Theory and Experiments on Social Norms and Social Image

by

Vera Louise te Velde

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Shachar Kariv, Chair
Professor Matthew Rabin
Professor Ernesto Dal Bó

Spring 2014

Theory and Experiments on Social Norms and Social Image

Abstract

Theory and Experiments on Social Norms and Social Image

by

Vera Louise te Velde

Doctor of Philosophy in Economics

University of California, Berkeley

Professor Shachar Kariv, Chair

This dissertation consists of three chapters exploring the role of social norms and social image in economic decision-making. A combination of theory and experiments is used to understand how certain social preferences might form, how they influence our individual choices through both internal and external motivation, and how they might aggregate into societal norms when individuals are externally pressured to appear to be doing the right thing.

The first chapter examines how individual beliefs, or personal norms, interact with the desire to maintain a positive social image when individuals disagree about what should be done in that situation. This commonly occurs in settings such as partisan politics or when moral norms are in flux. Traditional notions of social norms cannot describe these situations, and there is correspondingly no unambiguously "good" action that social pressure can promote. I develop two alternative, psychological game theoretic models that explain how social pressure affects behavior even if individuals do not agree on the norm. One channel is "approval seeking", in which individuals want their peers to approve of their actions. Another channel is "respect seeking", in which individuals want to be known for strict adherence to their personal norms. Approval seekers pool on one option as social pressure increases, thus creating the illusion of societal consensus. This can potentially leading to destructive posturing, in which damaging norms are perpetuated. Respect seekers, on the other hand, are less hypocritical the more social pressure increases, and are accordingly less willing to compromise. Respect is thus a less likely force to produce a societal consensus norm. These results demonstrate that using social pressure to promote a certain behavior may backfire if it targets the wrong kind of social image.

In the second chapter, Ulrike Malmendier, Roberto Weber and I explore the case of the norm of reciprocity, both positive and negative. Reciprocal behavioral has been found to play a significant role in explaining outcomes in many important economic domains. However, despite mounting empirical evidence, economists still struggle to converge on the correct model of the underlying motives. Existing theories posit internal preferences for the welfare of others, inequality aversion, or utility from repaying others' kindness. Recent evidence

reveals that 'one-sided' acts of kindness, not involving reciprocity, exhibit a large degree of reluctance, with people trying to avoid opportunities to act generously, suggesting that external factors such as social image, self image, or social pressure are important determinants of unilateral sharing. However, this revision of conventional theoretical motives for sharing has had little spillover to 'two-sided' reciprocity environments, where one individual responds to the actions of another. We review the literature on reciprocity and point to the relative lack of attention paid to external factors. We then present a novel experiment that explores the importance of internal versus external factors in driving reciprocal behavior. We find that, in a laboratory reciprocity setting (the double-dictator game), failure to account for external motives leads to a significant overestimation of internal motives such as fairness and altruism. We use the experimental data to illustrate the importance of combining reduced-form and structural analyses in disentangling internal and external determinants of pro-social behavior.

In the third chapter, Pamela Jakiela, Edward Miguel and I look deeper into the origins of norms. We combine data from a field experiment and a laboratory experiment to measure the causal impact of human capital on respect for earned property rights, a component of social preferences with important implications for economic growth and development. We find that higher academic achievement reduces the willingness of young Kenyan women to appropriate others' labor income, and shifts players toward a 50-50 split norm in the dictator game. This study demonstrates that education may have long-run impacts on social preferences, norms and institutions beyond the human capital directly produced. It also shows that randomized field experiments can be successfully combined with laboratory experiment data to measure causal impacts on individual values, norms, and preferences which cannot be readily captured in survey data.

To Matthew

Contents

List of Figures

List of Tables

Acknowledgments

I can't possibly thank my adviser, Shachar Kariv, enough for his endless encouragement and advice over the last several years. I am forever indebted to him for the seemingly thousands of hours spent editing my work and thousands of emails sent on my behalf during the job market.

On more than one occasion in the many supremely helpful discussions we had about my research, Matthew Rabin made an offhand remark that changed the course of my research when I fully comprehended it months later. He is the model economist I (futilely) strive to emulate.

Ulrike Malmendier, Edward Miguel, and Stefano DellaVigna, were endlessly thrilling to work with, and every encounter with any of them forced me to push the boundaries of my/our work in directions I never even anticipated. Pamela Jakiela and Roberto Weber were similarly wonderful coauthors who always willingly put in the extra effort required to communicate virtually.

This work also benefited from helpful conversations with countless others, including Ernesto Dal B, Gerard Roland, Richard Thaler, Eugene Caruso, Klaus Schmidt, Ted O'Donoghue, Charles Sprenger, Muriel Niederle, Doug Bernheim, Todd Rogers, David Hirschleifer, Aaron Bodoh-Creed and many seminar audiences. Patrick Allen and the rest of the department staff also made my life much easier by insulating me from Berkeley bureaucracy, and I am extremely grateful to the National Science Foundation and the Berkeley Center for the Economics of Demography and Aging for their financial support of my graduate studies.

My classmates and officemates, too many to name, were what made Berkeley the only place I wanted to be. They were always available to commiserate and reassure me that I wasn't alone in the deep end. After work hours, Joanna Noble, Ken Manne, and my family graciously put up with me after unwittingly being dragged along for this adventure. I can only hope that I will have the opportunity to reciprocate.

Finally, Matthew Peddie is the love of my life who had the astounding patience to put up with me as I was staring down the barrel of a graduation deadline. He has been my rock without whom I would not have made it through the last three years. In a more honest world, he would have his name on more than the dedication page.

Chapter 1

Heterogeneous norms How can social pressure influence behavior when people disagree?

1.1 Introduction

Over the last decade, social image and social pressure have emerged as key determinants of prosocial behavior in the economics literature. They have been exploited to increase voter participation in a cost effective way(Gerber, Green, and Larimer 2008), to increase donations to charity by up to 42% (DellaVigna, List, and Malmendier 2012), and to promote safe sex practices when financial incentives fail (Ashraf, Bandiera, and Jack 2012). But settings like voting and donating to charity, in which there is an essentially unanimously recognized "good" action, are rare: far more common are the decisions about which people disagree, such as *how* to vote, *how* to allocate resources, how to raise children, what dietary and religious habits to keep, or what customs to follow.

Existing models typically assume a unanimous social norm that governs behavior in a given setting and that individuals adhere more or less closely to that norm by trading off utility components such as consumption, image, and moral appropriateness. However, identifying the consensus norm is rarely straightforward, and it usually entails choosing an arbitrary point from the universe of opinions as the "norm".[1] Even simple laboratory settings such as dictator games, which exist in enough of a contextual vacuum to have an easily-identifiable norm, never perfectly map to actual allocation decisions. In the real world, matters are complicated by transfer costs, heterogeneous endowments and production abilities, heterogeneous preferences, and countless other factors. People do not even superficially agree

[1] As an example, DellaVigna, List, and Malmendier (2012) declare that giving $10 to charity is the "acceptable" minimum level of generosity. While this assumption helps to estimate a model of social pressure from real-world choices, which do display at mode at $10, it's clearly an ad hoc simplification of an issue that individuals have very different opinions about.

in other domains such as partisan political issues, child-rearing practices, dietary choices, religion, and lifestyle.

How can social norms work if some people believe voting is irrational, giving at the door is undesirable, or safe sex practices are a sign of weakness?[2] In such settings with heterogeneous norms, existing notions of social norms break down. In fact, different norms may pull in different directions: marching in a political protest can be admired as a principled move or shunned as antisocial rocking of the boat. Adopting new dress habits upon being promoted at work might be disrespected for being inauthentic or welcomed by those who prefer a formal dress code. In other words, the tradeoffs people are willing to make between self-interest, moral beliefs, and social image may be different depending whether they are more concerned with respect or approval.

Referring to "heterogeneous" norms is in fact an oxymoron by traditional definitions, which defines norms as unanimously recognized behavioral guidelines (Bicchieri 2006; Ostrom 2000).[3] In this paper, I propose to broaden this definition and to allow each individual to hold a distinct idea of the impartial appropriateness of each often. I redefine the consensus norm correspondingly to reflect an aggregate agreement that a particular action is the most moral. That is, personal norms prescribe a particular action as morally ideal, regardless of what others do. Descriptive norms (Bicchieri 2006; Cialdini, Kallgren, and Reno 1991) describe what people in fact do and are irrelevant. This rules out personal norms of the form "do what everyone else is doing".

This individualized notion of social norms requires a new understanding of social-image motivations, which is the second contribution of the paper. When people disagree about norms, the social pressure induced by norms becomes unclear. Being vegetarian might be a convincing signal of your concern for animal rights while still provoking disdain from the omnivorous majority. I therefore formulate two distinct types of social image, approval and respect, that are clearly defined even when norms are heterogeneous. The analysis shows that while traditional models of social-image motivations are extremely powerful in homogeneous norm domains, those resulting predictions constitute a knife-edge case within the broader universe of heterogeneous norms.

I propose a novel approach to modeling these two types of social image which are applicable when individuals honestly disagree about the correct course of action. The first type are the "respect seekers", who signal their adherence to their ideals. Respect seekers admire integrity, i.e. always doing what you personally believe to be right, and wish to be seen as having integrity. In their decisions, they attempt to signal their integrity, and similarly, they infer others' integrity and pass judgment on them accordingly. Others may have different ideals, but as long as those beliefs are applied impartially, respect seekers agree to disagree

[2] This paper concerns issues that people disagree about. As such, discussion of many sensitive topics is inevitable. The reader should not attempt to infer my personal beliefs from examples or hypothetical scenarios.

[3] Krupka and Weber (2013) broaden this definition to a set of commonly-recognized "moral appropriateness" ratings of *every* option.

and maintain mutual respect. Egalitarians and utilitarians, for example, respect each other's choices as long as they are made as though behind the veil of ignorance.

Since integrity is invisible must be inferred from choices, respect seekers would like to make choices that are most strongly attributable to high integrity. They might be vegan, for example, in order to show that they are willing to make personal sacrifices to stay true to their beliefs about animal cruelty. Or they might affiliate with groups (in politics, religion, or hobbies) that are known for being die-hard about their causes. In equilibrium, respect seekers interpret behavior based on the (rational expectations) likelihood that someone is faking his way to a good reputation, or conversely, that someone who appears to be selfish is in fact just doing what he believes in.

