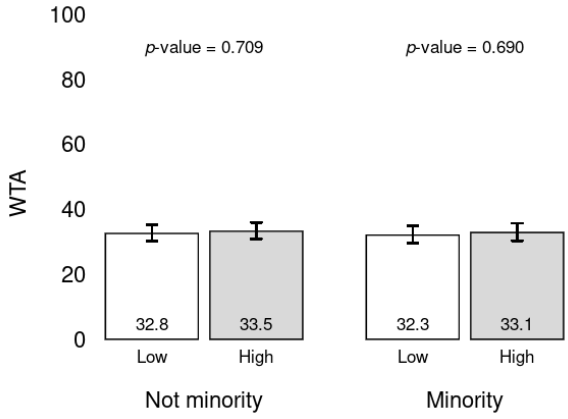
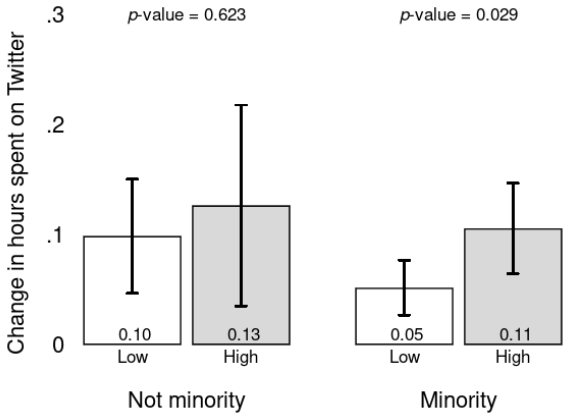


Figure A.20: Treatment effect on perceived experimenter's agenda

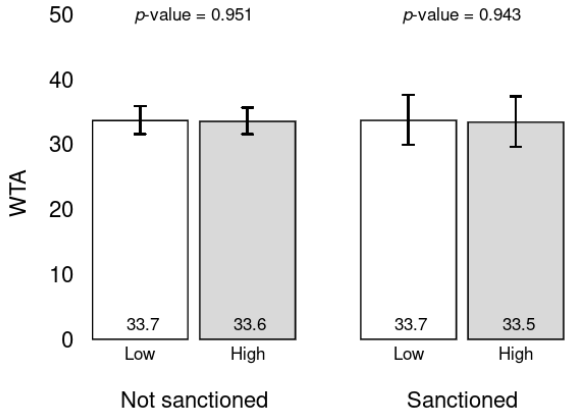
Notes: This figure presents means and 95% confidence intervals by treatment arm. The dependent variables are answers to the question "Do you think that the researchers in this study had an agenda?". The p-values come from independent OLS regressions.



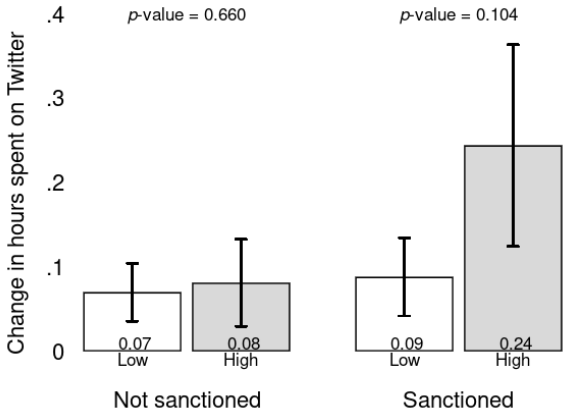
(a) WTA by minority status



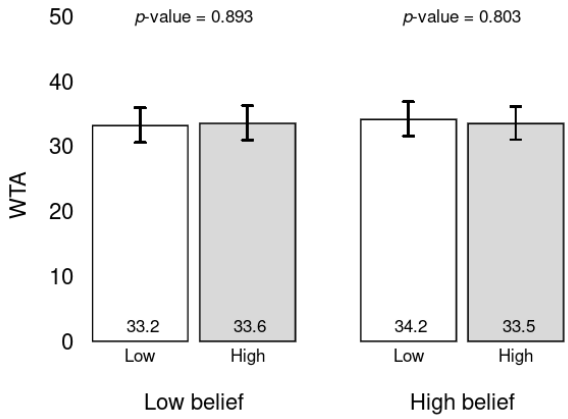
(b) Hours on Twitter by minority status



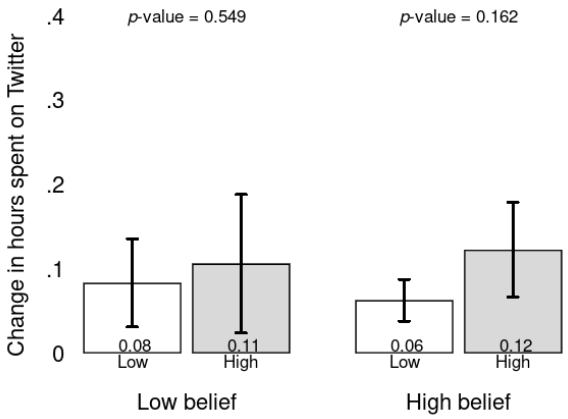
(c) WTA by previous sanctions



(d) Hours on Twitter by previous sanctions



(e) WTA by prior beliefs



(f) Hours on Twitter by prior beliefs

Figure A.21: Heterogeneity of WTA and hours on Twitter by minority status, previous sanctions, and priors

Notes: These figures present means and 95% confidence intervals by treatment arm and minority status. The p-values come from OLS regressions. Observations are reweighted to match Twitter users from the ATP on observables.

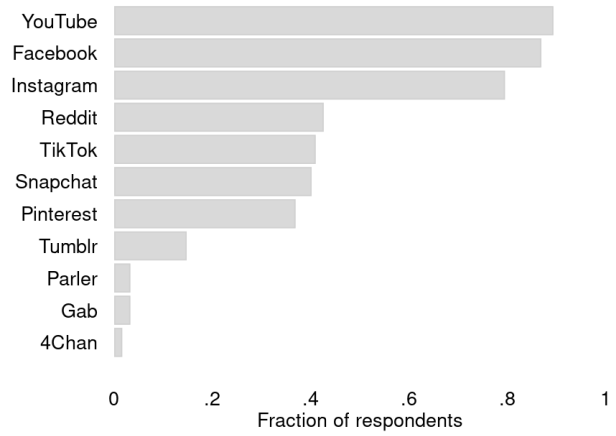


Figure A.22: Other platforms frequented by Twitter users

Notes: This figure presents the fraction of respondents who use other platforms besides Twitter.

Table A.17: Effects of information on WTA and time spent on Twitter

Panel A: Weighted (Twitter ATP)

	WTA			Time spent		
	(1)	(2)	(3)	(4)	(5)	(6)
High moderation	-0.3061	0.1035	0.1035	0.0542	0.0576*	0.0576*
	(2.234)	(2.056)	(2.056)	(0.033)	(0.034)	(0.034)
y Mean	33.57	33.57	33.57	0.10	0.10	0.10
y Std. Dev.	36.75	36.75	36.75	0.57	0.57	0.57
R^2	0.00	0.03	0.03	0.00	0.03	0.03

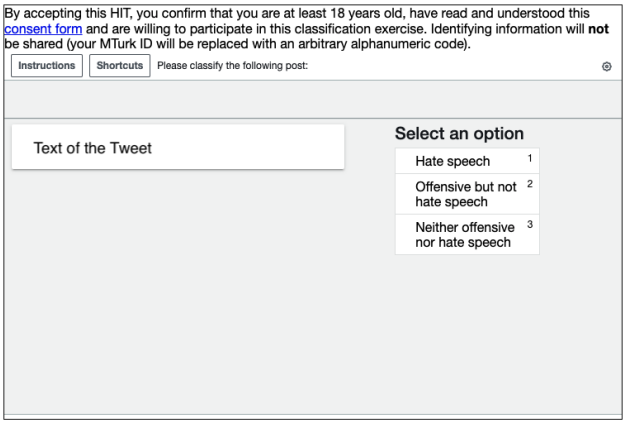
Panel B: Weighted (Social Media ATP)

	WTA			Time spent		
	(1)	(2)	(3)	(4)	(5)	(6)
High moderation	-0.0796	-0.0746	-0.0746	0.0400	0.0334	0.0334
	(2.126)	(2.093)	(2.093)	(0.042)	(0.036)	(0.036)
y Mean	34.98	34.98	34.98	0.10	0.10	0.10
y Std. Dev.	37.26	37.26	37.26	0.61	0.61	0.61
R^2	0.00	0.02	0.02	0.00	0.05	0.05
N	2998.00	2998.00	2998.00	1427.00	1427.00	1427.00

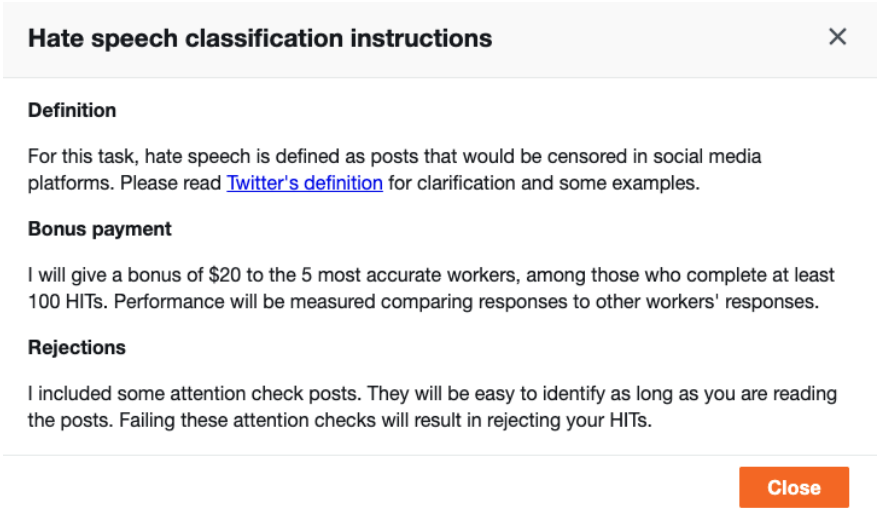
Panel C: Unweighted

	WTA			Time spent		
	(1)	(2)	(3)	(4)	(5)	(6)
High moderation	0.7230	0.7241	0.7241	0.0456	0.0461	0.0461
	(1.328)	(1.320)	(1.320)	(0.036)	(0.036)	(0.036)
y Mean	32.94	32.94	32.94	0.10	0.10	0.10
y Std. Dev.	36.35	36.35	36.35	0.68	0.68	0.68
R^2	0.00	0.02	0.02	0.00	0.01	0.01
Observations	2,998	2,998	2,998	1,427	1,427	1,427
Strata FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions of the WTA and the change in time spent on Twitter on a treatment indicator. Panel A reweights observations to match a representative sample of Twitter users on observables. Panel B reweights observations to match a representative sample of social-media users on observables. Panel C includes unweighted estimates. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016).



(a) Task screen



(b) Instructions

Figure A.23: MTurk task to classify posts as hate speech

APPENDIX B

APPENDIX TO CHAPTER 2

B.1 Theoretical Framework

In this section we derive the willingness to pay to report and solve for the equilibrium of the model. The payoff that consumers get for buying the product minus the payoff from not buying it is $u - p$, disregarding whether they report and whether they are the first ones to meet with the seller. Define $q(p) \equiv 1 - G_u(p)$ as the probability that individuals buy the product (expected demand). When individuals are the second ones to match with the seller, their report does not make a difference, so the value of reporting vs not reporting is just $-c_r$.

When they match first with the seller, the value of reporting relative to not reporting is:

$$\underbrace{\mathbb{E}(e(p)q(p))}_{\text{External payoff from a random meeting}} - \left(\underbrace{\frac{M-1}{M}\mathbb{E}(e(p)q(p))}_{\text{Random meeting}} + \underbrace{\frac{1}{M}e(p)q(p)}_{\text{The other Seller } p} \right) - c_r.$$

Intuitively, when the consumer does not report, the other consumer might meet the seller with $1/M$ chance and meets a random seller with the remaining $(M-1)/M$ chance. This past expression simplifies to:

$$\frac{1}{M}\left[\mathbb{E}(e(p)q(p)) - e(p)q(p)\right] - c_r$$

Because consumers match first with the seller with $1/2$ probability, the expected willingness to pay to report is given by Equation (2.1).

To find the equilibrium price distribution we need to find the sellers' optimal pricing strategy given consumers' reporting and buying strategies. Call $\sigma(p)$ the probability that a consumer reports price offer p after meeting. The expected profits of sellers that make price

offer p are:

$$\pi(p) = \frac{1}{M}\underbrace{\left[q(p)p - \sigma(p)\kappa + \underbrace{(1-\sigma(p))\frac{1}{M}q(p)p}_{\text{Meets with second consumer}}\right]}_{\text{Seller meets one consumer first}} + \frac{M-1}{M}\left(\frac{\mathbb{E}(\sigma(p))}{M-1} + \frac{1-\mathbb{E}(\sigma(p))}{M}\right)\underbrace{[q(p)p - \sigma(p)\kappa]}_{\text{Seller meets the second consumer}}$$

We can rewrite profits as $\pi(p) = \pi^0(p)(1-\sigma(p)) + \pi^1(p)\sigma(p)$; that is, as a linear combination between profits under a zero probability of reporting, $\pi^0(p)$, and profits when the probability of reporting is one, $\pi^1(p)$. These two functions are:

$$\begin{aligned}\pi^0(p) =& \frac{q(p)p}{M}\left(2 + \frac{\mathbb{E}(\sigma(p))}{M}\right)\\ \pi^1(p) =& \pi^0(p) - \frac{1}{M}\left(\frac{q(p)p}{M} + \kappa\left(2 - \frac{1}{M}[1-\mathbb{E}(\sigma(p))]\right)\right)\end{aligned}$$

When there is no reporting, profits are $2q(p)p/M$. Assume that second order conditions hold and let p^m be the monopolist price that maximizes this profit function. Notice that $\pi^0(p)$ and $\pi^1(p)$ are also maximized at p^m and hence $\pi^{0\prime}(p^m) = \pi^{1\prime}(p^m) = 0$. Then, it is easy to check that $\pi'(p^m) = -\sigma'(p)(\pi^0(p^m) - \pi^1(p^m)) < 0$, so profits are maximized with prices smaller than p^m.

Finally, we highlight the importance of understanding the mechanisms driving the external payoff. Consider a policy that fixes prices at level $\overline{p}$ and an alternative policy in which the government subsidizes purchases of the product for an amount $s(\overline{p})$. The government sets the subsidy such that the quantity in both cases is the same; that is, at a level equal to the monopoly price minus the controlled price $\overline{p}$. Up until now we have assumed that consumers get an external payoff $e(p)$ that is decreasing in the price of the third-party transaction. We now generalize this payoff to a new function $e(p^c, p^p)$ that captures the possibility of having

two different prices; the price paid by the other consumer p^c and the price received by the producer, p^p.

With both policies, the expected direct payoff of consumers is equal to the average utility among those who purchase the product, $2\int_{\overline{p}} udG_u u$. The aggregate external payoff, however, is $2q(\overline{p})e(\overline{p},\overline{p})$ with the price control, but $2q(\overline{p})e(\overline{p},\overline{p}+s(\overline{p}))$ with the subsidy. Because the subsidy is positive, the external payoff is lower with the subsidy if it is decreasing in the price that the seller receives. In other words, the welfare implications of both policies differ, even if they achieve the same equilibrium quantities, due to the presence of distaste for profits or markups.

B.2 Product Tracking Algorithm

To track goods and prices for our survey respondents we used the Rainforest API. It allowed us to get real-time data on availability, prices and comments on all products that are listed in the queries to "hand sanitizer" and "face mask".

The steps of the algorithm were:

1. Get the list of products that appear in the search results for the Hand Sanitizer and Face mask categories.[1]

2. Get information for each product: price, image, description, shipping date, etc.

3. Run an image classification algorithm to select which products were actually hand sanitizers and face masks

4. Process the text in the title, product description and product dimensions with regular expressions to extract and parse the number of units (fl oz, count, etc.)

1. Hand sanitizers can be found in product category 2265897011; see `https://www.amazon.com/handsanitizers/b?ie=UTF8&node=2265897011`. Likewise, face masks correspond to product catecories 6125377011, 8404646011 and 17864516011.

We collected search results on 7 dates, covering the 2 days that our survey lasted and 2 weeks before and after our experiment. We collect prices, listing titles and product images for all searches. The output from these queries included some "false positive" results, that is, not everything was truly one of the products we cared about. Since many products are advertised in multiple search categories (e.g., soaps in the hand sanitizer section), to avoid tracking and reporting incorrect items we classified 1200 results for"face mask" and 500 results for "hand sanitizer" with the help of Amazon MTurk workers to identify surgical face masks and alcohol based hand sanitizer gel. We used 3 labels to classify face masks: surgical masks, N-95 and not a mask. We used a binary label for hand sanitizer. These examples were then used to train a neural network classifier on PyTorch that used product images and text features from the product title as input to identify items of interest.

We used the pre-trained resnet50 model available in Torchvision to extract features from product images (see He et al. (2016)). To this convolutional model, we added two extra linear layers that allowed us to incorporate a vector of zeros and ones that identified the presence of particular words in the product title. The word-features used for each product model can be found in Table B.1. During the learning step, only the last linear layer of the resnet50 model and the two extra layers had their weights updated to fully take advantage of knowledge already incorporated in the pre-trained model. The trained model had an out-of-sample accuracy of 0.95 and cross-entropy loss of 0.23 for Hand Sanitizers while the respective quantities were 0.97 and 0.0957 for Masks.

Afterwards, we collected more detailed product characteristics from the filtered results, such as shipping dates, stock availability, product description and dimensions. As detailed on step 4 above, we used this information to convert prices into common units.

Table B.1: Extracted title features

Face Mask	'cloth', 'Surgical', 'Dust', 'respirator', 'dust', 'reusable'
Hand Sanitizer	'hand', 'gel', 'Purell', 'WIPES', 'TISSUES', 'paper', 'glo', 'GERM', 'lamp', 'uv', 'ULTRAVIOLET', 'IODINE', 'cotton', 'lotion', 'spray', 'air', 'holder', 'dispenser', 'soap'

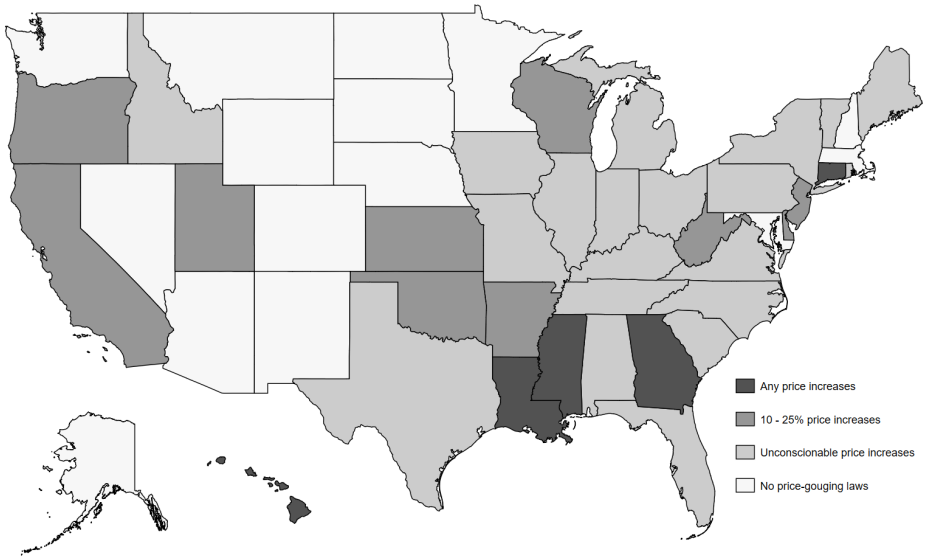


Figure B.1: Map of price gouging laws

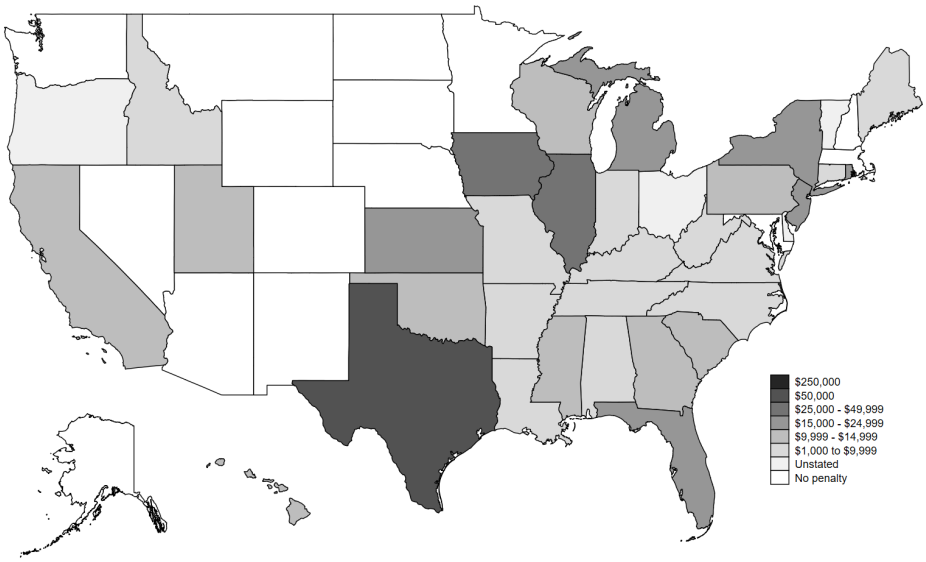


Figure B.2: Map of civil penalties for price gouging

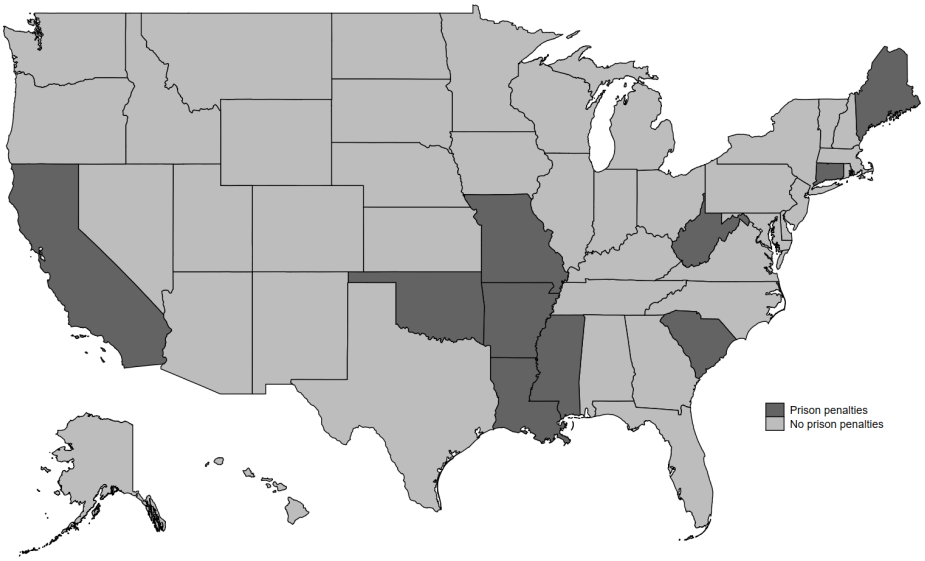
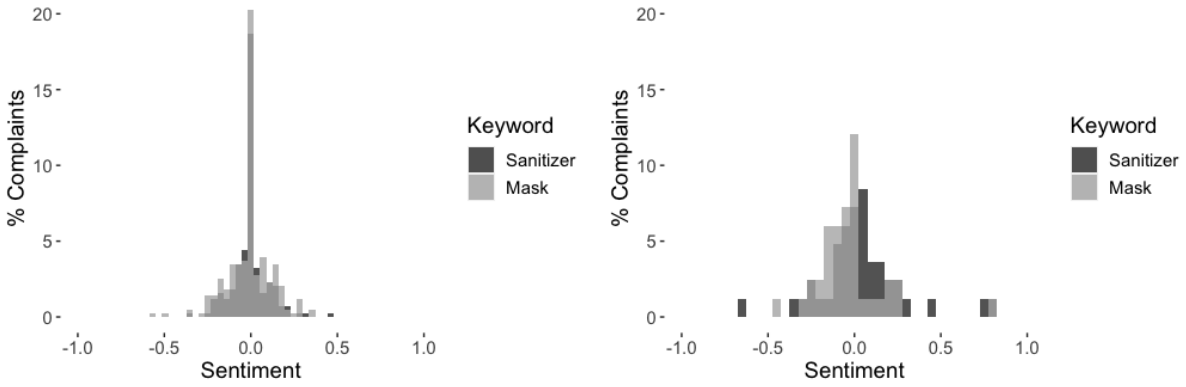


Figure B.3: Map of criminal penalties for price gouging



(a) Description **(b)** Suggested solution

Figure B.4: Distribution of sentiment in price gouging complaints

Notes: We calculate sentiment scores using the sentimentR package; see Naldi (2019) for a description and comparison with other sentiment lexicons. Sentiment ranges from -1 (negative) to 1 (positive). Mask/sanitizer complaints correspond to those that include the words 'mask' or 'sanitizer', respectively. We cannot reject the null of equality of distributions of description sentiments (Kolmogorov-Smirnov (KS) using Abadie (2002) bootstrap procedure with 10,000 resamples), with a p-value of 0.4015. Instead, we reject the null of equality of distributions of suggested solution sentiments, with a KS p-value of 0.0314. Moreover, we cannot reject the nulls of first and second order stochastic dominance (of sanitizer dominating masks) with p-values of 0.7540 and 0.6074, respectively.

Table B.2: Most frequent unigrams and bigrams in actual price gouging reports

Unigrams		Bigrams	
Description	Solution	Description	Solution
price	price	price gouge	price gouge
sell	gouge	toilet paper	stop price
gouge	fine	hand sanitizer	hold accountable
item	stop	normal price	toilet paper
paper	store	grind beef	fair price
store	people	dozen egg	gas price
egg	refund	grocery store	normal price
toilet	business	gas station	reasonable price
charge	time	oz bottle	regular price
pack	charge	paper towel	raise price
buy	low	gas price	fix income
purchase	sell	previously price	low price
mask	advantage	week ago	price increase
roll	item	lb bag	essential item
time	investigate	charmin toilet	grocery store
normal	product	mega roll	stop sell
hand	feel	raise price	gouge consumer
sanitizer	normal	regular price	gouge law
pay	crisis	covid pandemic	hand sanitizer
people	pandemic	price increase	hard time

Note: The table includes the most frequent words that appear in price gouging reports filed to the AGs of Idaho, Illinois, Missouri and Wisconsin. There are 1890 complaints in our sample (68 from ID, 102 from IL, 1271 from MO and 449 from WI). "Description" is the field where consumers detail the reason why they are submitting the complaint. "Solution" is the field where consumers express any relief/solution that they are requesting. We have solutions for 488 complaints. Missouri did not include a field to detail the requested solution. We exclude from the analysis common English stop words and lemmatize the words using the Hunspell dictionary. Unigrams denote single words and bigrams denote sequences of two adjacent words. Frequency is calculated counting occurrence across complaints.

B.3 Supplemental Evidence

B.4 Survey Instruments

B.4.1 Demographic questions

1. What is your U.S. ZIP code?

Table B.3: Willingness to pay to report is at least $5

	(1) WTPR ≥ 5	(2) WTPR ≥ 5	(3) WTPR ≥ 5	(4) WTPR ≥ 5	(5) WTPR ≥ 5
Seller Charges 27.50 to 30	0.13	0.13	0.13	0.12	0.12
	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)
Face Masks		-0.04	-0.04	-0.05	-0.05
		(0.03)	(0.03)	(0.04)	(0.04)
27.50 to 30 × Face Masks				0.02	0.02
				(0.05)	(0.05)
Constant	0.49	0.51	0.70	0.52	0.70
	(0.02)	(0.02)	(0.09)	(0.03)	(0.09)
Semi-Elasticity Estimate	0.15	0.15	0.15	0.14	0.14
Controls	NO	NO	YES	NO	YES
R-squared	0.017	0.018	0.041	0.018	0.041
Observations	1,391	1,391	1,391	1,391	1,391

Note: Table displays the effect of treatments on the probability of having a willingness to pay to report greater than or equal to five. Omitted category is hand sanitizer sold for $7.50 to $10.00. Controls include race indicators, gender indicator, age, income, education, and whether they chose to track either product, has purchased either item on Amazon in the past and whether they have Amazon Prime. Heteroskedasticity robust standard errors in parentheses.

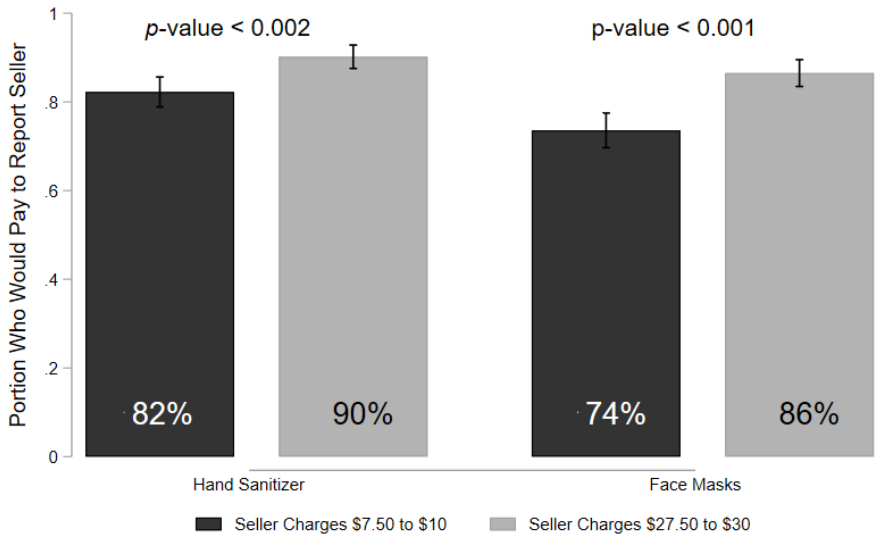
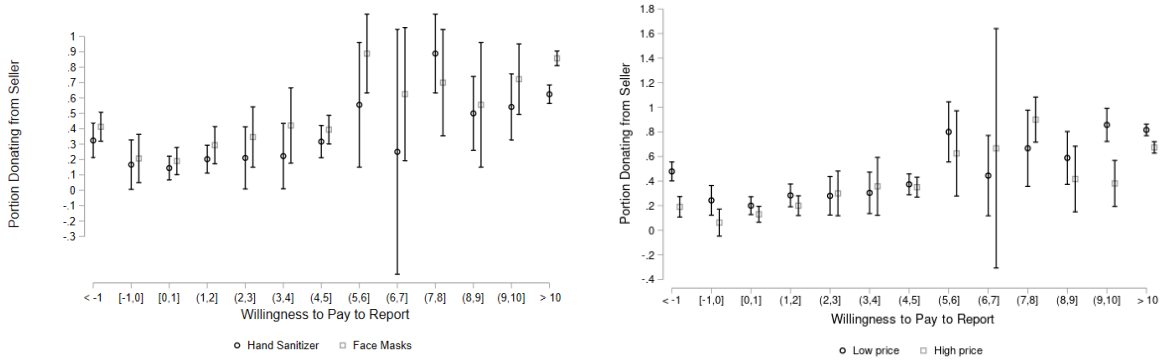


Figure B.5: Probability of choosing to report seller at any price

Notes: Figure displays the effect of treatments on the probability of choosing to report the seller at any price with 95% confidence intervals.



(a) Type of good **(b)** Seller price

Figure B.6: Relationship between willingness to report and propensity to donate

Notes: Panel (a) plots the average portion of subjects choosing to donate PPE within every willingness to report bin, by good. Panel (b) plots the average portion of subjects choosing to donate PPE within every willingness to report bin, by price.

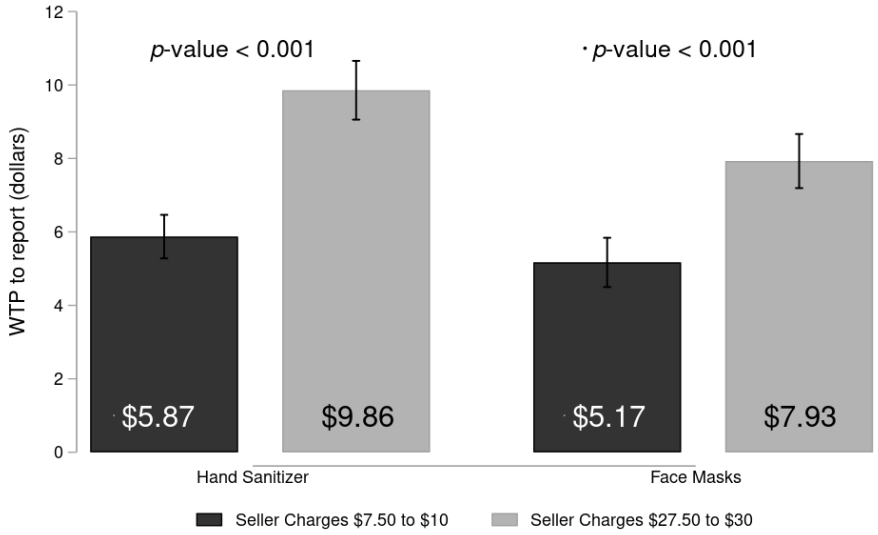


Figure B.7: Willingness to pay to report using a triangular distribution

Notes: This figure displays the average willingness to report sellers for price gouging at different prices separately by PPE type with 95% confidence intervals. We use the procedure in Allcott and Kessler (2019) to impute WTP from the results of the multiple price list. For each interior range, we assign the value of the midpoint. For the exterior unbounded ranges we assume a triangular distribution.

Table B.4: Willingness to pay to report using a triangular distribution

	(1) WTP	(2) WTP	(3) WTP	(4) WTP
Seller Charges \$27.50 to \$30	3.369	3.372	3.371	4.048
	(0.428)	(0.427)	(0.433)	(0.614)
Face Masks		-1.314	-1.364	-0.688
		(0.427)	(0.431)	(0.550)
Seller Charges \$27.50 to \$30 × Face Masks				-1.360
				(0.869)
Constant	5.521	6.177	9.857	9.552
	(0.271)	(0.332)	(1.391)	(1.386)
Elasticity Estimate	.36	.36	.36	.43
Controls	NO	NO	YES	YES
R2	0.043	0.049	0.067	0.069
Observations	1391	1391	1391	1391

Notes: This table shows regressions of individual willingness to pay to report on treatment dummies. We use the procedure in Allcott and Kessler (2019) to impute WTP from the results of the multiple price list. For each interior range, we assign the value of the midpoint. For the exterior unbounded ranges we assume a triangular distribution. Heteroskedasticity robust standard errors in parentheses. Controls include race indicators, gender indicator, age, income, education, and whether they chose to track either product, has purchased either item on Amazon in the past and whether they have Amazon Prime. Elasticity estimate calculated using the midpoint of seller price range.

2. What is your year of birth?

3. What is the highest level of school you have completed or the highest degree you have received?

    - Less than high school degree
    - High school graduate (high school diploma or equivalent including GED)
    - Some college but no degree
    - Associate degree in college (2 year)
    - Bachelor's degree in college (4 year)
    - Master's degree
    - Doctoral degree

Table B.5: Propensity to donate by WTPR

	(1) Donate	(2) Donate	(3) Donate	(4) Donate
WTPR $\geq$ 5	0.36	0.38	0.35	0.35
	(0.02)	(0.02)	(0.04)	(0.04)
Seller Charges 27.50 to 30		-0.11	-0.18	-0.19
		(0.02)	(0.04)	(0.04)
Face Masks		0.13	0.08	0.08
		(0.02)	(0.04)	(0.04)
Seller Charges 27.50 to 30 $\times$ Face Masks			0.09	0.09
			(0.05)	(0.05)
Seller Charges 27.50 to 30 $\times$ WTPR $\geq$ 5			0.05	0.06
			(0.05)	(0.05)
Constant	0.27	0.25	0.28	0.18
	(0.02)	(0.02)	(0.03)	(0.08)
Controls	NO	NO	NO	YES
R-squared	0.128	0.157	0.159	0.173
Observations	1,386	1,386	1,386	1,386

Notes: Table displays the effect of treatments on the propensity to donate along with the correlation between having a WTPR exceeding \$5 and donating. Omitted category is hand sanitizer sold for \$7.50 to \$10.00. Controls include race indicators, gender indicator, age, income, education, and whether they chose to track either product, has purchased either item on Amazon in the past and whether they have Amazon Prime. State laws is an indicator equal to 1 if the subject's state has laws against price gouging. Heteroskedasticity robust standard errors in parentheses.