The second type, the "approval seekers", want their actions to be admired. They aim to make their peers happy by going along with everyone else's beliefs, and they might abandon their own ideals in order to do so. Vice versa, an approval seeker wants everyone else to follow his personal beliefs and judges anyone who doesn't live up to his personal standard of behavior. Approval seekers judge actions, rather than inferred types, so they play a non-signaling game in which individuals simply choose the best trade-off between consumption, guilt, and image, without aiming to convincingly imitate other types. A vegetarian approval seeker might be tempted to go along with friends to a Brazilian buffet, but he would be happy to convert a friend to vegetarianism even if he knew it was for lack of enjoying meat rather than for heartfelt moral reasons.

For both respect seekers and approval seekers, image concerns become relatively more important as social pressure rises: as vegetarianism becomes the hot topic at the water cooler, approval seekers care more about whether others admire their dietary choices, and respect seekers care more about proving that their dietary choices reflect their true beliefs. Respect seekers respond to this pressure by trying harder to imitate high integrity types. They look for convincing ways to signal high integrity, possibly using costly signals to establish credibility. They might give up meat or purchase pricier cage-free eggs and free-range chicken, to signal that they care about animal rights, or they might join hunting clubs and spend their weekends protesting PETA in order to convince others that their meat-eating lifestyle is a moral choice. Choices can become *more* divided along moral lines, preventing conformity, as individuals become more hesitant to try passing off selfishness as a morally motivated decision. Compromise becomes harder to achieve as individuals become less willing to be seen conceding their norms. The "vegan before 6" Mark Bittman followers might turn vegan 24/7, while the apathetic omnivores join meat CSAs and become sausage connoisseurs.

Approval seekers, on the other hand, respond to social pressure by trying harder to please the majority. This leads to conformity: the more important image motivations are, the more power the majority has to sway choices. Approval-seeking vegetarians will hide their preferences and go along with friends to the sushi bar or give up on herbivorism altogether. But when possible, approval seekers may prefer to look for ways to avoid offending any one subgroup too much, leading to widespread compromise. Individuals of all beliefs might be able to agree that free-range chickens and farm-raised fish are fine to eat but that beef is not, thereby avoiding offending either the paleo dieters or the animal rights activists. In any case, high social pressure leads to deceptively uniform behavior, disguising the disagreement in the population.

I also find that approval seekers and respect seekers have differing abilities to sustain "sacrificial" equilibria in which individuals forego both consumption *and* guilt utility in order to attain a better social image. If approval-seeking vegetarians could claim a majority and then pressure everyone else to reluctantly follow, this could lead to better environmental or health outcomes. On the other hand, if a majority of approval seekers have bad intentions, this could create a disaster as they bully their peers down a terrible path. Respect seekers, on the other hand, are less likely to be able to sustain either kind of conformist equilibrium, for better or for worse.

Lastly, I show that approval and respect can interact when simultaneously relevant, such that social disapproval can *itself* serve as a signal of integrity. Omnivores might be able to prove that their habits are honorable merely by following them openly in a disapproving group of vegetarians.

The contrast between the outcomes of these two types, and the implied difficulty with using social pressure to influence behavior without understanding what type of social-image motivates people in that setting, is the fourth contribution of this paper. As described above, the two models of social image make distinct prescriptive predictions regarding persuasion and social pressure. For example, if politicians and voters are approval seekers who want to end legislative gridlock, they may do well to increase the salience of social image in order to motivate politicians to compromise in order to avoid offending any major constituent group too badly (so long as no single constituent group dominates the others). But if politicians are respect seekers, social pressure should be minimized if there is to be hope for agreement. In a somewhat different domain, a group seeking to foster group solidarity by extending membership to a certain number of the most committed individuals (or a company selling a product associated with a group's identity) should set a high membership price to attract respect seekers who wish to use group membership as a costly signal of integrity. But if the group stands for a minority opinion, only a low price will attract the same number of approval seekers, the few who feel strongly enough about being true to their norms to join the group regardless of the social-image cost. And in yet another domain, a development economist wishing to increase teacher attendance should be careful to identify the image motivations that teachers care about. If teachers derive respect from doing their job well, publicly honoring good work may be effective, but if they don't wish to be ostracized by other, less dedicated teachers, this approach may backfire.

These models are widely applicable to a variety of fields beyond these few examples, and I consider some possibilities in section 1.5. Until then, section 1.2 will discuss how this work fits with related literature, section 1.3 will formalize the two models of social image, and section 1.4 will explore the different behavior of approval seekers and respect seekers.

1.2 Background

This paper builds on the theoretical social preferences literature, which is accompanied by a vast empirical and experimental literature in economics and sociology on how social norms and image motivations influence behavior.

Economists have been trying to understand moral choice since Becker's (1974) model of pure altruism, in which individuals directly have utility over others' outcomes. This was soon modified to account for "warm glow" motivations (Andreoni 1989, 1990). Later models made moral behavior contingent on relative standing, reciprocity, or intentions (Bolton and Ockenfels 2000; Charness and Rabin 2002; Fehr and Schmidt 1999; Levine 1998; Rabin 1993).

More recent findings have suggested that we may not be so internally altruistic as these models would suggest.[4] First of all, reducing anonymity, by revealing individual decisions (even with noise) or by revealing the decisionmakers, can dramatically increase cooperation (Andreoni and Petrie 2004; Bohnet and Frey 1999b; Carpenter 2005; Cason and Khan 1999; Franzen and Pointner 2012; Hoffman et al. 1994; Koch and Normann 2008; Satow 1975; Sell and Wilson 1991; Soetevent 2005). People will even give up money to avoid being put in a position where they will have to anonymously make trade-offs between their own and others' payments (Broberg, Ellingsen, and Johannesson 2007; Dana, Cain, and Dawes 2006; Lazear, Malmendier, and Weber 2012; Malmendier, Velde, and Weber 2013).

Evidence from the field confirms these findings and highlights their extreme importance. Gerber, Green, and Larimer (2008) find that showing publicly available voter records to neighbors is the most cost effective way to increase vote turnout. In community health initiatives, social recognition has proven to be a more effective incentive than either small or large financial rewards (Ashraf, Bandiera, and Jack 2012), and Karlan (2012) shows that this cannot be explained by an altruistic desire to motivate others by example. In charitable giving, DellaVigna et al. (2013a) show that social pressure increases donations by up to 42%, and Meer (2011) and Rind and Benjamin (1994) show that the effect of social pressure is amplified by prior ties to the observer. There is even evidence that soccer referees favor home teams, but only when games are close (Garicano, Palacios-Huerta, and Prendergast 2005). In the workplace, Babcock et al. (2010) show that incentivizing teams can be more effective than individuals because inactive individuals don't want to let down their peers,

[4] This literature almost universally attributes these findings to social-image concerns and social pressure. However, a couple of other factors may account for some fraction of the effect. Self-image is understood to influence behavior similarly to social image when individuals use their own actions to self-signal (Bénabou and Tirole 2011) or when social pressure merely forces individuals to be more objective in their self-evaluation (Malmendier and Velde 2013; Rabin 1995). On the other hand, alleged effects of social image might have a much less cynical explanation, from beliefs-based altruism (Grossman 2013; Malmendier and Velde 2013; Velde 2013). If individuals don't care directly about their image, but do care about not hurting others' feelings, then having their decisions observed more closely could still push them to behave more generously. While acknowledging these potential confounding explanations, I will stick with the terminology of social image and social pressure in this paper, not least because other forms of image motivations (such as respect seeking) are not likely to be as confounded by these other explanations. Also see Posner and Rasmusen (1999) for related discussion on the various ways in which norms persist and are automatically enforced.

and Bandiera, Barankay, and Rasul (2005) show that altruism alone cannot account for this; only when workers are subject to social pressure due to peer monitoring does the effect occur. Bandiera, Barankay, and Rasul (2010), Falk and Ichino (2006), and Mas and Moretti (2009) further show that low-productivity workers improve when monitored by higher productivity workers and that this persists even if they can't observe the higher productivity workers, suggesting that social image must be explanation.

On the theory side, progress has been made, but we are still lacking a generally applicable model of social image. In many settings, however, these various models incorporating social image have found strong support. DellaVigna, List, and Malmendier (2012) show that people will give a token amount to door-to-door charity solicitors in order to avoid the social cost of saying no. Harbaugh (1998b), Harbaugh (1998a), and Glazer and Konrad (1996) model the prestige motivation for charitable giving and find empirical evidence from the desire of donors to be listed in elite categories of supporters, and Carpenter and Myers (2010) similarly find that volunteers with stronger image concerns prefer visible jobs leading to more social recognition. Andreoni and Bernheim (2009) find that dictator game players are motivated to signal their adherence in a 50-50 sharing norm, and Grossman (2012) develops a similar model to explain behavior in a binary choice environment. Bénabou and Tirole (2006) and Seabright (2009) use similar altruism signaling models to explore crowding out when financial incentives are given for prosocial behavior. Battigalli and Dufwenberg (2007) model moral choices as a way to avoid guilt. A plethora of lab experiments, field experiments, and empirical evidence support this theoretical approach (Broberg, Ellingsen, and Johannesson 2007; Dana, Cain, and Dawes 2006; Dana, Weber, and Kuang 2007; Lazear, Malmendier, and Weber 2012; Malmendier, Velde, and Weber 2013; Mas and Moretti 2009).

Most of this literature focuses on the particular case of altruism or charitable giving, but a few models are applicable to more general cases. Bernheim (1994) models conformity to an arbitrary fixed norm with a signaling model in which players infer each other's dedication to the norm from their choices. He shows that the signaling motivation endogenously produces pooling on the norm (similar to the main result in the dictator game model of Andreoni and Bernheim (2009)). Bénabou and Tirole (2011) model moral behavior through a self-signaling process, in which individuals invest in their identity by making easily remembered moral choices. Akerlof (1980) models social customs in which individuals feel social pressure to adhere to a norm and social pressure increases in the fraction of the population that believes in the norm. Akerlof and Kranton (2000) model choices as influenced by the identities we derive from membership in different social categories.