- Professional degree (JD, MD)

4. Choose one or more races/ethnicities that you consider yourself to be:

   - White or European American
   - Black or African American
   - Hispanic or Latino
   - Asian or Asian American
   - Other:

5. What is your approximate household annual income? Please indicate the answer that includes your entire household income in 2019 before taxes

Table B.6: Treatment effect on attention

	(1) Attention	(2) Attention	(3) Attention	(4) Attention
Seller Charges 27.50 to 30	-0.20	-0.20	-0.19	-0.17
	(0.02)	(0.02)	(0.02)	(0.03)
Face Masks		0.00	-0.00	0.02
		(0.02)	(0.02)	(0.02)
Seller Charges 27.50 to 30 × Face Masks				-0.04
				(0.04)
Constant	0.94	0.94	0.80	0.79
	(0.01)	(0.01)	(0.06)	(0.06)
Controls	NO	NO	YES	YES
R-squared	0.076	0.076	0.109	0.110
Observations	1,391	1,391	1,391	1,391

Notes: Table displays the effect of treatments on the propensity correctly answer the attention question. Omitted category is hand sanitizer sold for $7.50 to $10.00. Controls include race indicators, gender indicator, age, income, education, and whether they chose to track either product, has purchased either item on Amazon in the past and whether they have Amazon Prime. State laws is an indicator equal to 1 if the subject's state has laws against price gouging. Heteroskedasticity robust standard errors in parentheses.

- Less than $10,000
- $10,000 to $19,999
- $20,000 to $29,999
- $30,000 to $39,999
- $40,000 to $49,999
- $50,000 to $59,999
- $60,000 to $69,999
- $70,000 to $79,999
- $80,000 to $89,999
- $90,000 to $99,999
- $100,000 to $149,999
- $150,000 or more

Table B.7: Treatment effect on attentive subjects

	(1) WTPR	(2) WTPR	(3) Donate	(4) Donate
Seller Charges 27.50 to 30	-1.28	1.79	-0.09	-0.11
	(0.93)	(0.38)	(0.11)	(0.04)
Face Masks	0.29	-0.95	0.12	0.06
	(1.28)	(0.36)	(0.17)	(0.04)
Seller Charges 27.50 to 30 × Face Masks	-0.42	0.35	-0.01	0.12
	(1.44)	(0.54)	(0.18)	(0.06)
Constant	6.21	5.16	0.42	0.47
	(0.79)	(0.26)	(0.10)	(0.03)
Attentive	NO	YES	NO	YES
Elastcity Estimate	-0.798	1.117		
R-squared	0.017	0.051	0.017	0.019
Observations	219	1,172	214	1,172

Notes: Table displays the effect of treatments on the WTP to report and propensity to donate. Omitted category is hand sanitizer sold for $7.50 to $10.00. Odd Columns include the full sample of subjects. Even columns drop subjects who answered the attention question incorrectly. Controls include race indicators, gender indicator, age, income, education, and whether they chose to track either product, has purchased either item on Amazon in the past and whether they have Amazon Prime. State laws is an indicator equal to 1 if the subject's state has laws against price gouging. Heteroskedasticity robust standard errors in parentheses.

6. What is your sex? Male/Female

7. Have you purchased anything on Amazon in the last month? Yes/No

8. Do you have Amazon Prime? Yes/No

9. Have you bought online or in stores any of the following in 2020? Please select all that apply:

    - Hand sanitizer
    - Face masks
    - None of the above

Table B.8: Treatment effect on higher quality belief

	(1) Higher Quality	(2) Higher Quality	(3) Higher Quality	(4) Higher Quality
Seller Charges 27.50 to 30	-0.03	-0.03	-0.04	-0.06
	(0.02)	(0.02)	(0.02)	(0.03)
Face Masks		0.02	0.02	0.00
		(0.02)	(0.02)	(0.03)
27.50 to 30 × Face Masks				0.04
				(0.04)
Constant	0.23	0.22	0.32	0.33
	(0.02)	(0.02)	(0.07)	(0.07)
Controls	NO	NO	YES	YES
R-Squared	0.002	0.002	0.123	0.123
Observations	1,391	1,391	1,391	1,391

Notes: Table displays the effect of treatments on the propensity to claim that higher priced PPE is higher quality. Omitted category is hand sanitizer sold for $7.50 to $10.00. Controls include race indicators, gender indicator, age, income, education, and whether they chose to track either product, has purchased either item on Amazon in the past and whether they have Amazon Prime. State laws is an indicator equal to 1 if the subject's state has laws against price gouging. Heteroskedasticity robust standard errors in parentheses.

B.4.2 Quality/attention check questions

1. At which prices did we say we will buy and donate the product?

    - Between $7.50 and $10
    - Between $27.50 and $30

2. Do you think that $50 face masks or hand sanitizers have a higher quality than $5 ones? Yes/No

Table B.9: Treatment effect on subjects who think quality increases with price

	(1) WTPR	(2) WTPR	(3) Donate	(4) Donate
Seller Charges 27.50 to 30	1.73	-0.96	-0.11	-0.12
	(0.39)	(0.75)	(0.04)	(0.08)
Face Masks	-0.83	-1.20	0.10	-0.04
	(0.40)	(0.68)	(0.04)	(0.08)
Seller Charges 27.50 to 30 × Face Masks	0.13	1.74	0.06	0.20
	(0.56)	(1.04)	(0.06)	(0.12)
Constant	5.06	5.81	0.45	0.52
	(0.29)	(0.48)	(0.03)	(0.06)
Views More Expensive Goods as Higher Quality	NO	YES	NO	YES
Elastcity Estimate	4.406	-2.445		
R-Squared	0.043	0.012	0.024	0.012
Observations	1,093	298	1,088	298

Notes: Table displays the effect of treatments on the WTP to report and propensity to donate. Omitted category is hand sanitizer sold for $7.50 to $10.00. Odd columns include the full sample of subjects. Even columns drop subjects who answered the attention question incorrectly. Controls include race indicators, gender indicator, age, income, education, and whether they chose to track either product, has purchased either item on Amazon in the past and whether they have Amazon Prime. State laws is an indicator equal to 1 if the subject's state has laws against price gouging. Heteroskedasticity robust standard errors in parentheses.

Table B.10: Treatment effect by whether subjects found the price excessive

	(1) WTPR	(2) WTPR	(3) Donate	(4) Donate
Seller Charges 27.50 to 30	1.97	0.58	-0.08	-0.14
	(0.53)	(0.46)	(0.06)	(0.05)
Face Masks	-1.51	-0.86	0.00	0.08
	(0.62)	(0.43)	(0.07)	(0.05)
Seller Charges 27.50 to 30 × Face Masks	-0.56	1.12	0.10	0.11
	(0.86)	(0.61)	(0.09)	(0.07)
Constant	4.97	5.46	0.45	0.48
	(0.38)	(0.33)	(0.04)	(0.04)
Elastcity Estimate	0.260	0.068	-0.011	-0.016
R2	0.066	0.023	0.007	0.026
Observations	496	895	492	894

Notes: Heteroskedasticity robust standard errors in parentheses. Controls include race indicators, gender indicator, age, income, education, and whether they chose to track either product, has purchased either item on Amazon in the past and whether they have Amazon Prime. Elasticity estimate calculated using the midpoint of seller price range.

Table B.11: Willingness to pay to report by deaths (above median)

	(1) WTPR	(2) WTPR	(3) WTPR	(4) WTPR	(5) WTPR
Seller Charges 27.50 to 30	1.59	1.62	1.55	1.86	1.86
	(0.34)	(0.34)	(0.34)	(0.38)	(0.39)
deaths_abovemed	0.13	0.16	0.16	0.13	0.16
	(0.35)	(0.35)	(0.36)	(0.35)	(0.36)
Seller Charges 27.50 to 30 X High Deaths	-0.36	-0.41	-0.31	-0.38	-0.34
	(0.49)	(0.49)	(0.50)	(0.49)	(0.50)
Seller Charges 27.50 to 30 × Face Masks				-0.52	-0.59
				(0.35)	(0.36)
Constant	4.71	5.06	6.72	4.71	6.71
	(0.24)	(0.27)	(0.81)	(0.24)	(0.81)
Elasticity Estimate	0.20	0.20	0.19	0.23	0.23
Controls	NO	NO	YES	NO	YES
R-Squared	0.024	0.030	0.041	0.025	0.043
Observations	1,391	1,391	1,391	1,391	1,391

Notes: Table displays the effect of treatments on the willingness to pay to report. Omitted category is hand sanitizer sold for $7.50 to $10.00. Controls include race indicators, gender indicator, age, income, education, and whether they chose to track either product, has purchased either item on Amazon in the past and whether they have Amazon Prime. The outcome is the number of deaths due to Covid in the subject's state by the date of the experiment. Heteroskedasticity robust standard errors in parentheses.

Table B.12: Willingness to pay to report by deaths

	(1) WTPR	(2) WTPR	(3) WTPR	(4) WTPR	(5) WTPR
Seller Charges 27.50 to 30	1.34	1.36	1.28	1.58	1.56
	(0.31)	(0.31)	(0.31)	(0.36)	(0.37)
Deaths per Thousand	-0.13	-0.12	-0.12	-0.13	-0.13
	(0.67)	(0.68)	(0.69)	(0.67)	(0.69)
27.50 to 30 X Deaths per Thousand	0.11	0.01	0.26	0.05	0.19
	(0.92)	(0.92)	(0.93)	(0.92)	(0.93)
27.50 to 30 × Face Masks				-0.44	-0.53
				(0.35)	(0.36)
Constant	4.81	5.14	6.93	4.81	6.93
	(0.22)	(0.25)	(0.82)	(0.22)	(0.82)
Elasticity Estimate	0.17	0.17	0.16	0.19	0.19
Controls	NO	NO	YES	NO	YES
R-Squared	0.022	0.027	0.040	0.023	0.042
Observations	1,370	1,370	1,370	1,370	1,370

Notes: Table displays the effect of treatments on the willingness to pay to report. Omitted category is hand sanitizer sold for $7.50 to $10.00. Controls include race indicators, gender indicator, age, income, education, and whether they chose to track either product, has purchased either item on Amazon in the past and whether they have Amazon Prime. Deaths per 1000 is the number of deaths due to covid in the subject's state by the date of the experiment per 1000 people in the state. Heteroskedasticity robust standard errors in parentheses.

Table B.13: Willingness to pay to report by state law

	(1) WTPR	(2) WTPR	(3) WTPR	(4) WTPR	(5) WTPR
Seller Charges 27.50 to 30	1.07	1.00	0.90	1.27	1.15
	(0.69)	(0.69)	(0.71)	(0.71)	(0.73)
Price-Gouging is Illegal	-0.31	-0.36	-0.47	-0.31	-0.47
	(0.50)	(0.49)	(0.50)	(0.50)	(0.50)
Seller Charges 27.50 to 30 X Price-Gouging is Illegal	0.35	0.44	0.52	0.37	0.54
	(0.74)	(0.74)	(0.76)	(0.74)	(0.76)
Seller Charges 27.50 to 30 × Face Masks				-0.44	-0.54
				(0.35)	(0.36)
Constant	5.04	5.42	7.32	5.04	7.32
	(0.46)	(0.48)	(0.93)	(0.46)	(0.93)
Elasticity Estimate	0.13	0.12	0.11	0.16	0.14
Controls	NO	NO	YES	NO	YES
R-Squared	0.022	0.027	0.042	0.023	0.044
Observations	1,367	1,367	1,367	1,367	1,367

Notes: Table displays the effect of treatments on the willingness to pay to report. Omitted category is hand sanitizer sold for \$7.50 to \$10.00. Controls include race indicators, gender indicator, age, income, education, and whether they chose to track either product, has purchased either item on Amazon in the past and whether they have Amazon Prime. State laws is an indicator equal to 1 if the subject's state has laws against price gouging. Heteroskedasticity robust standard errors in parentheses.

Table B.14: Propensity to donate by deaths

	(1) Donation	(2) Donation	(3) Donation	(4) Donation	(5) Donation
Seller Charges 27.50 to 30	-0.07	-0.07	-0.07	-0.16	-0.16
	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)
Deaths per Thousand	0.05	0.05	0.04	0.05	0.04
	(0.07)	(0.07)	(0.08)	(0.07)	(0.08)
27.50 to 30 × Deaths per Thousand	0.00	0.02	0.02	0.02	0.05
	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)
27.50 to 30 × Face Masks				0.17	0.16
				(0.04)	(0.04)
Constant	0.49	0.43	0.45	0.49	0.45
	(0.02)	(0.03)	(0.09)	(0.02)	(0.09)
Semi-Elasticity Estimate	-0.08	-0.08	-0.08	-0.18	-0.19
Controls	NO	NO	YES	NO	YES
R-Squared	0.005	0.019	0.023	0.019	0.036
Observations	1,365	1,365	1,365	1,365	1,365

Notes: Table displays the effect of treatments on the propensity to donate. Omitted category is hand sanitizer sold for $7.50 to $10.00. Controls include race indicators, gender indicator, age, income, education, and whether they chose to track either product, has purchased either item on Amazon in the past and whether they have Amazon Prime. Deaths per 1000 is the number of deaths due to covid in the subject's state by the date of the experiment per 1000 people in the state. Heteroskedasticity robust standard errors in parentheses.

Table B.15: Propensity to donate by deaths (above median)

	(1) Donate	(2) Donate	(3) Donate	(4) Donate	(5) est5
Seller Charges 27.50 to 30	-0.04	-0.05	-0.05	-0.13	-0.13
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Face Masks		0.11			
		(0.03)			
Seller Charges 27.50 to 30 × Face Masks				0.16	0.16
				(0.04)	(0.04)
Constant	0.46	0.41	0.42	0.46	0.42
	(0.03)	(0.03)	(0.09)	(0.03)	(0.09)
Semi-Elasticity Estimate	-0.05	-0.06	-0.06	-0.15	-0.15
Controls	NO	NO	YES	NO	YES
R-Squared	0.007	0.019	0.024	0.020	0.036
Observations	1,386	1,386	1,386	1,386	1,386

Notes: Table displays the effect of treatments on the propensity to donate. Omitted category is hand sanitizer sold for $7.50 to $10.00. Controls include race indicators, gender indicator, age, income, education, and whether they chose to track either product, has purchased either item on Amazon in the past and whether they have Amazon Prime. High deaths is an indicator equal to one if the number of deaths due to covid in the subject's state by the date of the experiment per 1000 people in the state is above the median for the whole country. Heteroskedasticity robust standard errors in parentheses.

Table B.16: Propensity to donate by state law

	(1) Donation	(2) Donation	(3) Donation	(4) Donation	(5) Donation
Seller Charges 27.50 to 30	-0.05	-0.03	-0.04	-0.12	-0.12
	(0.07)	(0.07)	(0.08)	(0.08)	(0.08)
Price-Gouging Illegal	-0.06	-0.05	-0.06	-0.06	-0.06
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
27.50 to 30 X Price-Gouging Illegal	-0.02	-0.04	-0.02	-0.03	-0.03
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
27.50 to 30 × Face Masks				0.17	0.16
				(0.04)	(0.04)
Constant	0.55	0.48	0.51	0.55	0.51
	(0.05)	(0.05)	(0.10)	(0.05)	(0.10)
Semi-Elasticity Estimate	-0.06	-0.04	-0.05	-0.15	-0.14
Controls	NO	NO	YES	NO	YES
R-Squared	0.006	0.020	0.024	0.020	0.037
Observations	1,362	1,362	1,362	1,362	1,362

Notes: Table displays the effect of treatments on the propensity to donate. Omitted category is hand sanitizer sold for $7.50 to $10.00. Controls include race indicators, gender indicator, age, income, education, and whether they chose to track either product, has purchased either item on Amazon in the past and whether they have Amazon Prime. State laws is an indicator equal to 1 if the subject's state has laws against price gouging. Heteroskedasticity robust standard errors in parentheses.

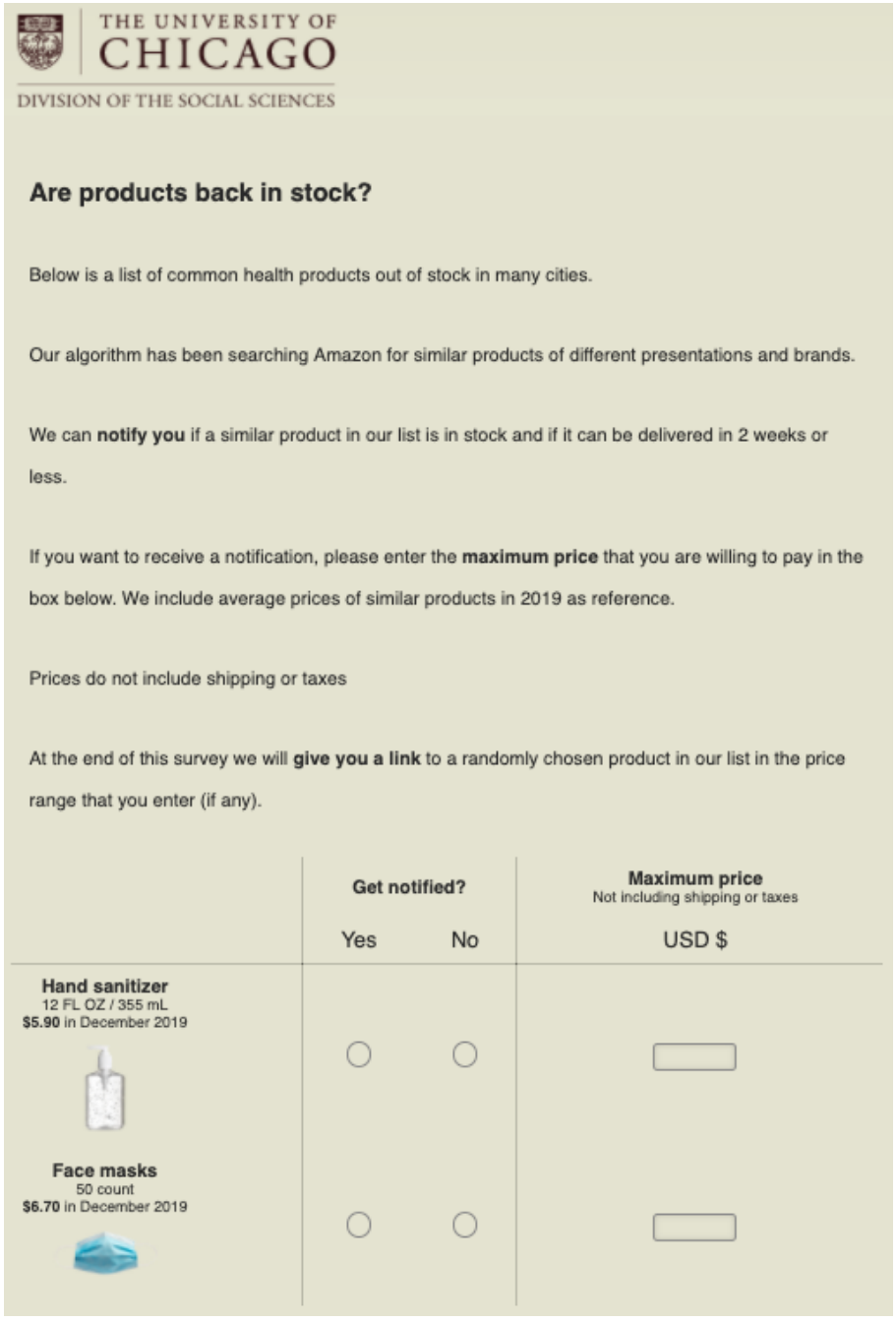


Figure B.8: Willingness to track the items

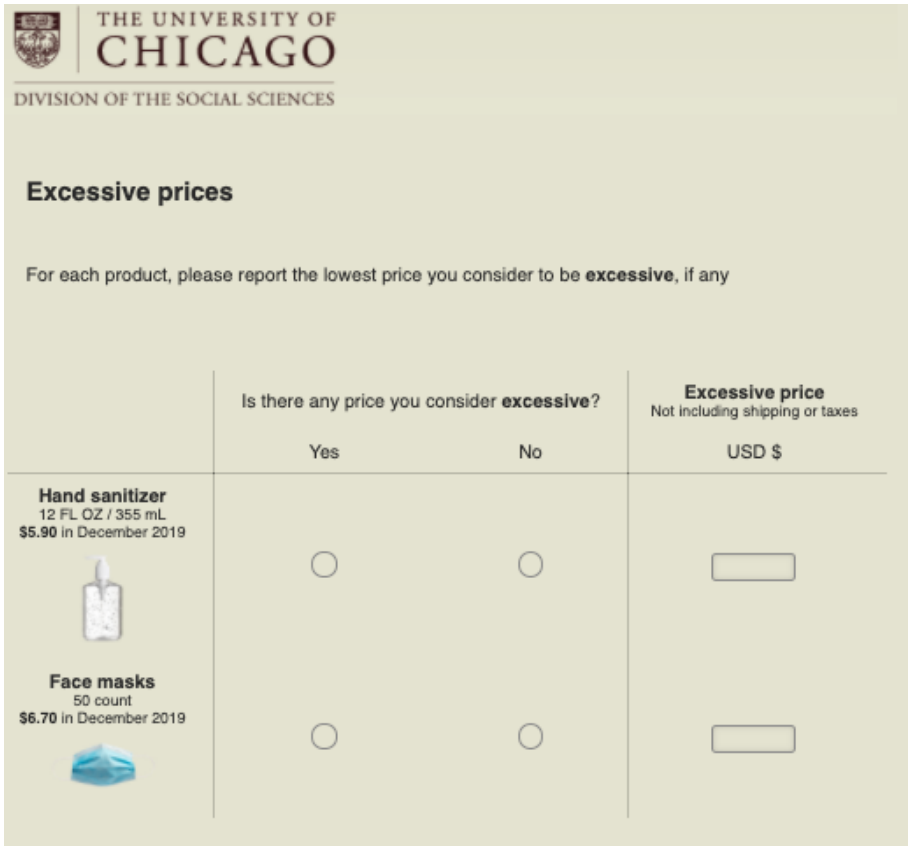
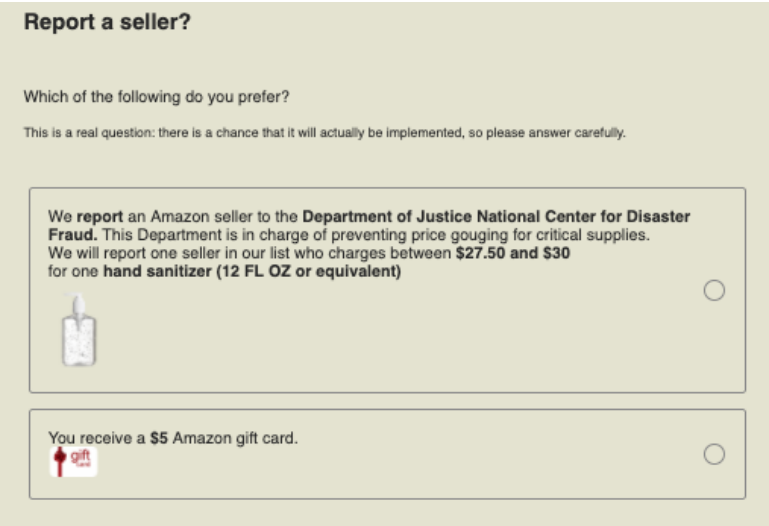


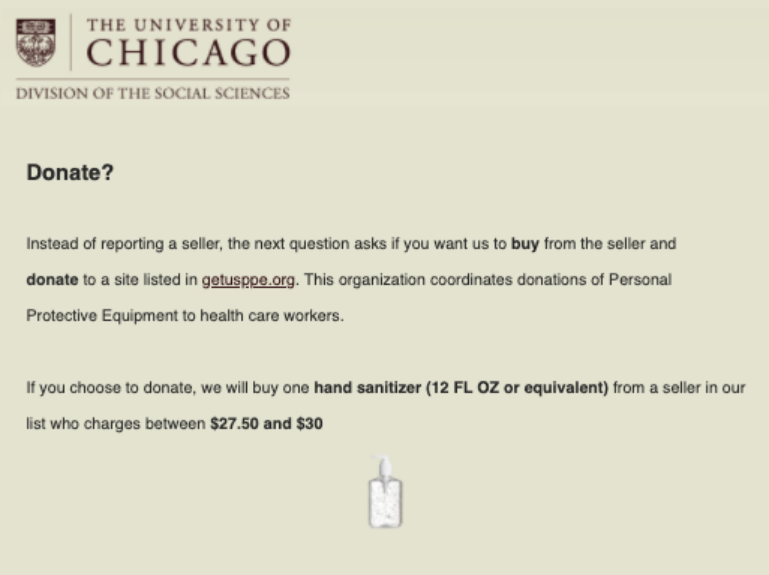
Figure B.9: Excessive prices

(a) Instructions

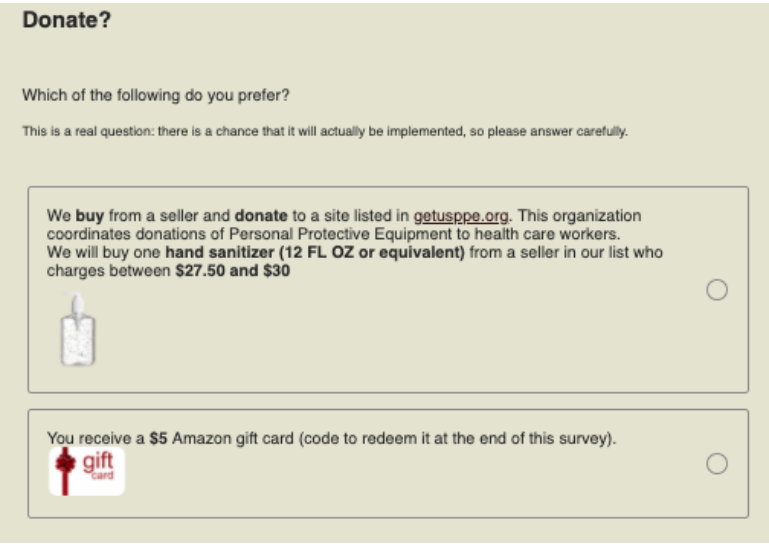


(b) Main question

Figure B.10: Elicitation of willingness to pay to report



(a) Instructions



(b) Main question

Figure B.11: Donation decision

APPENDIX C

APPENDIX TO CHAPTER 3

C.1 Welfare Cost of Taxing Cash: Alternative Interpretation

In this appendix, we give an alternative interpretation to the benchmark policy of Section 3.4.1 of taxing cash at rate τ and giving the proceeds as a lump-sum rebate ϱ. That benchmark policy is equivalent to a budget-neutral policy of a tax rate $\bar{\tau}$ to cash transactions and a subsidy to credit transactions at rate $\bar{s}$. To simplify the analysis we develop this equivalence for the case where there is no heterogeneity in the cash shares across goods, so that $\mathcal{A}$ has one element, namely $\bar{\alpha}$. In this case $\bar{\tau}$ and $\bar{s}$ satisfy:

$$1+\tau=\frac{1+\bar{\tau}}{1-\bar{s}} \text{ and } \bar{\tau}\, a_{\bar{\alpha}}=\bar{s}\, c_{\bar{\alpha}} \tag{C.1}$$

where $a_{\bar{\alpha}}$ and $c_{\bar{\alpha}}$ are cash and credit choices that corresponds to the original problem with tax in cash rebated lump-sum. To see that the equivalence, note that the first expression in (C.1) states that the two tax systems correspond to the same marginal rate of substitution, and that the second equation ensures that if feasibility, given by $a_{\bar{\alpha}}+c_{\bar{\alpha}}=y$, is satisfied in one case, then it is satisfied for the other tax system.

Figure C.1 plots the implied values of the budget neutral tax-subsidy $(\bar{\tau},\bar{s})$ equivalent to a range of values of our benchmark tax on cash $\tau=0.4$. The figure displays two set of taxes, with an elasticity of substitution $\eta=5$, and one pair $(\bar{\tau},\bar{s})$ for the share of credit $\bar{\alpha}=0.2$ and one for $\bar{\alpha}=0.4$.

Figure C.1: Budget neutral policy that taxes cash and subsidizes credit

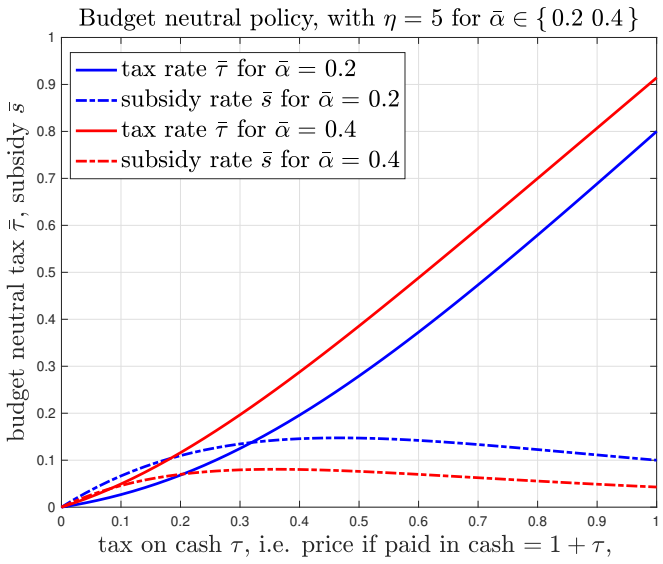
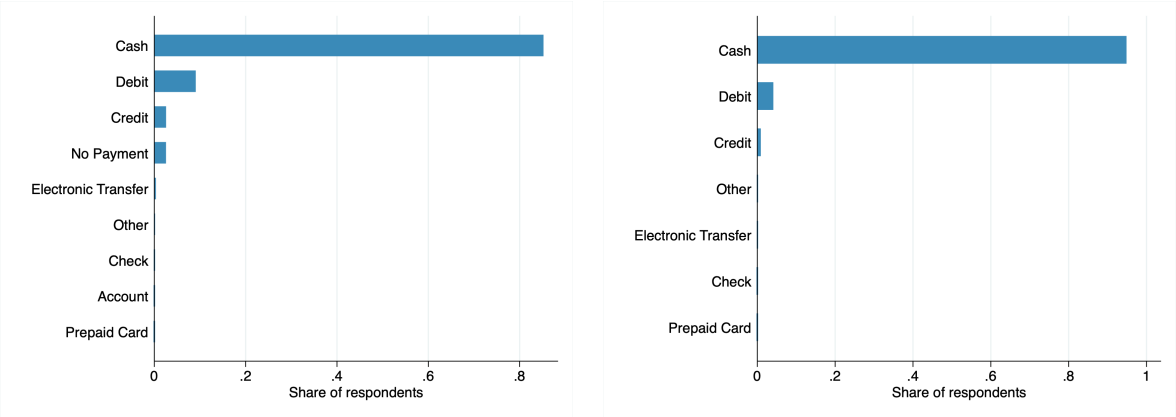


Figure C.2: Payment method by Amount

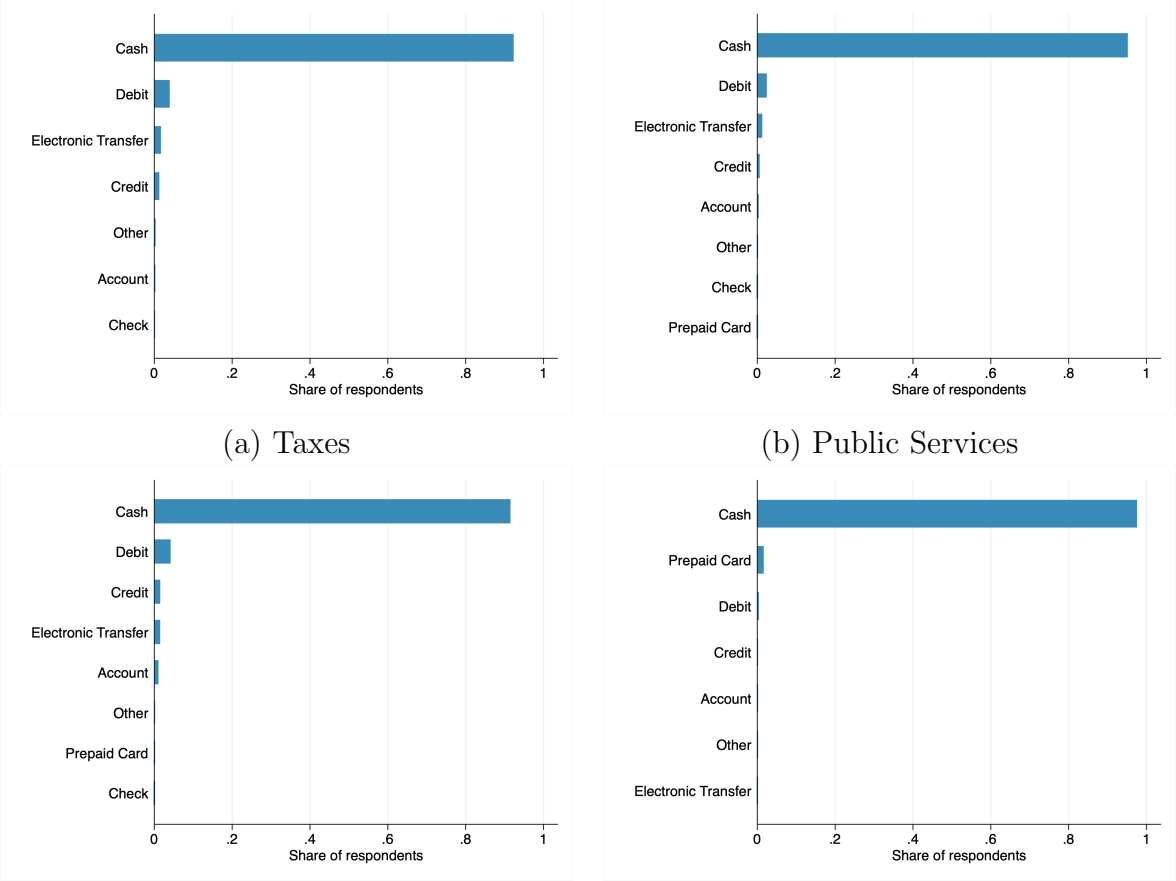


(a) 20 USD or More

(b) 20 USD or Less

Note: The figure shows the most frequent payment methods reported by households for payments 20 USD or more and for 20 USD or less. The data comes from the 2018 National Survey of Financial Inclusion (ENIF).

Figure C.3: Payment Method by Sector



(a) Taxes

(b) Public Services

(c) Private Services

(d) Transportation

Note: The figure shows the most frequent payment methods reported by households for different types of expenditures. The panels report expenditures on taxes, expenditures on public services (e.g. water, electricity), private services (e.g. cable, phone, internet), and transportation (e.g. taxi, bus). The data comes from the 2018 National Survey of Financial Inclusion (ENIF).