The preceding models have a common limitation, however: all assume that people unanimously agree on the best action. I correct this in each of my parallel models of social image.

The issue of wealth distribution is a key example that intuitively bridges the gap between traditional models of altruism and these models of moral beliefs. My model of respect seekers uses a signaling framework similar to Andreoni and Bernheim (2009), an influential study showing remarkable evidence for the role of social image in dictator game sharing. But in the simple dictator game, any impartial philosophy of redistribution agrees that the 50-

50 split is the morally correct action. No real-world redistribution decision falls within this knife-edge case. Even absent complications such as earned property rights, individuals exhibit substantial heterogeneity in their distributive preferences (Alesina and Glaeser 2004; Butler, Giuliano, and Guiso 2012; Cappelen et al. 2007; Charness and Rabin 2002; Dickinson and Tiefenthaler 2002; Fisman, Kariv, and Markovits 2007; Reuben and Riedl 2011). By reinterpreting the signaled types of Andreoni and Bernheim (2009) as integrity towards these *personal* redistributive ideals, however, I can generalize their model to apply to these more realistic settings with disagreement about notions of fairness. The respect seekers in my integrity signaling model must then jointly infer both integrity *and* ideals themselves in order to pass judgment on someone's actions, and judgment represents an inference of impartiality, rather than prosociality.

My model of approval seekers, on the other hand, uses a notion of social image similar to the one in Akerlof (1980). But while he, too, assumes that a particular fraction of the population believes in a norm and passes judgment on those who defect from that norm (and the higher the fraction is, the harsher is social pressure), I more broadly assume that every individual in a population judges others' actions according to the same moral standard as he judges his own actions. Since individuals have different ideals, it can be impossible to please everyone, but a maximal social image can be achieved by minimizing the harshness with which an average peer judges one's action. Often the way to do this is to follow the lead of the majority, and other times it pays to avoid offending anyone too extremely by taking a middle ground path. Even in settings without moral content, the desire to fit in can be described by the approval-seeking motive; approval seekers, but not respect seekers, might make transparently desperate attempts to stick with the crowd, as in Asch's (1955) or Milgram's (1963) seminal experiments, or conform apparently purely for the sake of conforming, as in Goeree and Yariv (2007).

In building on these models and focusing on the importance of social image that has emerged in the empirical and experimental literature, I will abstract from other aspects of social preferences, such as warm glow or altruism. This is primarily because they are intended to be portable across many situations which may or may not involve prosocial considerations. In particular cases, of course, these factors may be understood to influence the "consumption utility" of a particular choice, or the set of impartial norms that are actually held.

Additional background literature that is relevant to particular applications is mentioned along with discussion of those applications and in section 1.5.

1.3 Model

Consider a setting in which individuals within an (observant) population must take a morally contentious action. Different personal norms correspond to different ideal actions. Each option provides the individual a certain consumption utility, which includes the immediate costs and benefits of the action along with any expected long-run change in utility, such as the expected change in tax policy after volunteering to campaign for a particular candidate.

Externalities are not specified; this is discussed further in section 1.4. This setting is intended to intuitively capture a wide array of moral and customary decisions, such as whether to eat meat, which church to go to (if any), whether to send one's kids to private school, how to share resources, how to reciprocate kind actions, what to wear to work, what brand of shoes to buy, who to vote for, etc.[5]

Formally, an individual i is faced with a choice set X, which I will take to be a binary or ternary choice in the results below. Each option $x \in X$ leads to consumption utility $v(x)$, but in addition, i has a personal norm denoting ρ_i as the morally appropriate choice. When making a choice x_i that deviates from this ideal outcome, i feels guilt $G(x_i - \rho_i)$, which is additionally weighted by an integrity parameter t_i[6]. That is, each person has a two-dimensional type (t_i, ρ_i): an ideal choice and an integrity parameter that constitutes a weight on their personal norm.[7]

When types are specified in this two-dimensional way, there is a natural choice in the definition of social image: do people want others to share their moral beliefs and to signal that they share theirs with the group, or do they want to signal that they adhere to their own idea of right and wrong no matter what material consequences are involved and no matter what other people think? This is the intuition driving the formulation of two models of social image: respect seekers signal integrity, and approval seekers signal their moral ideals.[8]

Social image utility is an increasing function given by $H(m(x_i))$, where m is the image resulting from a given choice. The importance of image is determined by a social pressure parameter s, which enters utility as a weight on H. s summarizes the *shared* attributes of a situation that contribute to social pressure: visibility, audience size, and how harshly choices are judged in a particular setting. Image itself is the key ingredient for which I consider multiple possibilities, discussed below.

Altogether, person i has utility

$$U(x_i|t_i, \rho_i) = v(x_i) - t_i G(x_i - \rho_i) + sH(m(x_i)). \tag{1.1}$$

[5] While voting and similar choices are themselves private, it's reasonable to interpret many people's voting decisions as visible amongst their social group, as these topics frequently arise in conversation and there is substantial evidence that people dislike lying, in general and about voting in particular, e.g. Gneezy (2005) and DellaVigna et al. (2013b). Moreover there is evidence that social-image motivations strongly influence even anonymous choices, e.g. Lazear, Malmendier, and Weber (2012) or Malmendier, Velde, and Weber (2013), and, as mentioned in footnote 4, evidence that self-image motivations can operate similarly to social-image motivations when we use actions to self-signal (Bénabou and Tirole 2011; Rabin 1995).

[6] This formulation asserts that people differ in their norms but not in their guilt function G, whereas the personalized definition of norms described in the introduction would imply a fully individualized set of appropriateness ratings (given by the function G). This is a simplifying assumption that makes the model tractable.

[7] Rather than outright disagreement, this setup could represent *ambiguity* about the norm, such as in a new situation or an unfamiliar culture. Someone who is very confident about their belief in the norm ρ will experience guilt normally, while someone who has no idea what the norm is will not experience much guilt from any choice; that is, they will weight G with a very small t.

[8] Nonetheless, approval seekers are modeled without a signaling model, for reasons explained below.

Further assumptions may be helpfully motivated with an example.[9] Imagine that the relevant choice is to argue for a particular tax rate to fund a welfare system. The decisionmaker believes that ρ represents the best trade-off between equality and efficiency but would rather not pay such a high rate, so he might facetiously argue against social safety nets. He may not even have the option of voting for his most preferred tax rate, if only a few options are on the table. Others disagree with his moral beliefs, and some argue in favor of credible alternatives, while some argue for transparently selfish policies that no one believes are fair. Any of these morally contentious decisions can be understood to be contained within the general scenario specified by the following intuitive, non-restrictive assumptions:

Assumption 1. *i) X is indexed by $\mathbb{R}$; results below specify $X = \{x_1, x_2\} \in \mathbb{R}^2$ or $X = \{x_1, x_2, x_3\} \in \mathbb{R}^3$. $s \geq 0$.*

ii) $(t_i, \rho_i) \sim \phi$, continuous, with conditional distributions satisfying $\operatorname{supp} \phi_t = \mathbb{R}^+$ and $\operatorname{supp} \phi_\rho \subset X$. These distributions are commonly known.

iii) G is continuous and monotonically increasing in $|x_i - \rho_i|$ and is symmetric around 0.

iv) H is continuous and monotonically increasing in $m(x_i)$, and $\sup_m H(m) = \overline{H} < \infty$.

Continuity is assumed simply to guarantee existence of equilibrium; this could be relaxed in specific instances to allow, for example, an atom in the distribution of types.

I consider two hypothetical, distinct populations of people with two different types of social-image motivations, the *approval seekers* and the *respect seekers*. They respectively have social-image functions m_{as} and m_{rs}, and image utility functions H_{as} and H_{rs}, for which part 4 of assumption 1 holds separately.

Approval seekers and respect seekers are treated as disjoint populations for the sake of clearly contrasting the effects of the two kinds of social image, but of course in many situations these two motivations might operate simultaneously; Section 1.4 explores the interaction between approval seeking and respect seeking in these situations.

Approval seekers: Approval seekers derive utility from praise for their actions, and observers praise actions that agree with their personal norms. They are motivated to adhere as closely as possible to the ideals of the highest fraction of the population as possible. Note that approval seekers are not concerned with actually signaling either their ideal ρ or their integrity t; they only seek praise that depends on others' moral beliefs. This is superficially similar to wanting to signal that you share your beliefs with someone else, but I opt not to

[9] One could imagine more complicated utility specifications which, for example, relax additive separability or further individualize the component functions or allow for asymmetric guilt. I do not wish to commit to ruling out these possibilities, but I present the utility function in this way for ease of exposition. The key results come from contrasting the predictions of this model when m and H represent different types of social image, and generalizing the functional form further doesn't substantively contribute to that discussion.

use such a signaling model because it immediately leads to counterintuitive predictions: a vegetarian would approve of an admittedly-hypocritical lapsed-vegetarian friend just because they philosophically agree about the merits of vegetarianism. On the other hand, however, it's quite plausible that a vegetarian would be happy to convert an insincere meateater, or that Republican constituents would be happy to continue reelecting a moderate Democrat, so long as that representative denies any true pro-choice beliefs. These are scenarios in which individuals confer approval, but *not* respect (as defined below).

For approval seekers, $m(x)$ is defined as $m_{as}(x) = -\int_{-\infty}^{\infty}\int_{0}^{\infty}\phi(t,\rho)G(x-\rho)dtd\rho$. That is, each observer judges the choice of x according to his personal guilt function, and image is the negative of the average of these individual judgments over the full population. For example, if half of the population believes in ρ_1 and half the population believes in ρ_2, then $m(\rho_2) = -\frac{1}{2}G(\rho_2 - \rho_1)$. The best attainable image, $m_{as} = 0$, only occurs when perfectly adhering to a homogeneous norm.

Respect seekers: For respect seekers, social image is based on others' estimate m of the decisionmaker's integrity t. Formally, their image function is simply defined as $m_{rs}(x) = E[t|x]$, the rational inference of someone's t conditional on the choice x.[10] Individuals want to be seen as unhypocritical, whatever their personal beliefs. However, since integrity is not directly observable, inferences about integrity must be rational in equilibrium, and optimal choices must anticipate those equilibrium inferences. Note that observers do *not* care about the dictator's ρ. That is, observers do not judge norms themselves when multiple defensible options exist, but they do judge the decisionmaker's integrity or hypocrisy with respect to his beliefs. There are plausibly many such situations in which people care about respect but *not* approval. Outsiders admire some extreme religious groups such as the Amish or the Hasidic Jews for their dedication to their beliefs. Or, from the perspective of the actor rather than that of the audience, the actor might care only about respect even as the world disapproves. Suicide bombers are definitely not admired, but it's also hard to claim that their actions do not convincingly signal their dedication to their beliefs.[11]

Notice that for both approval seekers and respect seekers, there is a best possible image. Approval seekers can't do any better than to perfectly please everyone in the population, and respect seekers can't do any better than to be known to be perfectly impartial. It is intuitive that perfect image can't lead to unboundedly high utility; this is the motivation for the upper bound on H stated in part 4 of assumption 1. This assumption will also

[10] This assumption could be generalized to allow for inferences other than the mean, but this flexibility isn't important to the results in this paper, so I use the mean for expositional simplicity. However, all proofs go through just as well with other inference functions such as the median. Specifically, any inference function m, acting on the (rational, in equilibrium) distribution of types f that would choose the observed action is suitable so long as it satisfies the following properties: 1) m is continuous with respect to the weak topology on the space of distributions on $\mathbb{R}^+$, 2) $\min \operatorname{supp} f \leq m(f) \leq \max \operatorname{supp} f$ with strict inequalities if the support of f is nondegenerate, and 3) if f' first order stochastically dominates f, then $m(f') > m(f)$.

[11] I speculate that the unconvincing rhetoric labeling these people as cowards is in fact a somewhat facetious attempt to denigrate their *single* potentially admirable attribute.

provide mathematical utility by restricting equilibrium parameters to a compact space and guaranteeing existence of an equilibrium.

The dependence of utility on beliefs places the model in the realm of psychological game theory (Battigalli and Dufwenberg 2009; Geanakoplos, Pearce, and Stacchetti 1989). As long as types are exogenously assigned, however[12], approval seekers are not playing a strategic equilibrium in which inferences about types matter, so the tools from psychological game theory are not needed to analyze their outcomes. The respect seekers play a much more complicated signaling game, and analysis relies on that notion of psychological equilibrium.

Homogeneous Norms Before presenting the main results, a short note on the baseline setting with homogeneous norms is called for. Even in these scenarios, the two models lead to slightly different predictions. These differences arise fundamentally because respect seekers are playing a signaling game, while approval seekers are being judged directly for their actions rather than inferences based on those actions.

First of all, the respect-seeking model can generate pooling equilibria in situations where the approval-seeking model can't. This is demonstrated by Andreoni and Bernheim (2009) (or Bernheim's (1994) model of conformity) which can be seen as a (continuous choice) special case of the model of respect seekers. Their simpler setting with homogeneous, exogenous social norms shows that respect seekers playing a signaling game create endogenous discontinuities in their preferences that lead to pooling behavior. In particular, they show that respect seekers exhibit pooling on the 50-50 split in the dictator game, despite the fact that preferences have no discontinuities at this point.[13]

In their model, dictators choose an amount x between 0 and 10 dollars to keep, and the remainder of the ten dollar goes to their anonymous partners. A desire to signal adherence to the 50-50 norm attracts anyone who would like to choose nearly 50-50 to the exact 50-50 split, so that a pool occurs at the fair outcome and no one chooses nearby allocations. The reader is referred to the original paper for the proof, but intuitively speaking, this results from an endogenous discontinuity in the inference function: When a mass of high t types choose 50-50, the inferred type $E[t|x=5]$ must be the mean member of that pool, which then must be strictly larger than the inferred type upon observing $x = 5 \pm \epsilon$. These discontinuous inferences creates the discontinuous incentives that create the pooling behavior in the first place and ensure that anyone "nearby" will prefer to jump to the exactly fair choice. In equilibrium, increasing social pressure increases the size of this pool.

Approval seekers, on the other hand, merely have an increased motivation to get near 50-50 when social pressure increases, and their distribution of choices will smoothly approach 50-50 but will not discontinuously pool there. Say, for example, that consumption utility is linear and $g(x-5) = (x-5)^2$. Then approval-seeking dictators maximize $x - (t+s)(x-5)^2$, yielding $x^* = \frac{1}{2(t+s)} + 5$. Increasing social pressure moves the distribution of choices towards $x = 5$, but no pooling occurs, and other types will continue choosing nearby allocations as

[12] See Appendix A.1 for an extension that relaxes this exogeneity condition.

[13] Bernheim (1994) explicitly emphasizes this same difference, but that model is a bit harder to adapt to my notation, so I use the Andreoni and Bernheim (2009) setting instead.

well. The pooling outcomes generated by respect-seeker signaling only occur for approval seekers if we allow incidental structural features in their preferences.

Second of all, the respect-seeking model produces interdependent preferences more naturally than the approval-seeking model. Interdependent preferences exist when an individual's preferences depend on the preferences of others (Gul and Pesendorfer 2006; Postlewaite 2011). For example, someone might be more altruistic towards someone who is very altruistic (Levine 1998). There is indeed plenty of evidence that people's revealed preferences depend on how strictly their peers are adhering to a norm (Andreoni and Scholz 1998; Carman 2004; Cialdini 2003; Croson, Fatas, and Neugebauer 2005; Fischbacher, Gächter, and Fehr 2001; Frey and Meier 2004; Gino, Ayal, and Ariely 2009).

This behavior is easily understood in the respect seekers' signaling game, especially if we grant that isolated experiments do not reflect *equilibrium* outcomes; that is, while equilibrium beliefs must reflect actual choices, real-world choices reflect evolving, learned beliefs about others' strategies. Let's continue to use the dictator game example from above. Suppose a respect seeker first observes his peers playing a dictator game in which the the average gift is 0. He then knows that choosing to give 1 will cause his peers to infer that he has a fairly high t. If he then later observes his peers playing the same game and choosing an average gift of 3, he knows that to attain the same social image, he must give more than 3: the underlying distribution of types can't have changed, but each type is now sharing more, so he must also share more to maintain a particular image.

Even if isolated experiments *do* reflect equilibrium outcomes, interdependent preferences of respect seekers are easily understood. The signaling game that they play can have multiple equilibria (see section 1.4), depending on the underlying distribution of types, so individual preferences are contingent on the equilibrium that others are playing.

Note that modifying the definition of image for approval seekers such that actions are judged relative to others' actions will also allow for interdependent preferences. For respect seekers, however, this behavior arises endogenously.

Having completed the model setup, we can now turn to the main results describing how respect seekers and approval seekers behave in settings with heterogeneous norms.

1.4 Results

The following five subsections characterize the behavior of approval seekers and respect seekers facing several types of choices. Despite the simplicity of the discrete choice environments considered, this provides the key intuitions for the differences between the two types.

Equilibrium

First assume that each individual has exactly two options, x_1 or x_2, and one of two ideals, $\rho_1 = x_1$ or $\rho_2 = x_2$. Each option provides consumption utility $v(x_i)$ to the decisionmaker; WLOG assume $v(x_2) > v(x_1)$. Guilt is given by $G(x_1 - \rho_2) = G(x_2 - \rho_1) = G$. Additionally, assume that t is distributed according to ϕ_t, independently from ρ. ϕ_t has full support on $\mathbb{R}^+$ and is continuous, as required by assumption 1. A fraction $p_1 \in (0,1)$ of the population has $\rho = \rho_1$.

For concreteness, consider the following simple example. Imagine that the decisionmaker must choose between two allocations of wealth for himself and a partner. x_2 corresponds to $(3,0)$; that is, 3 units of utility for the decisionmaker and 1 for the partner. x_1 corresponds to $(1,1)$. Some subset of the population is utilitarian and believes x_2 is the fairer allocation. Another subset is egalitarian and believes that x_1 is fairer. Individuals who don't care too much about following their own norm would prefer to choose the personally advantageous allocation x_2. Guilt free (low t) approval seekers hope that ρ_2 is the prevailing norm so that they can gain approval without sacrificing material interests; guilt free respect seekers want to choose x_2 and convince their peers that they do so out of a true commitment to efficiency. High-t types, of both kinds, care enough about their norm to ignore these factors.

The examples described in previous sections lead to less stylized interpretations: x_1 could correspond to being a Democrat and voting for strongly redistributive policies. x_2 could correspond to being a Republican and voting for less redistribution for the sake of higher overall wealth. Or the choice could be between declaring war or risking national security, or between generous safety nets or a laissez-faire approach, or between equal rights for a minority group or discrimination.[14] Outside of politics, this setting could represent vegetarianism versus omnivorism, orthodoxy versus reform religious practices, sending girls to school versus keeping them home to help with housework, staying at home to raise children or taking an outside job, or dressing casually or formally for work.[15].

Proposition 1 describes equilibrium for approval seekers, who perform a straightforward utility maximization, and for respect seekers, which emerges in a more complicated signaling game. A signaling equilibrium consists of an action function $Q : [0,\infty] \times \{\rho_1, \rho_2\} \to \{x_1, x_2\}$, along with a perception function $P : \{x_1, x_2\} \to [0,\infty]$ with $P(x_i) = E[t|x_i]$. This follows the notion of a psychological equilibrium introduced by Geanakoplos, Pearce, and Stacchetti (1989). Equilibrium transfers must be optimal given P and inferences must be consistent with Q. Throughout this paper, I also restrict attention to equilibria satisfying the D1 criterion of Cho and Kreps (1987), which requires that inferences about types from disequilibrium actions must be reasonable in the sense that, roughly, all weight must be placed on the types

[14] In some cases, like the latter, strong opinions exist despite *no* utility consequences of either outcome for most people; in this case, the model still applies with $v(x_2) = v(x_1)$.

[15] Note that if v is heterogeneous, the model still applies unproblematically so long as v is observable. Someone arguing in favor of marriage equality will simply be interpreted differently if they are gay themselves.