C.2 Figures and Tables

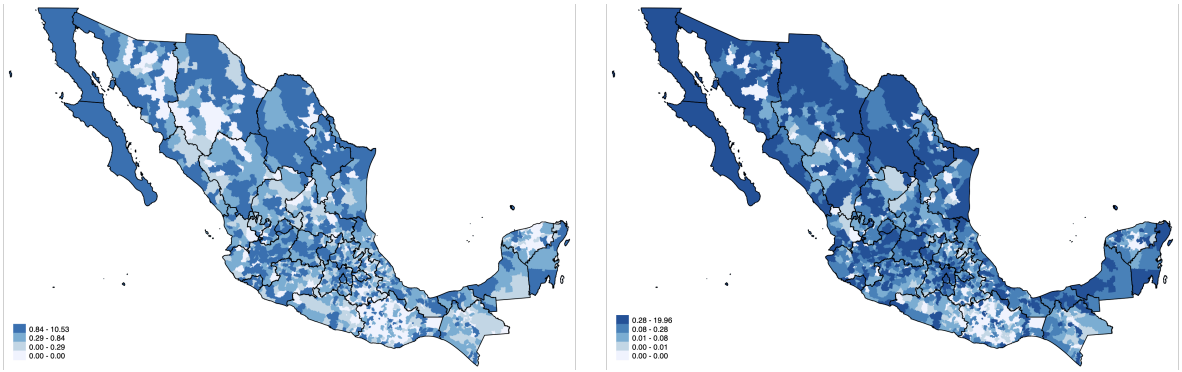
C.3 Card Shock: Semi-Dynamic Event Study

C.3.1 Card Shock: Share of Prospera Beneficiaries and in the Rollout

In this section we use the same specification used in Section C.3 but we consider the intensity of the treatment in the dependent variable. Specifically, we estimate the following equation:

$$\ln Y_{mt} = \alpha + \beta \ln \text{ShareProgresaRollout}_{mt} + \theta_m + \lambda_t + \zeta X_{mt} + \epsilon_{mt}, \tag{C.2}$$

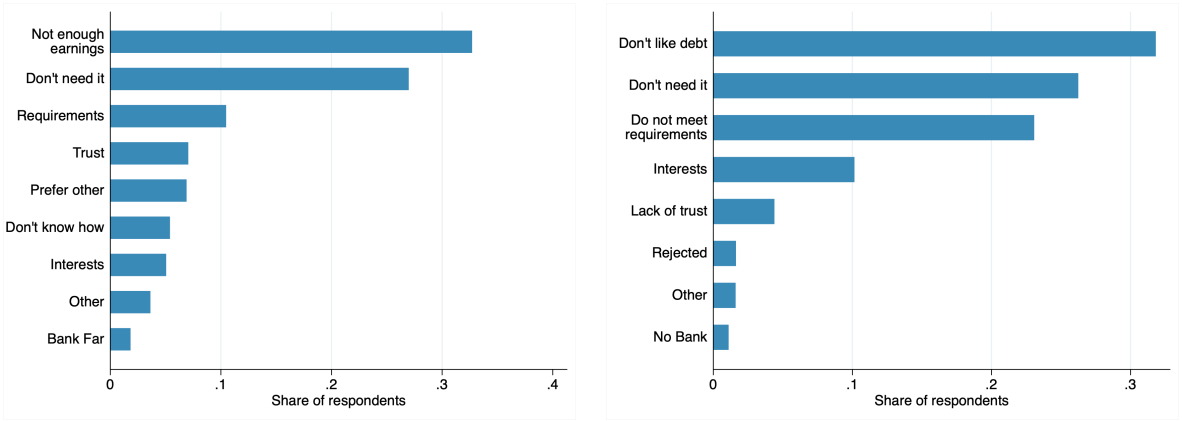
Figure C.4: Access to Financial Infrastructure



(a) ATM Transactions per Capita (b) POS Transactions per Capita

Note: Figure maps the number of ATM transactions per inhabitant and the number of POS transaction per inhabitant by municipality. Darker colors represent a higher numbers per capita. Data come from the Financial Inclusion Databases from the National Banking and Securities Commission (BDIF).

Figure C.5: Cash Users That Do Not Own an Account or a Card

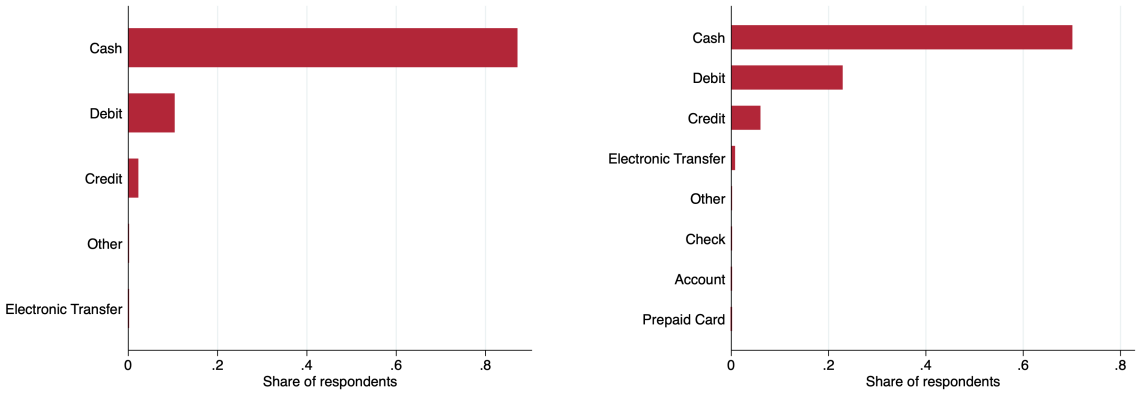


(a) Reason Not to Have an Account (b) Reason Not to Have a Card

Note: Panel (a) shows the most frequent reasons mentioned by households for not having an bank account. Panel (b) shows the most frequent reasons mentioned by households for not having a card. The data comes from the 2018 National Survey of Financial Inclusion (ENIF).

where $\text{ShareProsperaRollout}_{mt}$ is the ratio of the households participating in Prospera and in the rollout divided by the total number of households in the municipality.

Figure C.6: Payment method by Amount - Mixed Users



(a) 20 USD or Less (b) 20 USD or More

Note: The figure shows the most frequent payment methods reported by households for payments 20 USD or more and for 20 USD or less. The sample of households report owning a debit or a credit card. The data comes from the 2018 National Survey of Financial Inclusion (ENIF).

Table C.1: Share of Firms that Accept Debit Cards as Payment Method

Note: The table shows the share of firms that accept debit cards as payment method. Each cell indicates the share of firms within that cell that accept debit cards. The size bins are defined by the total number of employees: Micro (6-10), small (11-30), medium (30-100), large (100+). The data comes from the 2018 National Survey of Enterprise Financing (ENAFIN).

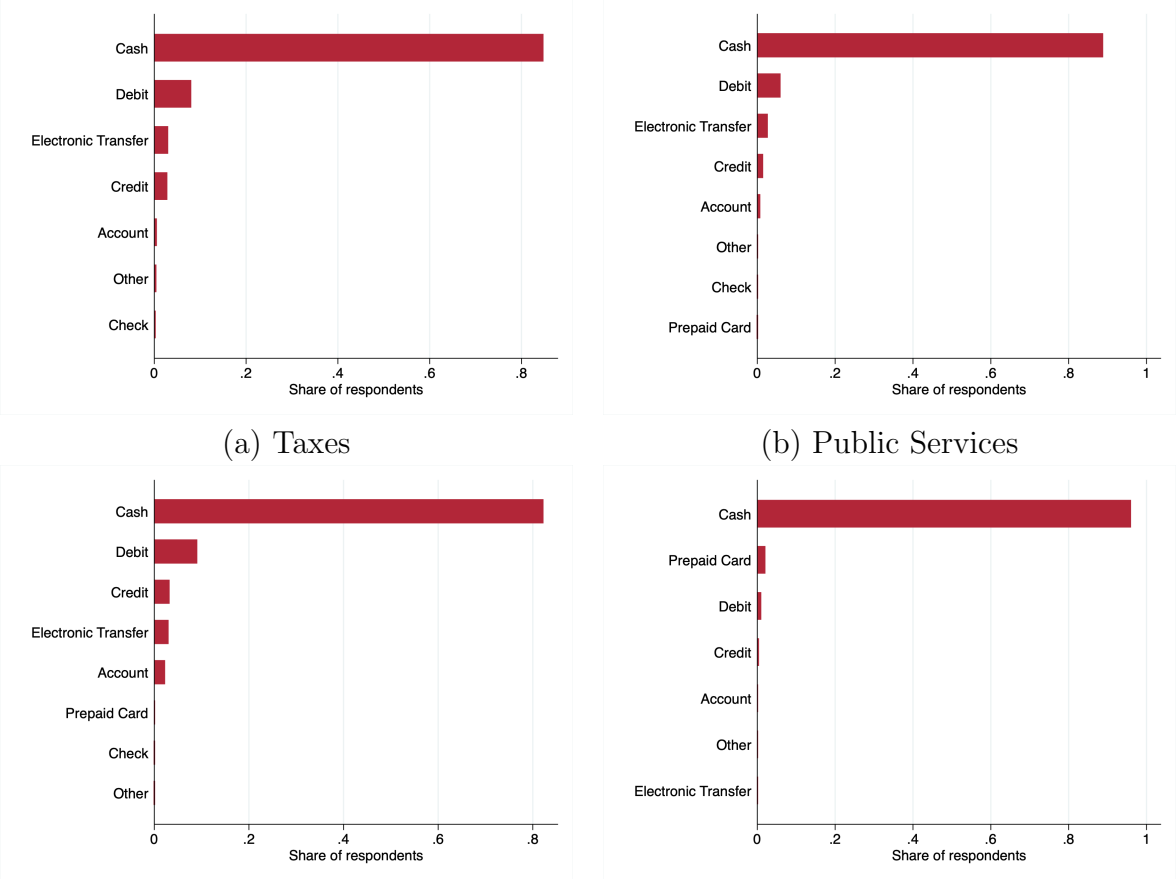
	Large	Medium	Small	Micro
Commerce	0.715	0.724	0.677	0.451
Construction	0.203	0.028	0.020	0.061
Manufacturing	0.055	0.156	0.134	0.142
Services	0.379	0.381	0.397	0.252

C.4 ATM-Sharing Agreements:

C.5 Data

Financial Inclusion Database (BDIF). The Financial Inclusion Databases (BDIF in Spanish) from the National Banking and Securities Commission (CNBV) consist of quarterly data gathered from commercial banks and other financial entities related to financial inclu-

Figure C.7: Payment method - Mixed Users

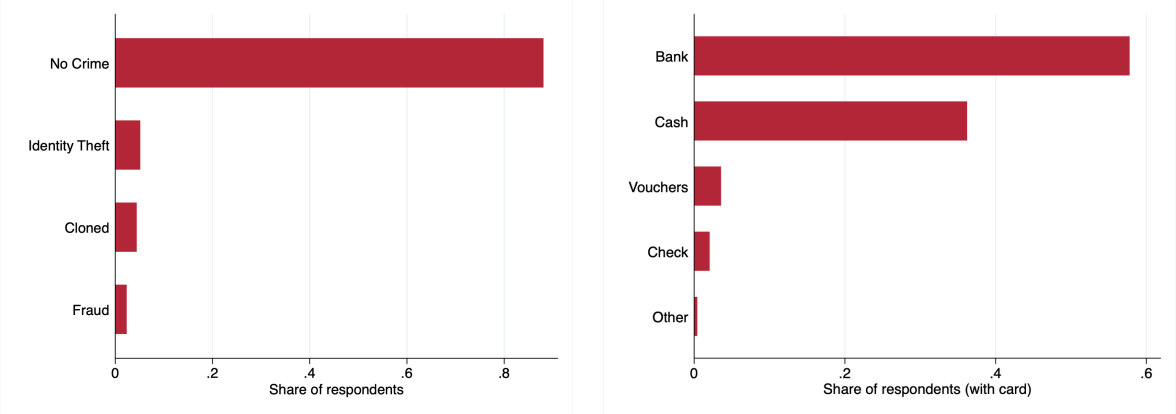


Note: The figure shows the most frequent payment methods reported by households for different types of expenditures. The panels report expenditures on taxes, expenditures on public services (e.g. water, electricity), private services (e.g. cable, phone, internet), and transportation (e.g. taxi, bus). The sample of households report owning a debit or a credit card. The data comes from the 2018 National Survey of Financial Inclusion (ENIF).

sion. The databases include variables such as bank branches, ATMs, point-of-sale terminals (POS), bank accounts and debit and credit cards. Data is disaggregated at the state and municipality level.[1] The data gathered for this paper is at the monthly level and corresponds to the period 2011-2019.

1. See the instructions in the manual R24-B 2422 *Información de variables operativas.*

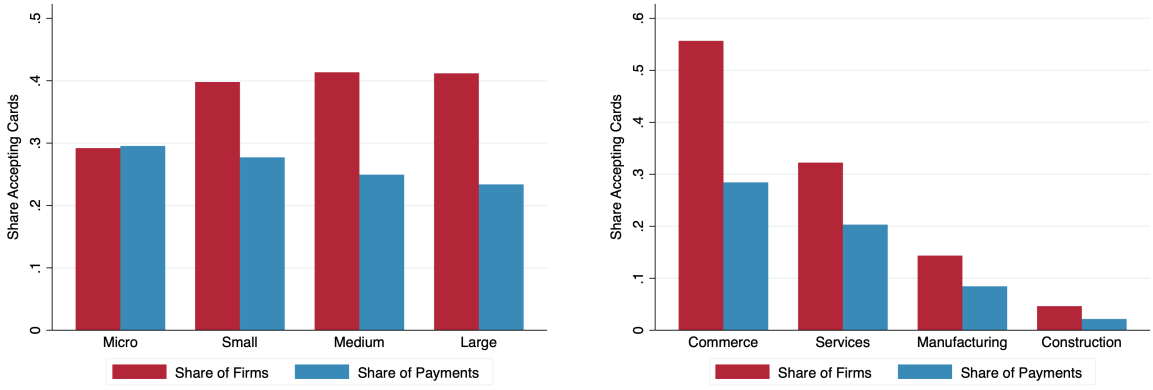
Figure C.8: Mixed Users: Crime and Wages



(a) Victim of a card related crime? (b) Method you get paid?

Note: Panel (a) shows the responses of households to the question "have you been victim of a credit card related crime?." Panel (b) shows the responses of households to the question "What payment method do you get paid in?." The sample of households report owning a debit or a credit card. The data comes from the 2018 National Survey of Financial Inclusion (ENIF).

Figure C.9: Share of Firms Accepting Card

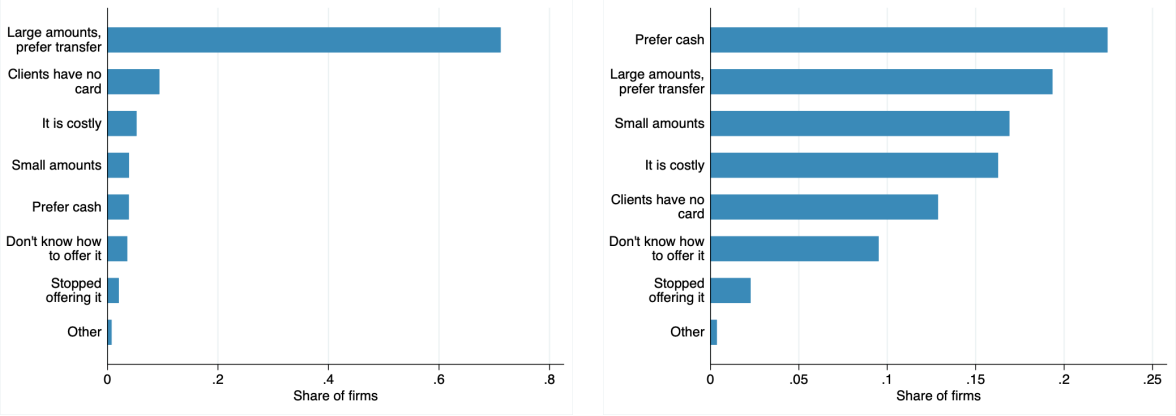


(a) Size (b) Sector

Note: The figure reports the share of firms that accept credit or debit cards and the share of total payments by size and by sector. The size bins are defined by the total number of employees: Micro (6-10), small (11-30), medium (30-100), large (100+). The data comes from the 2018 National Survey of Enterprise Financing (ENAFIN).

National Survey of Household Income and Expenditure (ENIGH). The National Survey of Household Income and Expenditure (ENIGH in Spanish) is a biannual household survey representative at the National level gathered by the National Institute of Statistics and

Figure C.10: Reasons For Not Accepting Card As Payment Method by Size



(a) Large Firm (b) Micro Firm

Note: The figure reports the most frequent reason stated by large and micro firms for not accepting cards as payment method. The data comes from the 2018 National Survey of Enterprise Financing (ENAFIN).

Table C.2: Share of Firms that Accept Credit Cards as Payment Method

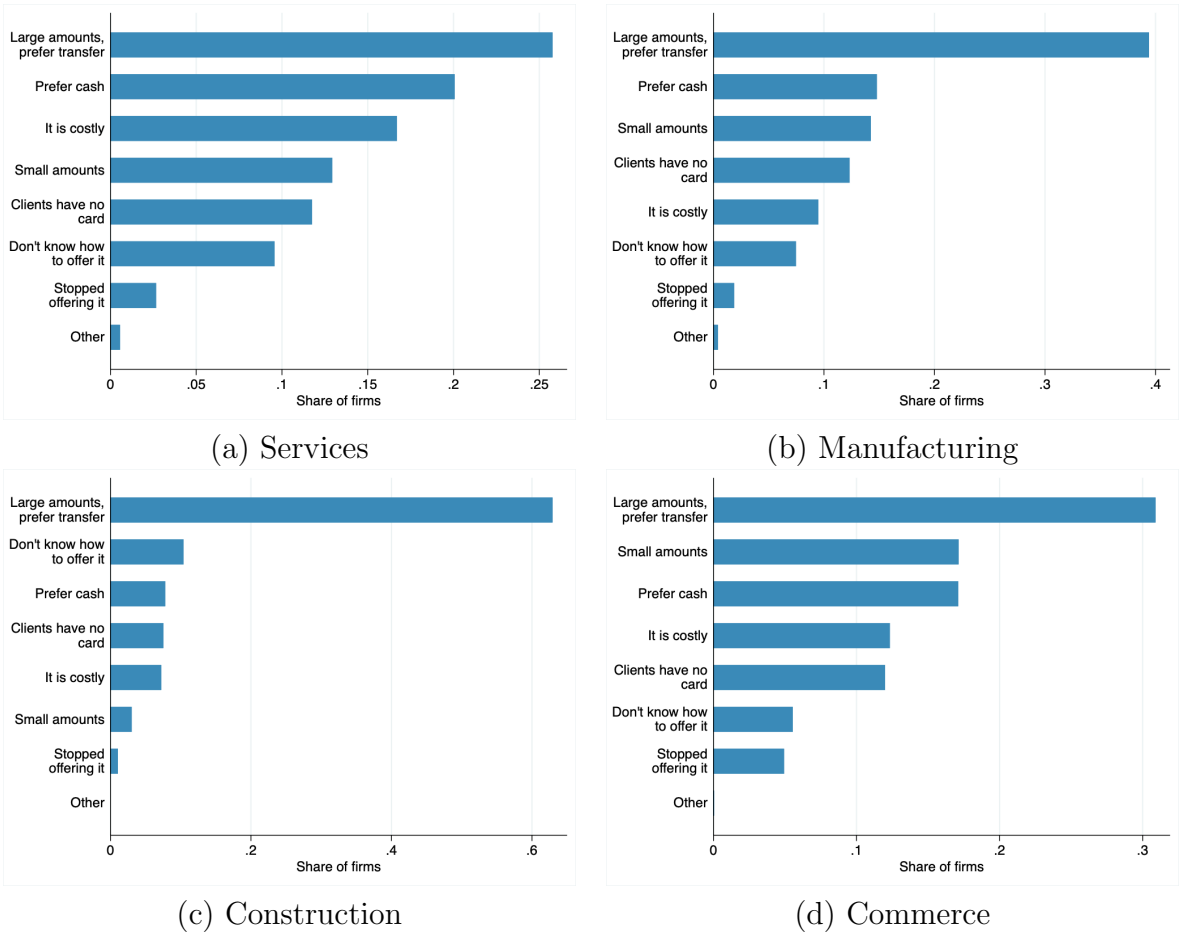
Note: The table shows the share of firms that accept credit cards as payment method. Each cell indicates the share of firms within that cell that accept credit cards. The size bins are defined by the total number of employees: Micro (6-10), small (11-30), medium (30-100), large (100+). The data comes from the 2018 National Survey of Enterprise Financing (ENAFIN).