On the other hand, if v is heterogeneous and independent from (t,ρ), so long as v_i falls in some bounded set we can simply divide each person's utility function by v_i in order to normalize consumption utility (after adding some positive constant if necessary to ensure that $\operatorname{supp} v_i \in \mathbb{R}^+$). This then alters the distribution of t to t_i/v_i and effectively renders each person differently responsive to social pressure. The limit-case results still hold. If v_i is correlated with types, using the same trick introduces correlation between normalized t and ρ, so we end up in the case from section 4 in which people are differentially responsive to s. Again, limit-case results still hold.

who would be tempted to deviate to that action for the widest range of mistaken beliefs.[16]

Proposition 1. *If $X = \{x_1, x_2\}$, $v(x_2) > v(x_1)$, and ρ is independent from t, the following hold:*

1. *For approval seekers, there is exactly one equilibrium outcome, in which sufficiently high-t types will adhere to their personal ideals and low types will choose whichever option yields a better combination of material and image utility. Choosing x_1, the more costly option, will lead to a higher social image* iff $p_1 > .5$.

2. *For respect seekers,*

   a) *There exists at least one pure strategy equilibrium, and all equilibria must take a form in which sufficiently high-t types will adhere to their personal ideals and low types will choose x_2. In these equilibria, choosing the costly action x_1 leads to a higher social image.*

   b) *This equilibrium is unique if 1) G is sufficiently large, 2) s is sufficiently small, 3) p_1 is sufficiently small, or 4) $\max \phi(t)$ is sufficiently small.*

The intuition for approval seekers is straightforward. Consumption and image utility are fixed for each option, so people who don't care very much about guilt choose the option with the higher sum of those factors, and people with high enough t stick to their beliefs. The intuition for respect seekers is subtler, since social image depends on aggregate behavior. First notice that imitation can only ever occur in one direction: if anyone with either ρ defects to the other option, everyone else with the same ρ and a lower t will have an even stronger reason to defect. But also, if someone with ρ_1 defects to x_2, then someone with the same level of t but with ρ_2 has an even stronger reason to choose x_2, because they feel less guilty doing so. Therefore, either low-t types with ρ_1 defect to x_2, and everyone with ρ_2 chooses x_2, or vice versa. But if the former is the case, then the average person choosing x_2 will have a *higher* t than the average person choosing x_1, so $m_2 > m_1$. In that case, anyone with ρ_2 no reason to choose x_1 because they will have higher consumption, guilt, *and* image utility by sticking with x_2. So, imitation can only occur in the direction stated in the proposition.

Qualitatively speaking, Proposition 1 says that for respect seekers, the cost of norm adherence is a key determining factor in aggregate behavior. Costly actions will dissuade those with low integrity, leading to a higher social image associated with that choice, and to an overall tendency to choose the cheaper option. This contrasts with the situation for approval seekers. Approval seekers care about the population distribution of personal norms. If most of the population has ρ_i, they are tempted to choose x_i in order to please their peers.

[16] Since $\operatorname{supp} \phi = \mathbb{R}^+$, there is never an off-equilibrium path choice since sufficiently high types will always choose in accordance with their ideal, so the D1 criterion does not refine the result. However, if t is assumed to have an upper bound, Proposition 1 still holds exactly as stated with only equilibria satisfying D1 considered (see Appendix B). Later results will also be substantively refined by the D1 criterion.

Costliness has no role in social image; it merely factors into individuals' decisions as they trade off cost, image, and guilt.[17]

For both approval seekers and respect seekers, since types with arbitrarily high integrity are assumed to exist in the population, no pooling equilibrium exists. (However, it can be shown that if t has finite support, pooling equilibria could exist while the exact statement of Proposition 1 still holds; see Appendix B.)

Costliness is not the *only* thing that determines the respect-seeking equilibrium, however: depending on the exact distribution of types and other parameter values, Proposition 1 does not rule out multiple equilibria.[18] But, Proposition 1 also describes some sufficient conditions for ensuring a unique equilibrium. These are quite conservative conditions, and the meaning of "sufficient" is of course jointly determined and contingent on all parameters and the distribution of t. But they are informative: If guilt is sufficiently powerful or social pressure sufficiently weak, the system approaches a non-signaling model in which only self-interest and morality determine decisions and hence a unique equilibrium exists. That is, the special case of zero social pressure is not a knife-edge case; the model behaves similarly in an (often large) region of parameters. Additionally, if p_1 is small enough or ϕ is "flat" enough, the inference function is essentially forced to behave convexly enough to ensure a unique equilibrium; see Appendix B for details.

Notice that this result implicitly states that when *norms* shift slightly in the population such that the majority belief changes, *behavior* of approval seekers can shift much more dramatically and suddenly. This may be apparent, for example, in the shifting tide of public opinion about marriage equality. Meta-surveys indicate that 2010 or 2011 was when a majority of Americans first supported marriage equality, but the shift has been slow and steady (Silver 2011). Support among senators, however, has changed much more dramatically, and more quickly than can be accounted for by turnover: only 15 senators openly supported marriage equality in 2011, and 51 did as of April 2013 (Matthews 2013). Since senators derive utility (re-election) exactly from pleasing the largest fraction of the population, approval seeking is a likely explanation for at least part of this phenomenon. Similar forces may be behind sudden changes in taboos, such as political correctness or corporal punishment. Behavior appears to be nearly unanimous, but beliefs are likely much more divided, and simply hidden due to social pressure.

More than the static equilibrium, however, we are interested in the dynamics of the model as s changes so that we might understand how social pressure influences norm adherence. I investigate these dynamics in the next subsection.

[17] Note that respect seekers, but not approval seekers, can display the counterintuitive behavior explored theoretically by Bénabou and Tirole (2006) and demonstrated by Ariely, Bracha, and Meier (2009) and Carpenter and Myers (2010): Extrinsic motivations may interact with image motivations in a negative way, because the signaling value of doing good is reduced when doing good is rewarded.

[18] Figure B.1 in Appendix B shows an example with three equilibria, but multiple equilibria are far from typical; the example in Figure 1.1 shows that it's quite possible for equilibrium to be unique at *all* levels of s.

Comparative statics in s

Proposition 2 summarizes the equilibrium dynamics for both approval seekers and respect seekers, as social pressure changes:

Proposition 2. *If $X = \{x_1, x_2\}$, $v(x_2) > v(x_1)$, and ρ is independent from t, the following hold:*

1. *For approval seekers, in the limit as social pressure increases, a consensus forms around x_1 (x_2) if $p_1 > .5$ ($p_1 < .5$), enforced by high social-image utility relative to x_2 (x_1).*
2. *For respect seekers, in the limit as social pressure increases, everyone strictly adheres to their personal ideal and $m_{rs}(x_1) > m_{rs}(x_2)$.*

Once again, the intuition for approval seekers is very straightforward: as social pressure increases, it's more likely that the option providing the best combination of consumption, guilt, and image utility is the one with the higher image, since that component is weighted by social pressure in the utility function. For respect seekers, the intuition for the limiting case comes from the fact that if there were any difference in image between the two options (i.e., if $m_1 > m_2$ strictly), high enough social pressure would lead to crowding on m_1. But we know imitation can't occur in this direction, just as in Proposition 1. The only possibility satisfying the equilibrium restrictions is that both options lead to the same image. Based on the structure of equilibria found in Proposition 1, the only way this is possible is if everyone sticks to their personal norms.

Qualitatively speaking, the result says that as social pressure increases, cost disparities between actions become irrelevant for respect seekers, and somewhat unintuitively, any action will lead to approximately the same image. On the other hand, approval seekers in the same scenario will become more and more conformist to the modal ideal, regardless of relative cost, as defectors become and more harshly shunned.

Figure 1.1 illustrate the results of Propositions 1 and 2, showing that social pressure can potentially push behavior of respect seekers and approval seekers in opposite directions and that different sets of equilibria are possible.

Let's return to wealth redistribution as an example of the contrast between respect seekers and approval seekers. Respect-seeking egalitarians believe that redistribution is fair despite selfishly wishing to vote Republican. Only the egalitarians who care enough about fairness will vote Democrat. Realizing this, observers admire the integrity of Democrats more than Republicans, even though most Republicans are equally committed to their utilitarian beliefs. If social pressure is very high, more egalitarians will vote Democrat due to the added benefit of a better social image. In the limit, all egalitarians vote Democrat, all utilitarians vote Republican, and remarkably, no one is suspected of hypocrisy. A Democratic campaign among respect seekers would do well to advertise the selfishness of Republican policies (making v salient) and to loudly accuse the wealthy of being hypocrites who vote by their pocketbooks (increasing s).

Figure 1.1: Approval seekers' versus respect seekers' choices

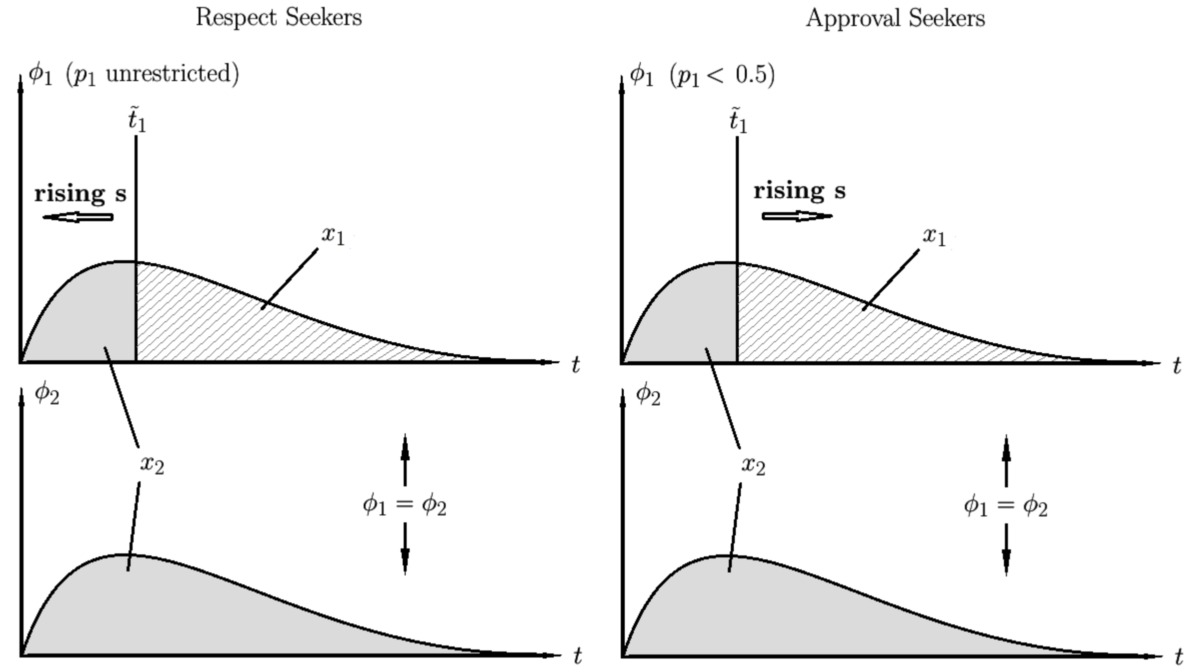


Illustration of choices made by the distributions of approval seekers and respect seekers. Types in the shaded regions of the distributions choose x_1, and types in the filled regions choose x_2. Increasing social pressure causes more approval seekers with ρ_1 to choose x_2 when $p_1 > 0.5$, but increasing social pressure (overall, although perhaps not for small changes) pushes respect seekers in the opposite direction.

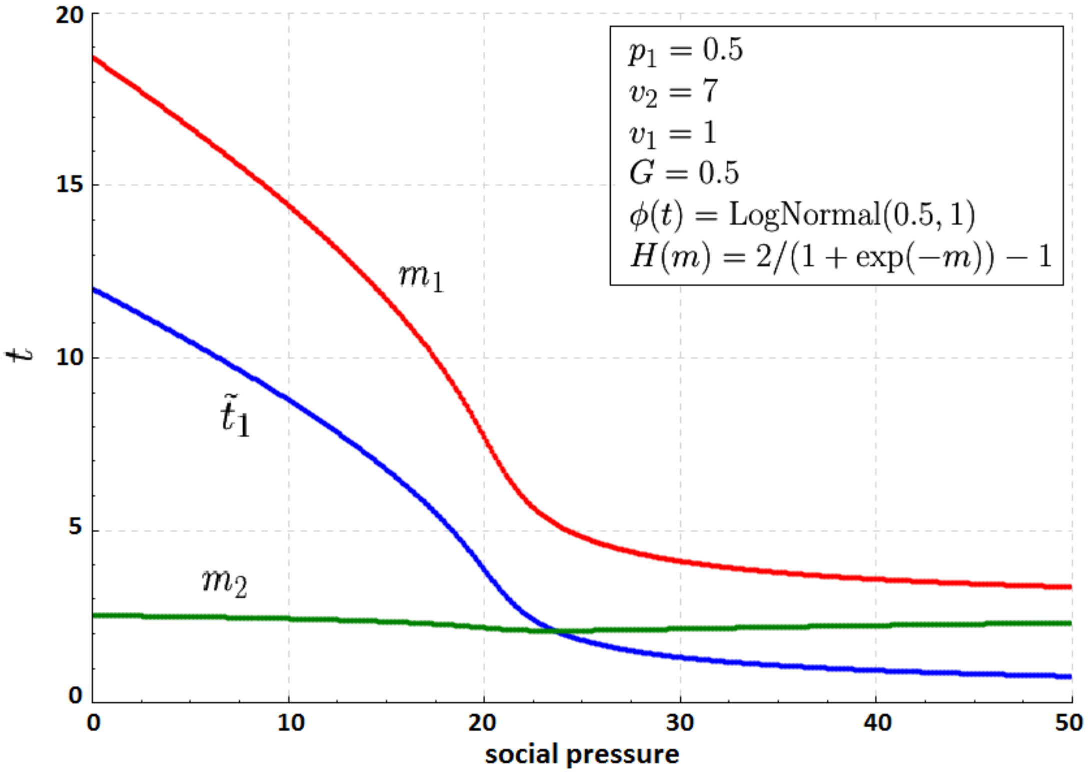
Approval seekers, on the other hand, don't try to discern each others' hypocrisy. If most people are born utilitarian, approval seekers will try to go along with the group, and more so the higher social pressure is. In the limit, only the most extreme egalitarians dare to follow their own moral compass in the face of extreme criticism. A Republican campaign among approval seekers would be advised to publicize the common wisdom of low redistribution (making p_1 salient), and loudly praise supporters for their loyalty (increasing s). This is the tactic employed recently by the Human Rights Campaign (HRC), as they've realized that their efforts to change minds about marriage equality directly haven't been very successful. As the tide of public opinion has shifted, they've switched to "trying to foster the sense that, whatever the justices decide, history has already ruled in favor of their cause" (Issenberg 2013).

The relative prestige of the two options as social pressure changes is also different for respect seekers and approval seekers. For respect seekers, in settings with extreme social pressure, the image associated with any choice is approximately the same in equilibrium. But when social pressure is lower, costly actions are uniquely admired as true signs of integrity.

Religion may be an example: religious fakery is judged harshly. (If you are skeptical, consider how often you have heard someone admit to knowingly disobeying God, rather than arguing for a religious interpretation that fits their actions?) As a result, members of reformed denominations are not assumed to be betraying their true orthodox beliefs simply because the rules are too onerous, and atheists are generally assumed to have abandoned religion for principled reasons. On the other hand, social pressure over dietary habits is not so strong that saying "I admire vegetarians, but I don't want to give up my steak." necessarily attracts horrified looks. In this domain, even though there are plenty of people who honestly think eating meat is the right and natural thing to do, vegetarians project an image of moral integrity more than omnivores.

Figure 1.2 shows an example of these dynamics, with parameters chosen so that a unique equilibrium exists at all levels of social pressure.

Figure 1.2: The effect of social pressure on respect seekers



An example of the model with the specified parameters. The solid curve shows the equilibrium cutoff value $\tilde{t}_1$ defining the minimum t for types with ρ_1 who choose x_1, as a function of social pressure s. As s rises, a smaller fraction of the population will act hypocritically. The bottom line shows m_2 and the top line m_1, both equilibrium values as a function of s. Note that as social pressure rises, the gap between m_1 and m_2 shrinks, so that in the limit either action will yield the same level of respect.

This characterizes the behavioral response to social pressure, but what are the welfare effects? This is explored in the next subsection.

Sacrificial equilibria

Without specifying material externalities, or how moral hypocrisy affects others, welfare isn't clearly defined in these models. For example, when discussing partisan politics, we would have to consider the imposition of one belief on the entire population in order to calculate welfare. But while it's easy to explain someone's defection to a profitable choice as a welfare-increasing choice (they are incurring guilt, but gaining consumption), it's much more surprising if an individual defects from his ideal and *sacrifices* material utility in order to do so. This individual would clearly prefer a lack of social pressure and is sacrificing utility in order to attain social image.[19] Define such a "sacrificial" equilibrium as follows:

Definition 1. *A population's equilibrium choices constitute a* sacrificial equilibrium *when some individuals with ρ_2 nonetheless choose the more costly option, x_1. (And an equilibrium is said to be more sacrificial when the fraction of the population who does this rises.)*

These sacrificial equilibria are very surprising from either the perspective of classical economics *or* from models of homogeneous norms. Nonetheless, Proposition 3 states that approval seekers are prone to sacrificial equilibria when the costly action is the majority norm, and moreso when social pressure rises. Respect seekers, on the other hand, are never able to sustain a sacrificial equilibrium.

Proposition 3. *If $X = \{x_1, x_2\}$, $v(x_2) > v(x_1)$, and ρ is independent from t, the following hold:*

1. *For approval seekers, if $p_1 > .5$, equilibrium is increasingly sacrificial when s is sufficiently high.*
2. *For respect seekers, equilibrium is never sacrificial.*

Intuitively, since approval seekers like to follow the crowd, sacrificial equilibria can be sustained in which individuals choose a costly option for no other reason than to win approval from the majority. Respect seekers' conformity is held in check by the rational expectations requirement for inferences, and no equilibrium can be sustained in which believers in ρ_2 nonetheless choose x_1 (see Proposition 1 for a refinement of this result, however). Approval seekers are thus more likely to be able to sustain an equilibria in which, for example, some parents invest more than they would like in their children's education, in order not to be looked down on by the other members of the PTA. But approval seekers are also more likely

[19] But such a sacrificial equilibrium may still be either good or bad for welfare: If individuals are being pressured into voting for higher funding for education than they believe is optimal, this is arguably good for long-term social welfare, but if individuals are being pressured into participating in a destructive religious war, this is clearly a bad outcome.

to become trapped in destructive equilibria. Some parents might spend more than they would like on sending their kids to private preschools, to the detriment of their ability to pay for higher education, in order to appear to be as devoted to their children as their neighbors.

The following three subsections examine the robustness of the results thus far, and in the process identify other several additional surprising phenomena.

Robustness

The initial analysis was simplified by the assumption that t and ρ are independent, which I relax here. Assume the same setting, but distinguish between the distributions of types t among those with ρ_1 and ρ_2: $t|\rho_1 \sim \phi_1$, $t|\rho_2 \sim \phi_2$. As before, fraction $p_1 \in (0,1)$ has ρ_1.

This leads to one norm being elite in the sense that it is associated with people of high integrity. While it's hard to measure whether beliefs and integrity are correlated in a particular domain, and hard to think of mechanisms that would produce such a correlation (despite the fact that many individuals might claim the moral high ground for their side!), the possibility must be addressed in order to rule out alternative sources for the behavior predicted above.