	Large	Medium	Small	Micro
Commerce	0.733	0.746	0.672	0.445
Construction	0.173	0.065	0.020	0.075
Manufacturing	0.068	0.165	0.137	0.151
Services	0.393	0.389	0.403	0.254

Geography (INEGI). It gives information on the characteristics of housing units and socio-demographic and economic characteristics of the household members. It provides detailed information about expenditures, such as the type of goods purchased and the method of payment, which are gathered using a diary. We use the latest survey corresponding to 2016.

National Survey of Financial Inclusion (ENIF). The National Survey of Financial Inclusion (ENIF in Spanish) is a triannual household survey representative at the National

Figure C.11: Reasons For Not Accepting Card As Payment Method by Sector



(a) Services (b) Manufacturing

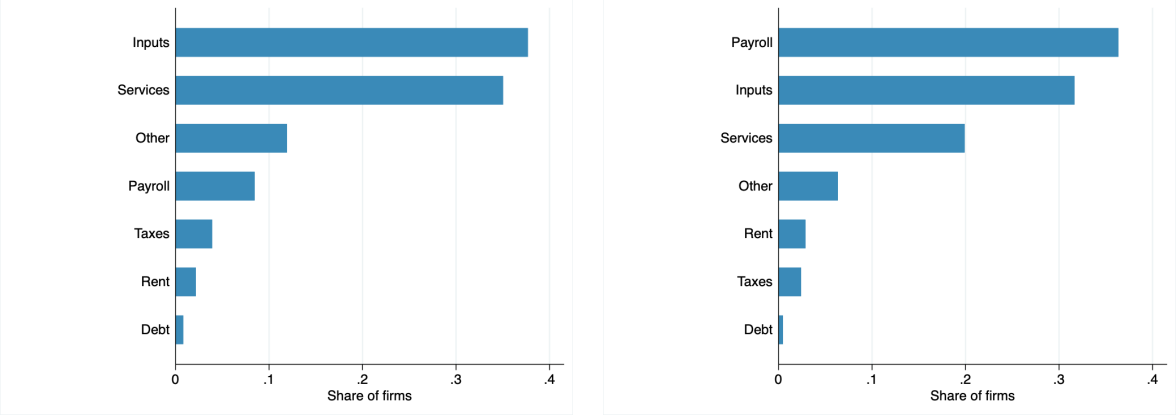
(c) Construction (d) Commerce

Note: The figure reports the most frequent reason stated by firms of different sectors for not accepting cards as payment method. The data comes from the 2018 National Survey of Enterprise Financing (ENAFIN).

level gathered by INEGI. It provides information about access and use of payment methods, saving products, loans and other financial products. We use the latest survey corresponding to 2016.

National Survey of Enterprise Financing (ENAFIN). The aim of the survey is to provide information related to the sources and use of financing mainly during the year 2017, as well as the needs of financial and banking services of enterprises, among other topics. Importantly for us, the survey contains information on whether the firms accept payment

Figure C.12: Share of Payments Made in Cash by Size



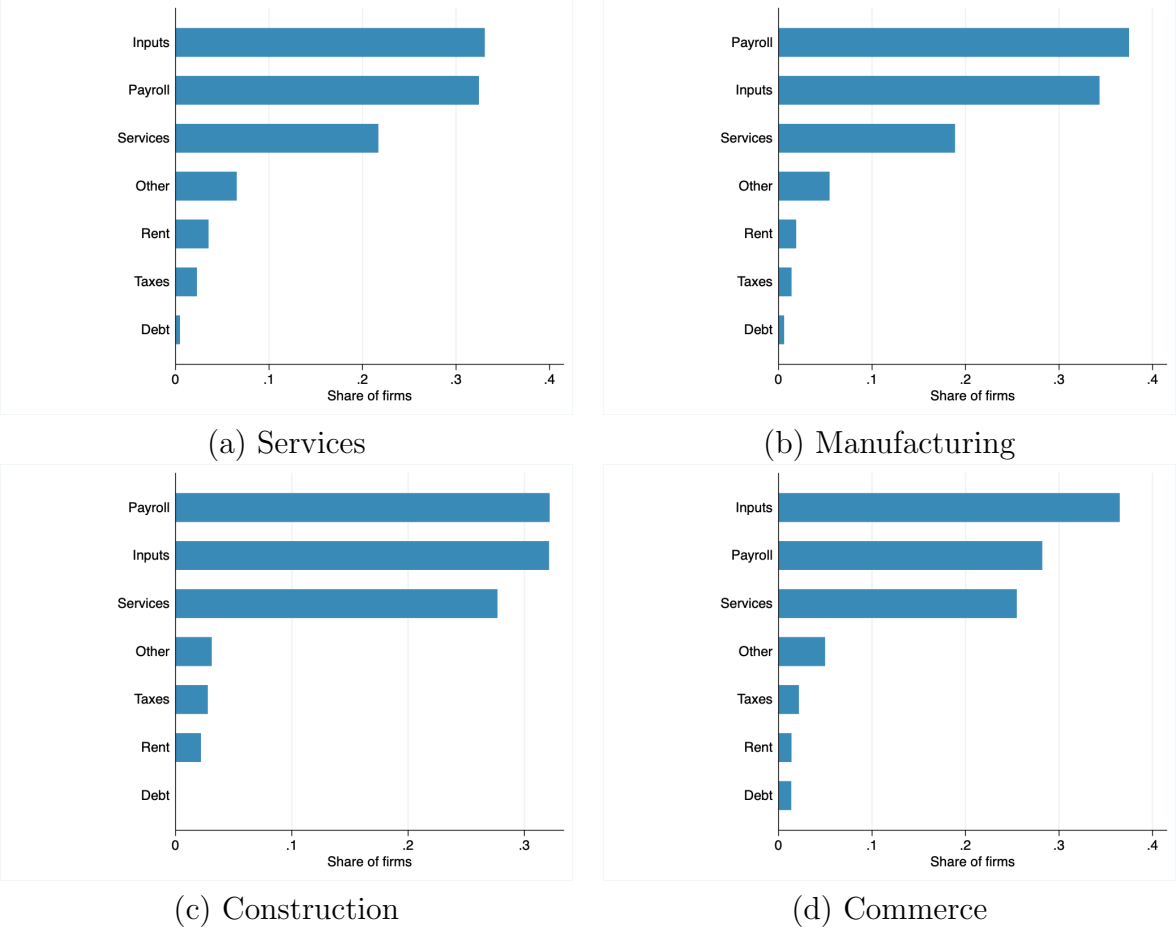
(a) Large (b) Micro

Note: The figure reports the share of payments made by firms in cash by type of payment. Panel (a) shows firms with more than 100 employees and Panel (b) firms with 6-10 employees. The data comes from the 2018 National Survey of Enterprise Financing (ENAFIN).

methods other than cash. It also has information on the payment methods firms use to conduct their own payments such as paying salaries, renting capital, or paying for services, taxes, and other financial obligations. The survey is representative at the national level and by size of locality. For the latter, there are only two ranges according to the number of inhabitants: from 50,000 to 499,999 and of 500,000 and more. Information is collected from the construction, manufacturing, trade and private non-financial services sectors including transport. The data classifies firms by their number of employees: micro 6-10, small 10-30, medium 30-100, and large firms with more than 100 employees.

National Employment Survey (ENOE). The National Employment Survey (ENOE), conducted by the National Institute of Statistics and Geography (INEGI), is the main source of labor related statistics in Mexico. The data gathered by the survey on a quarterly basis and it is representative at the level of locations of less than 2.500 inhabitants. The economically active population, used as control in some of our estimations, includes people who during the reference period carried out or had an economic activity (employed population) or actively

Figure C.13: Share of Payments Made in Cash by Sector



(a) Services (b) Manufacturing

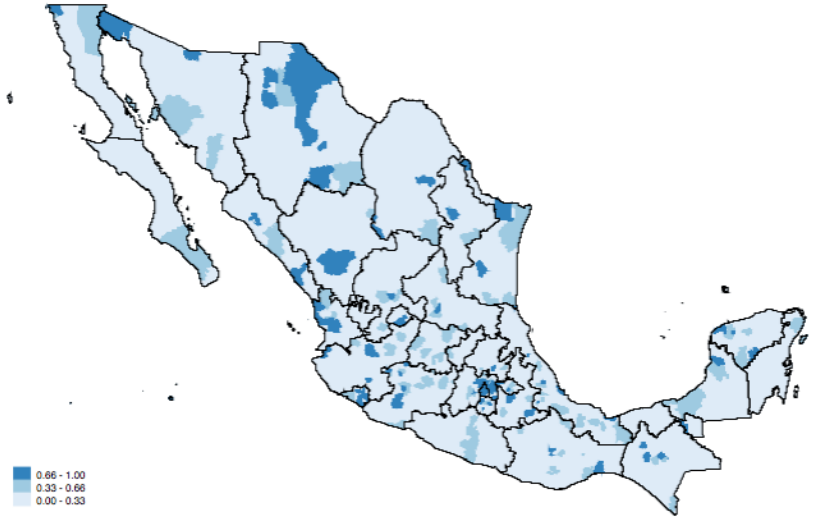
(c) Construction (d) Commerce

Note: The figure reports the share of payments made by firms in cash by type of payment. Panels (a)-(c) show the responses of firms in services, manufacturing, construction, and commerce respectively. The data comes from the 2018 National Survey of Enterprise Financing (ENAFIN).

sought to carry out one at some moment of the month prior to the day of the interview (unemployed population).

Criminal Incidence from the Executive Secretariat of the Public Security National System (SESNSP). The criminal incidence reported by the SESNSP refers to the alleged occurrence of crimes recorded in previous investigations initiated or investigation files, reported by the Attorney General's Offices and Attorney General's offices in the case of the common law and by the Attorney General's Office in the federal jurisdiction. The data

Figure C.14: Share of Beneficiaries in the Rollout by Municipality (2012)



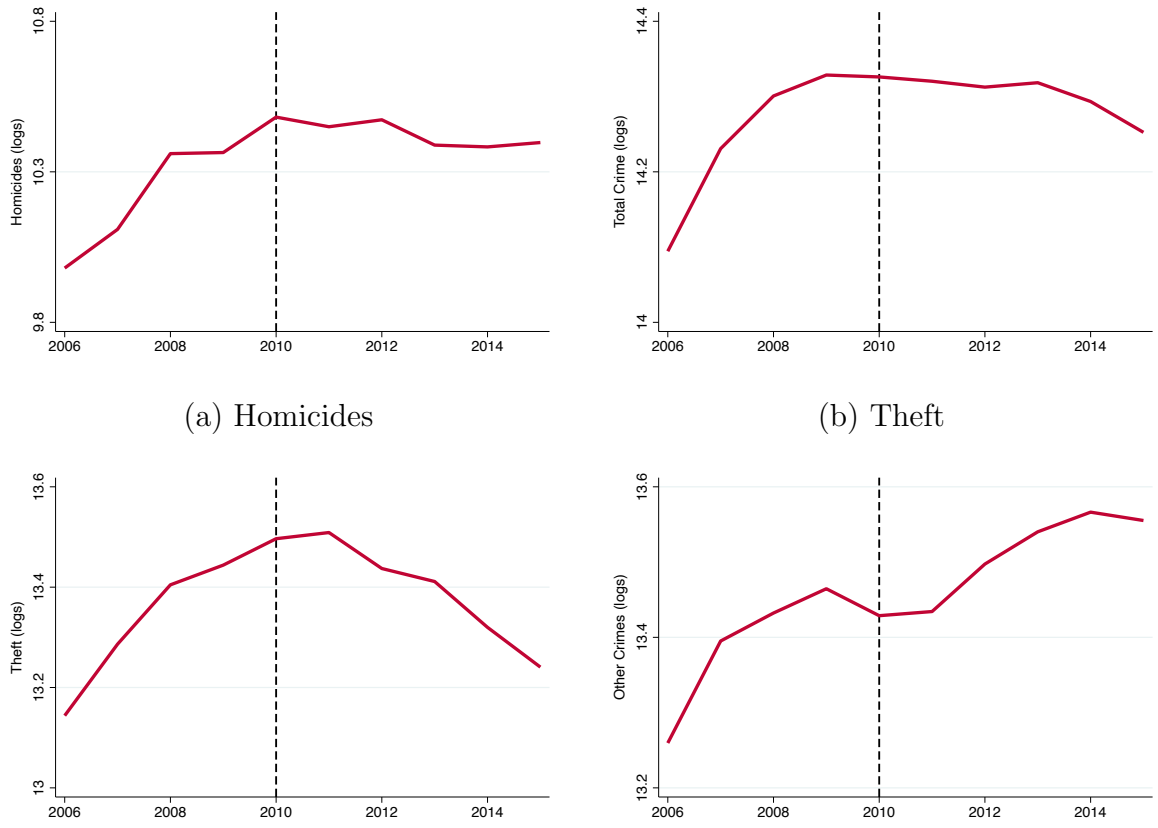
Note: The map shows the share of beneficiaries that were part of the rollout of debit cards by municipality. The shares are calculated for 2012. The data comes from the administrative data of the Prospera program.

contains violent crimes (e.g. sexual assault, murder), property crimes (e.g. robery/burglary), other crimes (e.g. extortion, fraud).

Administrative data from Prospera. Prospera provided confidential data at the municipality level by two-month payment period level. The data include the number of beneficiaries in the municipality and the payment method by which they are paid. Examples of payment methods include cash, bank account without debit card, and bank account with debit card. These data, which span 2007–2015 and all 2,457 of Mexico's municipalities.

State and Municipal Public Finances (EFIPEM). The National Institute of Statistics and Geography (INEGI) provides information on the public finances of each municipality at the annual level. The data includes the taxes collected by each municipality in a calendar year including estate taxes, property taxes, production taxes, consumption taxes, new cars and motor vehicle taxes, and gasoline taxes.

Figure C.15: Crime



(a) Homicides

(b) Theft

(c) Thefts

(d) Other Crimes

Note: The figure shows the evolution of the total number crimes, homicides, thefts, and other crimes from 2006 to 2015. From 2006-2010, the figures use information from the State and Municipal Databases (SIMBAD). From 2011 onward, the figures use data from the Executive Secretariat of the Public Security National System (SESNSP). The dashed line indicates the transitions across data sets.

Statistics of Registered Deaths. The statistics of registered deaths are produced by the National Institute of Statistics and Geography (INEGI). The statistics are based on death certificates. They have detailed information of the causes of death, including deaths from homicide, and the date the death occurred. They also include detailed information on the place of death at the locality level.

Table C.3: Effect of Card Shock on Debit and Credit Cards

Note: The table reports the results for the coefficient of β after estimating (3.2). The dependent variable in Columns (1), (2), (9) and (10) is the logarithm of debit cards. Columns (3), (4), (11) and (12) use debit cards excluding those given as part of the Prospera program through Bansefi. Columns (5), (6), (13) and (14) use credit cards and Columns (7), (8), (15) and (16) use the sum of debit cards and credit cards as dependent variable. We use the inverse hyperbolic sine transformation in all cases. The controls we use include income per capita, total employment, and total population, and the total number of families in the Prospera program. The specifications that are weights use the total population in the municipality. We use Driscoll and Kraay standard errors in all specifications.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Debit		Debit Not Prospera		Credit		Total Cards	
Card Shock	0.313***	0.299***	0.493***	0.464***	0.185*	0.179**	0.196***	0.201***
	(0.073)	(0.058)	(0.153)	(0.117)	(0.086)	(0.064)	(0.054)	(0.039)
Obs.	6,181	5,212	6,181	5,212	6,181	5,212	6,181	5,212
Mun.	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	N	N	N	N	N	N	N	N
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
	Debit		Debit Not Prospera		Credit		Total Cards	
Card Shock	0.163***	0.132***	0.205***	0.167***	0.140**	0.115**	0.133***	0.114***
	(0.031)	(0.029)	(0.047)	(0.040)	(0.058)	(0.050)	(0.030)	(0.026)
Obs.	6,181	5,212	6,181	5,212	6,181	5,212	6,181	5,212
Mun.	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	Y	Y	Y	Y	Y	Y	Y	Y

CONAPO Population Estimates. We use the Mexican population estimates from the National Population Council. The estimates are at the annual level and at the municipality level. The estimates are constructed using Mexican Census and Intercensal Surveys, which are carried out to update socio-demographic information at the midpoint between censuses.

Annual Crime Statistics. We use crime statistics from 2005-2010 collected from the National Institute of Statistics and Geography (INEGI) and available at the State and Municipal Databases (SIMBAD). The data is based on registered crimes collected from local

Table C.4: Effect of Card Shock on Debit and Credit Cards (Log)

Note: The table reports the results for the coefficient of β after estimating (3.2). The dependent variable in Columns (1) and (2) is the logarithm of debit cards. Columns (3) and (4) use debit cards excluding those given as part of the Prospera program through Bansefi. Columns (5) and (6) use credit cards and Columns (7) and (8) use the sum of debit cards and credit cards as dependent variable. We use the natural logarithm in Columns (2), (4), (6) and (8). We use the inverse hyperbolic sine transformation for values greater than zero in the dependent variable in Columns (1), (3), (5), and (7). The controls we use include income per capita, total employment, and total population, and the total number of families in the Prospera program. The specifications that are weights use the total population in the municipality. We use Driscoll and Kraay standard errors in all specifications.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Debit		Debit Not Prospera		Credit		Total Cards	
Card Shock	0.096***	0.066***	0.071**	0.071**	0.108**	0.108**	0.114***	0.114***
	(0.024)	(0.024)	(0.030)	(0.030)	(0.050)	(0.050)	(0.026)	(0.026)
Obs.	5,149	5,149	5,103	5,103	5,189	5,189	5,212	5,212
Mun.	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Weights	Y	Y	Y	Y	Y	Y	Y	Y
Specification	IHS	Log	IHS	Log	IHS	Log	IHS	Log

Table C.5: Effect of Card Shock on Homicides (Locality)

Note: The table reports the results for the coefficient of β after estimating (3.2) at the locality level and at bi-monthly frequency. The dependent variable is the logarithm of the total number of homicides. We use the inverse hyperbolic sine transformation in all cases. We control for the total number of families in the Prospera program in each locality. We use Driscoll and Kraay standard errors in all specifications.