The qualitative behavior of approval seekers does not change (although the different distribution of course changes the exact fraction of the population that chooses each option), since part 1 of Proposition 1 does not depend on independence of ρ and t. Proposition 4 describes the behavior of respect seekers:

Proposition 4. *If $X = \{x_1, x_2\}$, $v(x_2) > v(x_1)$, and t and ρ are correlated, then for respect seekers:*

1. *There exists an equilibrium in which all respect seekers with ρ_1 (ρ_2) choose x_1 (x_2), in addition to types with sufficiently low t and ρ_2 (ρ_1).*

2. *As social pressure increases, in the limit, if $E[\phi_1] > E[\phi_2]$, everyone with ρ_1 will choose x_1 and those with ρ_2 and sufficiently low t will also choose x_1. The same applies if subscripts are reversed: the role of costliness is irrelevant in the limit.*

3. *If $E[t|\rho_1] > E[t|\rho_2]$, social pressure that is sufficiently high can sustain a sacrificial equilibrium.*

Part 1 is very similar to part 2 of Proposition 1, but allows for imitation to occur in either direction, since now low-t types may wish to choose *either* x_1 for the sake of a better social image, or x_2 for the sake of a better payoff. Higher social pressure drives them to the former, which in the limit serves to equalize the social image associated with different choices. Since in this limit imitation can only occur in one direction, this demonstrates that, contrary to section 2, taking costly actions isn't effective at signaling integrity if that choice is not usually strongly believed in. In the degenerate case, doing something that is *no* one's ideal, such as burning money, can't possibly signal integrity.

Part 3 establishes that even respect seekers are capable of sustaining sacrificial equilibria. These equilibria are substantially different from approval seekers' sacrificial equilibria (see section 1.4), however: approval seekers are sacrificial when the majority believe in a costly policy or action; respect seekers are sacrificial when particularly high integrity types disproportionately believe in a costly action. And, high enough social pressure can lead almost an entire population of approval seekers to choose the sacrificial action, but rational expectations limit sacrificial behavior among respect seekers to only a few low t types, no matter how high social pressure gets.

Proposition 1 is critical to keep in mind when predicting behavior based on the results of the previous subsections, since it is, after all, very difficult to know *a priori* whether a particular norm is correlated with integrity. If egalitarians and utilitarians have the same integrity, as in the previous subsection, then increasing social pressure has the unambiguous consequence of preventing egalitarians from defecting and voting Republican. But, if utilitarians tend to care more about fairness, then increasing social pressure may cause the opposite shift in voting patterns, as more egalitarians defect to the Republican party in an effort to be seen as being committed to growing the overall pie. A Democratic campaign in this world, if they followed the advice of the previous subsection, would only exacerbate the problem of defecting egalitarians.

From an economist's perspective, the complementary caution has to do with interpreting data. Empirical and experimental studies will be needed to discover how the two models in this paper play out in the real world. In these studies, even if the factors that are abstracted from in these models are carefully held constant, inferences can't be based on rigid assumptions about unobservable parameters. We will need to measure integrity and personal norms along with observed choices.

Unanimously immoral options

Another natural robustness check on the previous results is to allow other options than the ones that correspond to norms. At the very least, a binary choice often admits a third option: abstention. In other cases, opposing sides often have the opportunity to compromise on an option that neither believes in but both can accept. As it turns out, this doesn't substantially change the picture painted above, but does lead to new insights on the nature of compromise.

I analyze a ternary choice setting, but the intuition of the results would also apply to any richer choice set or a larger discrete set of norms. Consider a setting in which x_i is chosen from $\{x_1, x_2, x_3\}$. ρ_i is either $\rho_1 = x_1$ or $\rho_3 = x_3$, and x_2 is a middle ground option: $G(x_2 - x_1) = G(x_2 - x_3) = G_1$ and $G(x_1 - x_3) = G(x_3 - x_1) = G_2 > G_1$. This assumption simplifies the math but doesn't impact the results below. As before, fraction $p_1 \in (0, 1)$ have ρ_1 and the remainder have ρ_3. Without loss of generality, assume that $v_3 > v_1$. As in section 1, ρ and t are independent, which allows us to focus exclusively on the potential for compromise.

For concreteness, wealth redistribution is again a useful example. x_1 may be to vote Democrat, and x_3 to vote Republican, while x_2 is not to vote at all. From the perspective of policy makers, x_1 could be proposing strong redistribution, x_3 could be proposing a Republican alternative, and x_2 could be brokering an imperfect compromise. Note that neither failing to vote nor a hacked compromise or half-measure is an ideal outcome according to anyone, but that people with any norm are likely to be happier with that middle option than with letting the other side have its way.

Proposition 5 describes the equilibrium:

Proposition 5. *If $X = \{x_1, x_2, x_3\}$ with $G(x_2 - x_1) = G(x_2 - x_3) = G_1$ and $G(x_1 - x_3) = G(x_3 - x_1) = G_2 > G_1$, and ρ is uncorrelated with t, then the following hold:*

1. *For respect seekers, at least one equilibrium exists, in which high integrity types adhere to their norms. Either low types with either norm defect to the middle option, or low types with ρ_1 defect while ρ_3 is adhered to perfectly. In the latter case, low types with ρ_1 either all defect to x_3, or mid-level types defect to the middle and low types defect to the opposite ideal.*

2. *For approval seekers, a unique equilibrium exists in which high-t types adhere to their norm and lower types either all defect to one option, or mid-level types defect to x_2 and lower types defect to another option.*

The intuition behind these results is quite similar to that of Proposition 1, but the addition of a third option, when v_2 falls between v_1 and v_3, provides some midlevel types with a defection that is profitable without inducing too much extra guilt. Or, of v_2 is greater than either v_1 or v_3, low-t types with either norm would of course prefer to defect to x_2 than the other side.

Figure 1.3 illustrates two possibilities for respect seekers, one in which all low-t types defect to the profitable middle option, and one in which everyone with ρ_3 always chooses x_3. There is an additional possibility in which neither type chooses x_2. Equilibrium takes the same static form for approval seekers but the arrows may point in any direction.

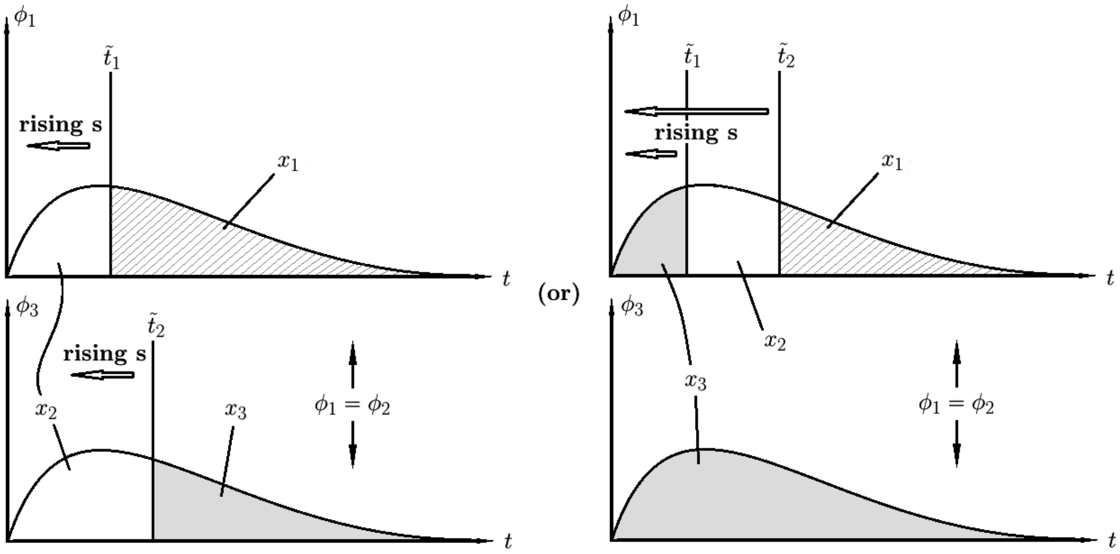
Proposition 6 provides the comparative static result analogous to section Proposition 2.

Proposition 6. *If $X = \{x_1, x_2, x_3\}$ with $G(x_2 - x_1) = G(x_2 - x_3) = G_1$ and $G(x_1 - x_3) = G(x_3 - x_1) = G_2 > G_1$, and ρ is uncorrelated with t, then the following hold:*

1. *As social pressure rises, in the limit, all respect seekers adhere closer to their own norms, with only very low integrity types with one norm defecting to the other norm and no one choosing the compromise option.*

2. *As social pressure rises, in the limit, either all approval seekers pool on the majority's ideal action x_1 or x_3, or they all compromise by choosing x_2.*

Figure 1.3: Respect seeker compromise



Two (non-comprehensive) possibilities for respect seekers with an opportunity for compromise. The left side possibility occurs when types with either ρ choose x_2 but x_2 disappears as social pressure increases. The right side shows another possibility when ρ_3 is perfectly adhered to.

As before, approval seekers are more prone to conformity than respect seekers. Interestingly, the two groups have different motives for choosing the middle option. Respect seekers will *only* choose to compromise if it offers a large enough monetary reward, and that option is necessarily abandoned in the limit as s approaches infinity, because since no high types will ever choose it, it inevitably has a less favorable image than one of the other options. Approval seekers, on the other hand, may choose the middle option if it's better to partly appease everyone than to perfectly please one group and displease the other, even if compromising is costly.

These results are clearly relevant for understanding group decision making. Going back to the voting example, while it's almost universally agreed that you *should* vote, regardless of your political beliefs, you have the option to vote for either party or to abstain in any particular election. A respect seeker might be tempted to abstain to avoid the hassle, but this action would never be socially rewarded. On the other hand, an egalitarian approval seeker might be tempted to abstain if he doesn't want to be shunned by his Republican friends, or may even vote Republican. If he did abstain, his friends of all beliefs would react by saying "Well, he didn't vote, but at least he didn't vote for the wrong side." As social pressure rises, more approval seekers might avoid confrontation this way, and voter turnout would plummet. This has also been strategically used by the HRC as they promote the idea that history has already decided in favor of marriage equality: "It just deflates them. People who may disagree with [gay marriage] but believe it may happen anyway are hard people to mobilize" (Issenberg 2013).

Simultaneous respect and approval motives

A final natural robustness check on the results of sections 1.4 through 1.4 is to allow respect-seeking and approval-seeking motivations to interact. This allowance turns out not to substantially change the earlier results, but does lead to new insights.