	(1)	(2)	(3)	(4)	(5)	(6)
Card Shock	0.0073***	0.0053*	0.0092**	0.0064**	0.0047*	0.0084**
	(0.003)	(0.003)	(0.004)	(0.002)	(0.003)	(0.004)
Observations	540,594	540,594	529,653	540,594	540,594	529,653
Municipality	Y	Y	Y	Y	Y	Y
Controls	N	N	N	Y	Y	Y
Period	Y	N	N	Y	N	N
State × Period	N	Y	N	N	Y	N
Municipality × Period	N	N	Y	N	N	Y

Table C.6: Effect of Card Shock on Homicides (Municipality)

Note: The table reports the results for the coefficient of β after estimating (3.2). The dependent variable in Columns (1), (2), (9) and (10) is the logarithm of homicides using data from INEGI based on death certificates. Columns (5), (6), (13) and (14) use the logarithm of homicides using data from SESNSP based on criminal cases. We use the inverse hyperbolic sine transformation in all cases. Columns (3), (4), (11) and (12) use homicide rate per 10,000 persons from INEGI as dependent variable. Columns (7), (8), (15) and (16) use homicide rate per 10,000 persons from SESNSP as dependent variable. The controls we use include income per capita, total employment, and total population, and the total number of families in the Prospera program. The specifications that are weights use the total population in the municipality. We use Driscoll and Kraay standard errors in all specifications.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	INEGI		SESNSP		INEGI		SESNSP	
	Homicides		Homicides		Homicide Rate		Homicide Rate	
Card Shock	-0.0156	-0.0144	0.1712	0.1533	0.0810**	0.0763*	0.3208	0.3039
	(0.053)	(0.052)	(0.118)	(0.147)	(0.033)	(0.034)	(0.185)	(0.234)
Observations	3,672	3,149	3,149	3,149	3,517	3,028	3,028	3,028
Municipality	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	N
Weights	N	N	N	N	N	N	N	N
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
	INEGI		SESNSP		INEGI		SESNSP	
	Homicides		Homicides		Homicide Rate		Homicide Rate	
Card Shock	0.0911	0.0923	0.2849*	0.3153**	0.0427	0.0393	0.2394**	0.3203***
	(0.065)	(0.062)	(0.138)	(0.116)	(0.030)	(0.029)	(0.093)	(0.090)
Observations	3,149	3,149	3,149	3,149	3,028	3,028	3,028	3,028
Municipality	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	Y	Y	Y	Y	Y	Y	Y	Y

criminal courts and includes information on total thefts, homicides, injuries, damages, sex crimes, and kidnaps. The rest of the crimes are classified as other crimes. The data is at the annual level and is available at the municipality level.

Table C.7: Effect of Card Shock on Theft

Note: The table reports the results for the coefficient of β after estimating (3.2). The dependent variable in Columns (1), (2), (7) and (8) is the logarithm of total thefts. We use the inverse hyperbolic sine transformation in all cases. Columns (3), (4), (9) and (10) use theft rate per 10,000 persons. Columns (5), (6), (11) and (12) use the logarithm of theft divided by total crimes. We again use the inverse hyperbolic sine transformation in all cases. The controls we use include income per capita, total employment, and total population, and the total number of families in the Prospera program. The specifications that are weights use the total population in the municipality. We use Driscoll and Kraay standard errors in all specifications.

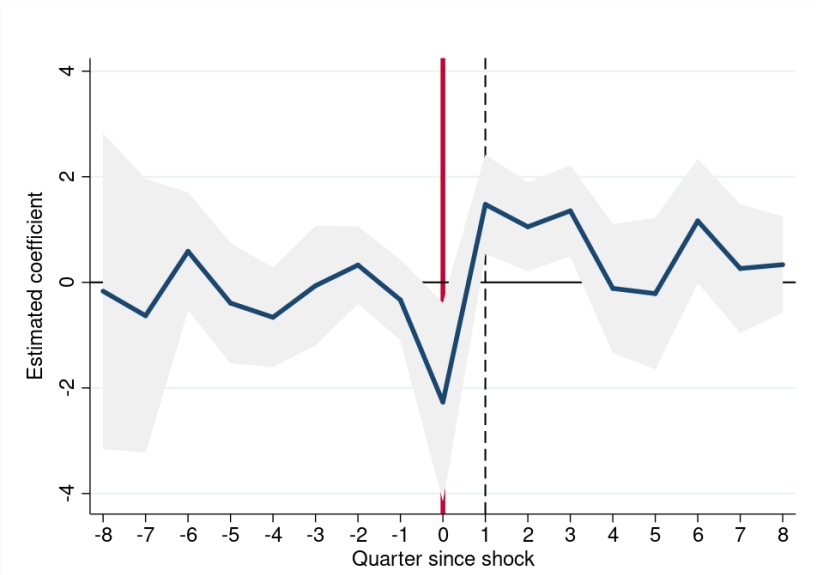
	(1)	(2)	(3)	(4)	(5)	(6)
	Thefts		Theft Rate		Theft/Crime	
Card Shock	-0.0478	-0.0088	-0.9098*	-0.7936*	-0.0123**	-0.0065
	(0.035)	(0.049)	(0.446)	(0.395)	(0.004)	(0.004)
Observations	3,505	3,027	3,505	3,027	3,452	2,989
Municipality	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y
Weights	N	N	N	N	N	N
	(7)	(8)	(9)	(10)	(11)	(12)
	Thefts		Theft Rate		Theft/Crime	
Card Shock	-0.0405	-0.0238	-0.8949	-0.3181	-0.0099	-0.0065
	(0.024)	(0.013)	(0.753)	(0.940)	(0.006)	(0.005)
Observations	3,505	3,027	3,505	3,027	3,452	2,989
Municipality	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y
Weights	Y	Y	Y	Y	Y	Y

Table C.8: Effect of Card Shock on Total Crime

Note: The table reports the results for the coefficient of β after estimating (3.2). The dependent variable in Columns (1)-(4) is the logarithm of total crimes. We use the inverse hyperbolic sine transformation in all cases. Columns (5)-(8) use crime rate per 10,000 persons. The controls we use include income per capita, total employment, and total population, and the total number of families in the Prospera program. The specifications that are weights use the total population in the municipality. We use Driscoll and Kraay standard errors in all specifications.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Crimes				Crime Rate			
Card Shock	0.0182	0.0377	0.0100	0.0164	-0.8181	-1.1457	-1.8768	-1.3784
	(0.027)	(0.036)	(0.020)	(0.014)	(1.153)	(1.087)	(1.337)	(1.426)
Observations	3,505	3,027	3,505	3,027	3,505	3,027	3,505	3,027
Municipality	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	N	N	Y	Y	N	N	Y	Y

Figure C.16: Alternative Lags and Leads for the Agreement Shock in the Bartik 1st Stage



Note: The graph shows the evolution of the growth rate ($d\ln$) of ATM transactions before and after ATM sharing agreements. The figures plot the coefficients of γ_k after estimating regressions of the form $d\ln w_{mt+k} = \gamma^w \sum_i \sum_j E_{ijt} d\ln p_{ijt} z_{ijm0} + \theta^w_m + \lambda^w_t + \epsilon^w_{mt}$, where k represent quarters after the shock. The red line marks the quarter or year in which an agreement occurred. The dashed line corresponds to $k = 0$, the lagged agreement dummies that we use in Equation 3.4. The gray area depicts the 95% confidence interval using standard errors clustered at the municipality level.

Table C.9: Effect of Card Shock on Informality

Note: The table reports the results for the coefficient of β after estimating (3.2). The dependent variable in Columns (1), (2), (9), and (10) is the logarithm of informal workers. The dependent variable in Columns (5), (6), (13), and (14) is the logarithm of self-employed workers. We use the inverse hyperbolic sine transformation in all cases. The dependent variable in Columns (3), (4), (11), and (12) is the ratio of informal workers and the total population of the municipality. The dependent variable in Columns (7), (8), (15), and (16) is the ratio of informal workers and the total population of the municipality. The controls we use include income per capita, total employment, and total population, and the total number of families in the Prospera program. The specifications that are weights use the total population in the municipality. We use Driscoll and Kraay standard errors in all specifications.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Informality		Informality Rate		Self-Employed		Self-Employed Rate	
Card Shock	-0.0118	-0.0013	0.0004	-0.0009	-0.0129	-0.0026	0.0002	-0.0004
	(0.033)	(0.009)	(0.003)	(0.002)	(0.037)	(0.025)	(0.002)	(0.002)
Observations	6,225	6,224	6,225	6,224	6,225	6,224	6,225	6,224
Municipality	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	N	N	N	N	N	N	N	N
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
	Informality		Informality Rate		Self-Employed		Self-Employed Rate	
Card Shock	0.0108	0.0089	0.0012	0.0017	0.0020	-0.0005	-0.0002	-0.0000
	(0.020)	(0.010)	(0.002)	(0.002)	(0.024)	(0.017)	(0.001)	(0.001)
Observations	6,225	6,224	6,225	6,224	6,225	6,224	6,225	6,224
R-squared	0.949	0.993	0.814	0.896	0.928	0.965	0.657	0.704
Municipality	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	Y	Y	Y	Y	Y	Y	Y	Y

Table C.10: Effect of Card Shock on Local Taxes

Note: The table reports the results for the coefficient of β after estimating (3.2). The dependent variable in Columns (1)-(4) is the logarithm of local taxes. The dependent variable in Columns (5)-(8) is the ratio of taxes and the total population of the municipality. The controls we use include income per capita, total employment, and total population, and the total number of families in the Prospera program. The specifications that are weights use the total population in the municipality. We use Driscoll and Kraay standard errors in all specifications.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Taxes				Taxes/Population			
Card Shock	-0.0125	-0.0086	-0.0084	-0.0060	3.0979	3.9631	7.5085	6.1952
	(0.013)	(0.019)	(0.015)	(0.016)	(3.184)	(3.252)	(5.607)	(5.545)
Observations	3,382	2,895	2,895	2,895	3,382	2,895	2,895	2,895
Municipality	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	N	N	Y	Y	N	N	Y	Y

Table C.11: Effect of Card Shock on Debit and Credit Cards

Note: The table reports the results for the coefficient of β after estimating (C.2). The dependent variable in Columns (1), (2), (9) and (10) is the logarithm of debit cards. Columns (3), (4), (11) and (12) use debit cards excluding those given as part of the Prospera program through Bansefi. Columns (5), (6), (13) and (14) use credit cards and Columns (7), (8), (15) and (16) use the sum of debit cards and credit cards as dependent variable. We use the inverse hyperbolic sine transformation in all cases. The controls we use include income per capita, total employment, and total population, and the total number of families in the Prospera program. The specifications that are weights use the total population in the municipality. We use Driscoll and Kraay standard errors in all specifications.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Debit		Debit Not Prospera		Credit		Total Cards	
Progresa	2.542***	2.387***	2.813**	2.837***	2.068***	2.526***	1.961***	2.163***
× Rollout	(0.496)	(0.504)	(0.951)	(0.644)	(0.450)	(0.413)	(0.369)	(0.483)
Observations	6,181	5,212	6,181	5,212	6,181	5,212	6,181	5,212
Municipality	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	N	N	N	N	N	N	N	N
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
	Debit		Debit Not Prospera		Credit		Total Cards	
Progresa	1.561***	1.192***	1.054**	0.782	2.386***	2.466***	1.523***	1.366***
× Rollout	(0.340)	(0.277)	(0.482)	(0.461)	(0.534)	(0.553)	(0.272)	(0.243)
Observations	6,181	5,212	6,181	5,212	6,181	5,212	6,181	5,212
Municipality	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	Y	Y	Y	Y	Y	Y	Y	Y

Table C.12: Effect of Card Shock on Debit and Credit Cards (Log)

Note: The table reports the results for the coefficient of β after estimating (3.2). The dependent variable in Columns (1) and (2) is the logarithm of debit cards. Columns (3) and (4) use debit cards excluding those given as part of the Prospera program through Bansefi. Columns (5) and (6) use credit cards and Columns (7) and (8) use the sum of debit cards and credit cards as dependent variable. We use the natural logarithm in Columns (2), (4), (6) and (8). We use the inverse hyperbolic sine transformation for values greater than zero in the dependent variable in Columns (1), (3), (5), and (7). The controls we use include income per capita, total employment, and total population, and the total number of families in the Prospera program. The specifications that are weights use the total population in the municipality. We use Driscoll and Kraay standard errors in all specifications.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Debit		Debit Not Prospera		Credit		Total Cards	
Progresa	1.308***	1.312***	0.464**	0.4686**	2.313***	2.315***	1.366***	1.367***
× Rollout	(0.246)	(0.247)	(0.188)	(0.187)	(0.527)	(0.527)	(0.243)	(0.242)
Observations	5,149	5,149	5,103	5,103	5,189	5,189	5,212	5,212
Municipality	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Weights	Y	Y	Y	Y	Y	Y	Y	Y

Table C.13: Effect of Card Shock on Homicides (Municipality)

Note: The table reports the results for the coefficient of β after estimating (C.2). The dependent variable in Columns (1), (2), (9) and (10) is the logarithm of homicides using data from INEGI based on death certificates. Columns (5), (6), (13) and (14) use the logarithm of homicides using data from SESNSP based on criminal cases. We use the inverse hyperbolic sine transformation in all cases. Columns (3), (4), (11) and (12) use homicide rate per 10,000 persons from INEGI as dependent variable. Columns (7), (8), (15) and (16) use homicide rate per 10,000 persons from SESNSP as dependent variable. The controls we use include income per capita, total employment, and total population, and the total number of families in the Prospera program. The specifications that are weights use the total population in the municipality. We use Driscoll and Kraay standard errors in all specifications.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	INEGI Homicides		SESNSP Homicides		INEGI Homicide Rate		SESNSP Homicide Rate	
Share Progresa	-0.0875	-0.4188	-0.2699	-0.4325	0.3537	0.2915	-2.1235	-2.4744
× Rollout	(0.226)	(0.237)	(0.638)	(0.666)	(0.347)	(0.367)	(1.585)	(1.536)
Observations	3,672	3,149	3,149	3,149	3,517	3,028	3,028	3,028
Municipality	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	N
Weights	N	N	N	N	N	N	N	N
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
	INEGI Homicides		SESNSP Homicides		INEGI Homicide Rate		SESNSP Homicide Rate	
Share Progresa	0.2176	0.2932	3.2159**	3.3644***	0.8094*	0.8576**	4.2859**	3.7177**
× Rollout	(0.454)	(0.425)	(0.979)	(0.991)	(0.378)	(0.372)	(1.303)	(1.389)
Observations	3,149	3,149	3,149	3,149	3,028	3,028	3,028	3,028
Municipality	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	Y	Y	Y	Y	Y	Y	Y	Y

Table C.14: Effect of Card Shock on Theft

Note: The table reports the results for the coefficient of β after estimating (C.2). The dependent variable in Columns (1), (2), (7) and (8) is the logarithm of total thefts. We use the inverse hyperbolic sine transformation in all cases. Columns (3), (4), (9) and (10) use the theft rate per 10,000 persons. Columns (5), (6), (11) and (12) use the logarithm of theft divided by total crimes. We again use the inverse hyperbolic sine transformation in all cases. The controls we use include income per capita, total employment, and total population, and the total number of families in the Prospera program. The specifications that are weights use the total population in the municipality. We use Driscoll and Kraay standard errors in all specifications.

	(1)	(2)	(3)	(4)	(5)	(6)
	Thefts		Theft Rate		Theft/Crime	
Share Progresa	-0.6041	-0.3237	0.6341	0.4615	0.0551	0.1106
× Rollout	(0.370)	(0.399)	(4.573)	(5.089)	(0.048)	(0.070)
Observations	3,505	3,027	3,505	3,027	3,452	2,989
Municipality	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y
Weights	N	N	N	N	N	N
	(7)	(8)	(9)	(10)	(11)	(12)
	Thefts		Theft Rate		Theft/Crime	
Share Progresa	0.2872	0.4112	28.5657	34.6705**	0.2176	0.2678
× Rollout	(0.435)	(0.396)	(16.700)	(14.559)	(0.147)	(0.147)
Observations	3,505	3,027	3,505	3,027	3,452	2,989
Municipality	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y
Weights	Y	Y	Y	Y	Y	Y

Table C.15: Effect of Card Shock on Total Crime

Note: The table reports the results for the coefficient of β after estimating (C.2). The dependent variable in Columns (1)-(4) is the logarithm of total crimes. We use the inverse hyperbolic sine transformation in all cases. Columns (5)-(8) use the crime rate per 10,000 persons. The controls we use include income per capita, total employment, and total population, and the total number of families in the Prospera program. The specifications that are weights use the total population in the municipality. We use Driscoll and Kraay standard errors in all specifications.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Crimes				Crime Rate			
Share Progresa	-0.6388**	-0.4620	-0.2425	-0.2503	-7.3589	-16.5165**	26.7512*	27.4000**
× Rollout	(0.234)	(0.260)	(0.170)	(0.151)	(8.126)	(6.829)	(13.091)	(10.839)
Observations	3,505	3,027	3,505	3,027	3,505	3,027	3,505	3,027
Municipality	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	N	N	Y	Y	N	N	Y	Y

Table C.16: Effect of Card Shock on Informality

Note: The table reports the results for the coefficient of β after estimating (C.2). The dependent variable in Columns (1), (2), (9), and (10) is the logarithm of informal workers. The dependent variable in Columns (5), (6), (13), and (14) is the logarithm of self-employed workers. We use the inverse hyperbolic sine transformation in all cases. The dependent variable in Columns (3), (4), (11), and (12) is the ratio of informal workers and the total population of the municipality. The dependent variable in Columns (7), (8), (15), and (16) is the ratio of informal workers and the total population of the municipality. The controls we use include income per capita, total employment, and total population, and the total number of families in the Prospera program. The specifications that are weights use the total population in the municipality. We use Driscoll and Kraay standard errors in all specifications.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Informality		Informality Rate		Self-Employed		Self-Employed Rate	
Share Progresa	-0.0937	0.0937	0.0090	0.0228	-0.2776	-0.1225	-0.0045	-0.0014
× Rollout	(0.275)	(0.060)	(0.024)	(0.014)	(0.322)	(0.192)	(0.018)	(0.016)
Observations	6,225	6,224	6,225	6,224	6,225	6,224	6,225	6,224
Municipality	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	N	N	N	N	N	N	N	N
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
	Informality		Informality Rate		Self-Employed		Self-Employed Rate	
Share Progresa	-0.0831	0.1495**	0.0190	0.0349**	-0.1703	0.0203	0.0024	0.0057
× Rollout	(0.200)	(0.067)	(0.020)	(0.014)	(0.272)	(0.153)	(0.016)	(0.014)
Observations	6,225	6,224	6,225	6,224	6,225	6,224	6,225	6,224
Municipality	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	Y	Y	Y	Y	Y	Y	Y	Y

Table C.17: Effect of Card Shock on Local Taxes

Note: The table reports the results for the coefficient of β after estimating (C.2). The dependent variable in Columns (1)-(4) is the logarithm of local taxes. The dependent variable in Columns (5)-(8) is the ratio of taxes and the total population of the municipality. The controls we use include income per capita, total employment, and total population, and the total number of families in the Prospera program. The specifications that are weights use the total population in the municipality. We use Driscoll and Kraay standard errors in all specifications.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Taxes				Taxes/Population			
Progresa	-0.113	-0.080	0.099	0.102	-136.488**	-178.104**	-229.309*	-241.709**
× Rollout	(0.134)	(0.152)	(0.189)	(0.166)	(40.994)	(55.615)	(100.090)	(92.197)
Observations	3,382	2,895	2,895	2,895	3,382	2,895	2,895	2,895
Municipality	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	N	N	Y	Y	N	N	Y	Y

Table C.18: List of ATM-sharing agreements

Note: The table includes the list of 24 ATM sharing agreements approved by the Bank of Mexico between 2014 and 2019. Agreements 1 and 2 only apply to Banjército customers. Banregio joined Agreement 3 on June 27th, 2016. The CNBV has no financial data of Accendo (Agreement 18). The data comes from the 2019 *Informe Anual sobre las Infraestructuras de los Mercados Financieros* (Annual Report on Financial Market Infrastructures) of the Bank of Mexico. The reduction in fee uses the maximum fee before the agreement and the minimum fee after the agreement and it includes withdrawals and balance checks fees.

Agreement	Date	Banks	% Reduction in fee
1	November 24, 2014	Banjército, Banamex	100
2	November 24, 2014	Banjército, BBVA Bancomer	100
3	April 20, 2015	Bajío, Inbursa, Scotiabank, Banregio	60
4	June 30, 2015	Afirme, Bajío	100
5	November 24, 2015	Afirme, BanCoppel	50
6	January 20, 2016	Afirme, Scotiabank	100
7	September 19, 2016	Afirme, Inbursa	100
8	September 28, 2016	Scotiabank, Mifel	100
9	October 18, 2016	Multiva, American Express	100
10	November 15, 2016	Scotiabank, Actinver	100
11	January 24, 2017	Scotiabank, BanCoppel	50
12	January 24, 2017	Scotiabank, Intercam	60
13	March 28, 2017	Bansefi, Banjército	100
14	July 28, 2017	Scotiabank, Famsa	60
15	October 9, 2017	Bajío, Famsa	60
16	February 28, 2018	Scotiabank, Autofin	46.6
17	February 28, 2018	Scotiabank, Multiva	62.5
18	April 24, 2018	Bancoppel, Accendo	7.2
19	April 24, 2018	Actinver, Multiva	100
20	October 16, 2018	Azteca, Multiva	100
21	October 30, 3018	Bajío, Intercam	61.3
22	January 30, 2019	Azteca, Mifel	100
23	May 13, 2019	Azteca, Bajío	100
24	October 21, 2019	Afirme, Azteca	100

Table C.19: Effect of ATM-Sharing Agreements on ATM Withdrawals and Debit Cards

Note: The table reports the results for the coefficient γ after estimating Equation 3.5. Observations are at the municipality-quarterly level. The dependent variable in Columns (1) - (4) is the quarterly change in the logarithm of the total ATM withdrawal count. Columns (5) - (8) use the quarterly change in the logarithm of debit card contracts. We use the inverse hyperbolic sine transformation in all cases. The controls we use include income per capita, total employment, and total population. The specifications with weights use the total population in the municipality on the pre-period. Standard errors are clustered at the municipality level (called Report R2422).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	ATM Withdrawals				Debit			
Bartik	0.0820	0.3507	1.4766***	1.6770***	0.3193	0.3253	1.0940	0.6504
	(0.519)	(0.690)	(0.476)	(0.600)	(0.502)	(0.655)	(0.981)	(1.093)
Observations	34,415	20,695	34,397	20,695	34,415	20,695	34,397	20,695
Municipality	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	N	N	Y	Y	N	N	Y	Y

Table C.20: Effect of ATM-Sharing Agreements on Homicides

Note: The table reports the results for the coefficient γ after estimating Equation 3.5. Observations are at the municipality-quarterly level. The dependent variable in Columns (1), (2), (9) and (10) is the quarterly change in the logarithm of homicides using data from INEGI based on death certificates. Columns (5), (6), (13) and (14) use the quarterly change in the logarithm of homicides using data from SESNSP based on criminal cases. Columns (3), (4), (11) and (12) use the quarterly change in the logarithm of the homicide rate per 10,000 persons from INEGI as dependent variable. Columns (7), (8), (15) and (16) use the quarterly change in the logarithm of the homicide rate per 10,000 persons from SESNSP as dependent variable. We use the inverse hyperbolic sine transformation in all cases. The controls we use include income per capita, total employment, and total population. The specifications with weights use the total population in the municipality on the pre-period. Standard errors are clustered at the municipality level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	INEGI Homicides		SESNSP Homicides		INEGI Homicide Rate		SESNP Homicide Rate	
Bartik	-2.5716	-1.6411	-2.6014	-2.3389	-1.7108*	-1.7961	-0.9848	-1.5428
	(3.023)	(4.825)	(3.484)	(2.570)	(0.926)	(1.246)	(1.717)	(1.084)
Observations	34,479	20,710	34,479	20,710	34,451	20,710	34,451	20,710
Municipality	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	N	N	N	N	N	N	N	N
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
	INEGI Homicides		SESNSP Homicides		INEGI Homicide Rate		SESNP Homicide Rate	
Bartik	-0.2100	-0.1447	-2.6315	-2.9215	-1.8310*	-1.9552	-1.3776	-1.3154
	(3.9812)	(4.6458)	(3.1261)	(3.1378)	(1.0982)	(1.2745)	(1.2816)	(1.3769)
Observations	34,452	20,710	34,452	20,710	34,451	20,710	34,451	20,710
Municipality	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	Y	Y	Y	Y	Y	Y	Y	Y

Table C.21: Effect of ATM-Sharing Agreements on Theft

Note: The table reports the results for the coefficient γ after estimating Equation 3.5. Observations are at the municipality-quarterly level. The dependent variable in Columns (1), (2), (7) and (8) is the quarterly change in the logarithm of total thefts. Columns (3), (4), (9) and (10) use the quarterly change in the logarithm of the theft rate per 10,000 persons. Columns (5), (6), (11) and (12) use the quarterly change in the logarithm of theft divided by total crimes. We use the inverse hyperbolic sine transformation in all cases. The controls we use include income per capita, total employment, and total population. The specifications with weights use the total population in the municipality on the pre-period. Standard errors are clustered at the municipality level.

	(1)	(2)	(3)	(4)	(5)	(6)
	Thefts		Theft Rate		Theft/Crime	
Bartik	1.1889	0.5210	1.4190	1.0828	0.4698*	0.6397**
	(1.615)	(2.626)	(1.544)	(2.416)	(0.252)	(0.253)
Observations	34,479	20,710	34,451	20,710	31,890	19,713
Municipality	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y
Weights	N	N	N	N	N	N
	(7)	(8)	(9)	(10)	(11)	(12)
	Thefts		Theft Rate		Theft/Crime	
Bartik	-1.8427	-3.7118	-1.2211	-2.7815	0.0890	-0.0525
	(2.6112)	(3.4255)	(2.1863)	(2.8910)	(0.3958)	(0.4797)
Observations	34,452	20,710	34,451	20,710	31,875	19,713
Municipality	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y
Weights	Y	Y	Y	Y	Y	Y

Table C.22: Effect of ATM-Sharing Agreements on Theft to Pedestrians

Note: The table reports the results for the coefficient γ after estimating Equation 3.5. Observations are at the municipality-quarterly level. The dependent variable in Columns (1), (2), (7) and (8) is the quarterly change in the logarithm of total thefts to pedestrians. Columns (3), (4), (9) and (10) use the quarterly change in the logarithm of the pedestrian theft rate per 10,000 persons. Columns (5), (6), (11) and (12) use the quarterly change in the logarithm of theft divided by total crimes. We use the inverse hyperbolic sine transformation in all cases. The controls we use include income per capita, total employment, and total population. The specifications with weights use the total population in the municipality on the pre-period. Standard errors are clustered at the municipality level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Ped. Thefts		Ped. Theft Rate		Ped. Theft/Crime		Ped. Theft/Theft	
Bartik	-4.1044*	-2.5892	-2.7902	-1.3942	-0.1043	-0.0176	-0.4934	-0.2392
	(2.446)	(2.416)	(1.806)	(1.213)	(0.090)	(0.072)	(0.314)	(0.261)
Obs.	34,479	20,710	34,451	20,710	31,890	19,713	28,716	18,678
Mun.	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	N	N	N	N	N	N	N	N
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
	Ped. Thefts		Ped. Theft Rate		Ped. Theft/Crime		Ped. Theft/Theft	
Bartik	-6.0923**	-6.1498*	-2.8432**	-2.7275	-0.0587	-0.0453	-0.1797	-0.1087
	(2.8218)	(3.4541)	(1.4173)	(1.7334)	(0.0904)	(0.1065)	(0.2775)	(0.3193)
Obs.	34,452	20,710	34,451	20,710	31,875	19,713	28,701	18,678
Mun.	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	Y	Y	Y	Y	Y	Y	Y	Y

Table C.23: Effect of ATM-Sharing Agreements on Total Crime

Note: The table reports the results for the coefficient γ after estimating Equation 3.5. Observations are at the municipality-quarterly level. The dependent variable in Columns (1) - (4) is the quarterly change in the logarithm of total crimes. Columns (5) - (8) use the quarterly change in the logarithm of the crime rate per 10,000 persons. We use the inverse hyperbolic sine transformation in all cases. The controls we use include income per capita, total employment, and total population. The specifications with weights use the total population in the municipality on the pre-period. Standard errors are clustered at the municipality level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Crimes				Crime Rate			
Bartik	-1.0344	-2.3154	-2.5049	-3.8172	-0.9004	-2.1698	-2.2962	-3.5126
	(1.293)	(2.187)	(1.943)	(2.497)	(1.250)	(2.105)	(1.756)	(2.270)
Observations	34,479	20,710	34,452	20,710	34,451	20,710	34,451	20,710
Municipality	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	N	N	Y	Y	N	N	Y	Y

Table C.24: Effect of ATM-Sharing Agreements on Informality

Note: The table reports the results for the coefficient γ after estimating Equation 3.5. Observations are at the municipality-quarterly level. The dependent variable in Columns (1), (2), (9), and (10) is the quarterly change in the logarithm of informal workers. The dependent variable in Columns (5), (6), (13), and (14) is the quarterly change in the logarithm of self-employed workers. The dependent variable in Columns (3), (4), (11), and (12) is the quarterly change in the logarithm of the ratio of informal workers and the total population of the municipality. The dependent variable in Columns (7), (8), (15), and (16) is the quarterly change in the logarithm of the ratio of self-employed workers and the total population of the municipality. We use the inverse hyperbolic sine transformation in all cases. The controls we use include income per capita, total employment, and total population. The specifications with weights use the total population in the municipality on the pre-period. Standard errors are clustered at the municipality level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Informality		Informality Rate		Self-Employed		Self-Employed Rate	
Bartik	-6.8998*	-2.7589***	-0.3055	1.0403	-8.0917***	-3.6340	-0.1529	0.4224
	(3.536)	(0.918)	(0.192)	(0.938)	(2.457)	(2.887)	(0.146)	(0.452)
Obs.	20,759	20,710	20,754	20,710	20,759	20,710	20,754	20,710
Mun.	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	N	N	N	N	N	N	N	N
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
	Informality		Informality Rate		Self-Employed		Self-Employed Rate	
Bartik	-6.4160	-1.6562	-0.1731	1.0269	-4.7946	0.6080	-0.0224	0.4637*
	(4.2298)	(1.2684)	(0.2742)	(0.6686)	(3.8115)	(1.8885)	(0.1554)	(0.2529)
Obs.	20,755	20,710	20,754	20,710	20,755	20,710	20,754	20,710
Mun.	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	Y	Y	Y	Y	Y	Y	Y	Y

Table C.25: Effect of ATM-Sharing Agreements on Local Taxes

Note: The table reports the results for the coefficient γ after estimating Equation 3.5. Observations are at the municipality-yearly level. The dependent variable in Columns (1) - (4) is the quarterly change in the logarithm of local taxes. Columns (5) - (8) use the yearly change in the logarithm of the ratio of taxes and the total population of the municipality. We use the inverse hyperbolic sine transformation in all cases. The controls we use include income per capita, total employment, and total population. The specifications with weights use the total population in the municipality on the pre-period. Standard errors are clustered at the municipality level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Taxes				Taxes/Population			
Bartik	-0.8494	-0.6120	-2.5409**	-2.8440**	-0.8176	-0.5366	-2.5222**	-2.7954**
	(0.636)	(0.941)	(1.118)	(1.385)	(0.629)	(0.912)	(1.104)	(1.370)
Obs.	6,031	3,822	6,028	3,822	6,028	3,822	6,028	3,822
Mun.	Y	Y	Y	Y	Y	Y	Y	Y
Period	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y
Weights	N	N	Y	Y	N	N	Y	Y

Table C.26: Summary of Rotemberg Weights and Over Id Tests

Note The table reports statistics on the Rotemberg weights. We aggregate the weights of a given industry across years as described in Goldsmith-Pinkham et al. (2020). Panel A reports the share and sum of negative weights. Panel B reports the top 5 agreements (ATM-card) according to their Rotemberg weights $\hat{\alpha}_k$ and g_k is equal to the national-level agreement shock $E_{kt}d\ln p_{kt}$. $\hat{\beta}_k$ is the coefficient of the just-identified 2SLS regression of the growth rate of pedestrian theft on the growth rate of ATM withdrawals using as instrument the agreement shares. Agreement share is the average agreement share multiplied by 100 for legibility. Panel C reports how $\hat{\beta}_k$ varies with the positive and negative Rotemberg weights. Panel D reports estimates of the 2SLS estimates, the growth rate of pedestrian theft on the growth rate of ATM withdrawals. Column TSLS (Bartik) uses the Bartik instrument. Column TSLS uses each agreement share (times time period) separately as instruments. The overidentification test corresponds to Hansen's J statistic. Standard errors are reported in parentheses and p-values in brackets.

Panel A: Negative and positive weights

	Sum	Mean	Share
Negative	-0.555	-0.035	0.263
Positive	1.555	0.050	0.737

Panel B: Top 5 Rotemberg weight agreements

	$\hat{\alpha}_k$	g_k	$\hat{\beta}_k$	Ag. Share
Scotiabank-Bancoppel	0.135	0.500	-1.576	0.206
Banjército-Bansefi	0.119	1.000	-0.051	0.004
Banco Azteca-Banco del Bajío	0.094	1.000	-4.691	0.000
Afirme-Scotiabank	0.092	1.000	19.352	0.003
Banco Ahorro Famsa-Scotiabank	0.090	0.600	-5.651	0.001

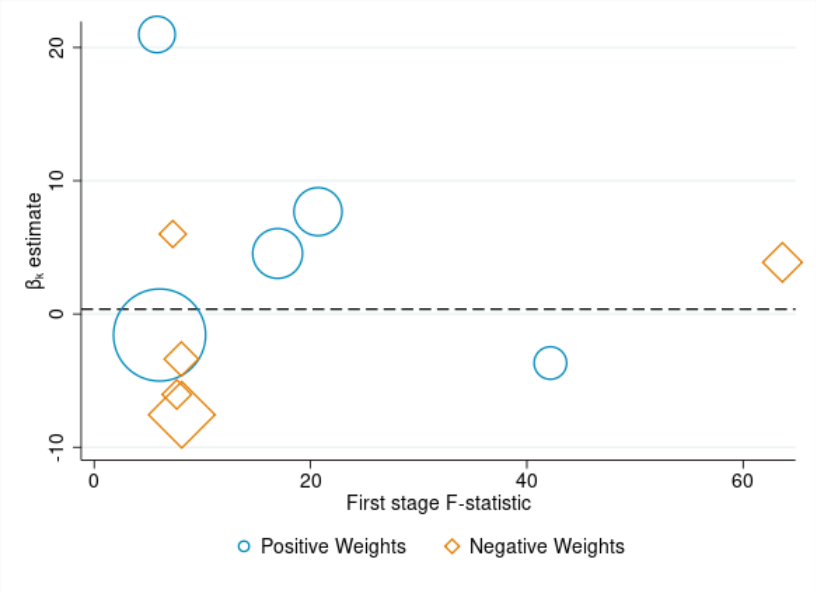
Panel C: Estimates of β_k for positive and negative weights

	α-weighted Sum	Share of overall β	Mean
Negative	-1.682	-1.969	-0.335
Positive	2.536	2.969	4.345

Panel D: Overidentification test

TSLS (Bartik)	TSLS	Over Id test
-3.67	0.36	1099.04
(2.29)	(0.18)	[0.38]

Figure C.17: Heterogeneity of β_k



Note: The graph shows the relationship between each instruments' $\hat{\beta}_k$, first-state F-statistics, and the Rotemberg weights. $\hat{\beta}_k$ is the 2SLS coefficient from the regression of pedestrian theft growth on ATM withdrawal growth rate instrumented by each share. Each point is a separate instrument's estimate (agreement share). The figure plots the estimated $\hat{\beta}_k$ for each instrument on the y-axis and the estimated first-stage F-statistic on the x-axis. The size of the points are scaled by the magnitude of the Rotemberg weights, with the circles denoting positive Rotemberg weights and the diamonds denoting negative weights. The horizontal dashed line is plotted at the value of the overall $\hat{\beta}$ reported in the TSLS column in Table C.26. The figure excludes instruments with first-stage F-statistics below 5.