I model simultaneous respect- and approval-seeking motivations in the natural manner:

$$U(x_i) = v(x_i) - tG(x_i - \rho_i) + s_{as}H_{as}(m_{as}(x_i)) + s_{rs}H_{rs}(m_{rs}m(x_i))$$

First note that $H_{as}(x_i)$ is a fixed quantity, so that at a particular level of social pressure, low-t individuals will choose the option with higher $v(x_i) + s_{as}H_{as}(m_{as}(x_i))$, analogously to Proposition 1. But as s_{as} increases from 0, this direction of imitation might change.

At higher levels of s_{as}, however, the utility from approval dwarfs consumption utility, and imitation must occur in the direction of higher approval. At the same time, increasing s_{rs} reduces imitation, as people try to convincingly signal their integrity. The net effect may or may look like *either* of the equilibria described in Proposition 1:

Proposition 7. *If individuals are motivated by both approval and respect, $X = \{x_1, x_2\}$, and ρ and t are uncorrelated, the following hold:*

1. *Low-t types choose* $\arg\max v(x_i)$ *for sufficiently small s_{as} but choose* $\arg\max H_{as}(m_{as}(x_i))$ *for larger s.*

2. *At a fixed level of s_{as}, increasing s_{rs} in the limit leads to perfect adherence to personal norms.*

3. *At a fixed level of s_{rs}, increasing s_{as} in the limit leads to perfect conformity with majority opinion.*

4. *As s_{as} and s_{rs} simultaneously increase, in the limit, equilibrium can take* either *of the two forms above.*

One key difference arises from section 1.4: accepting social disapproval can *itself* be a signal of integrity.

This interaction between image motivations, perhaps pushing in opposite directions, can explain the seemingly arbitrary costly signals that individuals take to declare their identity convincingly. For example, teenagers might wear goth clothing in order to be shunned by the majority, thereby convincingly signaling to their friends that they are devoted to their social group. This rhetoric is also frequently used to promote evangelism. In the Sermon on the Mount, Jesus states states (emphasis mine) "Blessed are those who are *persecuted because of righteousness*, for theirs is the kingdom of heaven. Blessed are you when people insult you, persecute you and falsely say all kinds of evil against you because of me." (Matthew 5:11). In a Mormon parable, the righteous man is similarly identified as the one who withstands the strongest pressure: "This is the only righteous man in the country; and there are seven of the biggest devils trying to turn him out of his path, and they all cannot do it" (Young 1858). This rhetoric also permeates popular culture in D.C. Talk's double-platinum Grammy-winning Christian rock album's hit single *Jesus Freak*, which (unintentionally) ironically implies that believing its message will lead to social ostracism: "I don't really care if they label me a Jesus freak - there ain't no disguising the truth."

Combining both types of social image into a single model clarifies the prescriptive discussion of the previous subsections. "Social pressure" is a general concept combining many things into a single parameter, and these components don't necessarily increase or decrease unilaterally. It may be possible to target approval-based social pressure separately from respect-based social pressure. See section 1.5 for further discussion on this point.

This concludes the theoretical results. The next section briefly discusses how these results apply to various real world domains, pointing the direction to empirical work on heterogeneous norms.

1.5 Discussion

The models in this paper are applicable to a wide range of domains. Most decisions governed by moral guidelines, customs and traditions, or fads and fashions (a very wide range!) fall into an ambiguous domain in which people disagree about appropriate actions. In addition to the various examples referred to throughout the discussion, these models can be used to understand behavior such as middle school fads, child-rearing practices and taboos, religious practice, etiquette, local customs, and so on ad infinitum. I highlight a few possibilities of particular interest to economists here.

Political Economy: In addition to the obvious applications to partisan politics and persuasion, which have been alluded to throughout the paper, Propositions 5 and 6 may specifically be relevant to political polarization. Polarization has been shown to be increasing in the United States, at least among political elites (McCarty et al. 2006), over at least the last half century. While attention has focused on potential demographic reasons for increasing polarization (for a review, see Layman, Carsey, and Horowitz (2006)), this model points to social pressure as another possible explanation. If legislators are respect seekers and are appealing to the same constituency (such as federal-level or local-level politicians), increasing polarization results from increasing social pressure. On the other hand, if legislators are approval seekers, increasing polarization most likely results from *decreasing* social pressure. Or, if increasing polarization is primarily occurring among politicians with differing constituencies (such as state representatives in Congress), this could indicate approval-seeking behavior in which higher social pressure leads politicians to act more beholden to local interests[20].

[20] Future work will explore how respect- and approval-seeking motivations function when people have

In any of these situations, the source of changing social pressure could come from any number of sources, such as media penetration, political literacy, changing formal or informal institutions that incentivize party loyalty, information technology, etc. But in any case, the theory points in this direction as a possibility to be checked empirically, or at minimum forms a theoretical basis for existing theories of polarization that can be thought of as based on social pressure. In fact, it could also explain a situation in which polarization increases among political elites despite fairly constant polarization within the electorate, which some authors claim is the case in the United States (Fiorina and Abrams 2008) although the evidence is somewhat mixed (Abramowitz and Saunders 2008). This poses a puzzle for many potential explanations for polarization (again, see Layman, Carsey, and Horowitz (2006) for a review) but can be understood within these models of social pressure, perhaps with a similar mechanism as that proposed to explain the dramatic shift in support for marriage equality discussed in section 1.4.

Development: A particularly important domain that can potentially put the models in this paper to good use is development economics. Many development initiatives are beholden to or stalled by local norms which either interfere with incentives or are directly the cause of behavior that an initiative aims to change. For example, interventions designed to reduce HIV transmission must deal with different norms for safe sex practices (Macintyre, Brown, and Sosler 2001), pregnancy prevention relies on strong norms allowing women to demand the use of contraception (Caldwell and Caldwell 1987), sanitation systems that dramatically reduce child mortality depend on usage norms (Kremer, Miguel, and Thornton 2009), women can only advance in society if their parents are willing to send them to school (Fuller, Singer, and Keiley 1995), and wealth accumulation and capital investment are only possible if the property rights of the culture allow it (Svensson 1998). The list of such examples is endless.[21]

While many individuals may agree with the alternative norms that are promoted by a development initiative, they are still beholden to the social-image motivations that enforce the local norm. Understanding the mechanics of these motivations is critically important to designing effective, persuasive interventions. The above results show that if individuals are respect seekers and the desired action is costly, increasing visibility may encourage those who agree to comply. This is the approach taken, for example, by Cameron, Gertler, and Shah (2013), which attempts to reduce open defecation by making usage of the sanitation system visible. On the other hand, if individuals are approval seekers and the desirable norm has yet to take significant hold, or if individuals are respect seekers and the *un*desired action is costly, peer visibility and social pressure should be minimized. An initiative relying on social pressure to encourage good behavior may *backfire*.

The case of Kremer, Miguel, and Thornton (2009) contains a convenient illustration of the relative effectiveness of an education initiative when there is more or less heterogeneity in norms. They analyze the effect of NGO scholarships offered to girls in two districts in

different peer groups or face multiple peer groups at different times.

[21] Aldashev et al. (2011) also discusses some of these examples through the lens of wanting to change norms, but models norm adjustment through direct institutional changes.

Kenya, granted based on test scores, on education attainment. In the Busia district, 90% of teachers claimed that parents of students were positive towards the NGO, while in the Teso district this number was only 58%. There was therefore more pressure in the Teso district to drop out of the program or refuse the scholarship (as many schools, and one scholarship winner, did). In the end, not only did this high attrition rate jeopardize many students' chances at a scholarship, the impact on education attainment from the scholarship opportunity was much smaller in the Teso district. Part of this gap might be explained by respect-seeking students who were refusing a valuable program in order to signal dedication to their belief that outsiders are not to be trusted, or by approval-seeking students who caved to community pressure to resist outsiders.

Speculating in another domain, imagine that a development initiative were designed to encourage better sanitation habits, using social recognition as an incentive, as has been proven effective in other contexts such as community health and contraception (Ashraf, Bandiera, and Jack 2012; Ashraf, Bandiera, and Lee 2013). *Even if* locals express support for the program, they may not admit wanting to keep their beliefs secret, perhaps to avoid embarrassment, perhaps to stay in good graces with the organization that is bringing money to the community, or perhaps simply out of apathy. Social pressure in this case may *reduce* sanitation. Conversely, reducing social pressure or undermining the prestige of the desired action might discourage good sanitation habits among respect seekers.[22] Understanding local norms is always an important part of program design, but these models of heterogeneous norms clarify important specific pitfalls to ensure against.

If a beneficial norm has yet to take hold to *any* significant degree, despite seemingly significant costs to the individual, the development economist might also look to the results on sacrificial equilibria in section 3. Social pressure might be maintaining a destructive norm of bad sanitation, or violent civil conflict, or expensive religious sacrifice, or refusing contraception. And if social pressure is too high, it might be difficult to break the cycle. Take, for example, the norm of having expensive funerals. In several African countries, this has clearly reached the status of a destructive, welfare reducing norm; funeral costs are a major cause of families falling into poverty (Case et al. 2008; Krishna et al. 2004). In this case, reducing social pressure is unambiguously the way to improve the situation.[23]

Specific image targeting: It's unlikely that only respect-seeking or approval-seeking motivations are in effect in any particular setting, although the relative strengths surely vary. Political activists then may wish to refer to the model of section 1.4 by *separately* targeting respect or approval. Take, for example, the efforts to enroll young people in health insurance under the new Affordable Care Act. The partisan reputation of the ACA may discourage

[22] This is similar to the crowding out effect of extrinsic incentives on intrinsic altruism identified in Ariely, Bracha, and Meier (2009) and Carpenter (2005).

[23] This view is further strengthened by an extension of the model discussed in Appendix A.1, which allows norms to be chosen endogenously. These results show that a destructive norm is only sustainable when social pressure is high, and that another welfare-improving equilibrium is always possible. Rather than triggering a switch to this other equilibrium (by trying to change minds directly, for example) a development economist may wish to reduce social pressure to the point that the destructive norm is not sustainable at all.

young Republicans from enrolling, on principle, despite the fact that taking advantage of the new subsidies is clearly the profitable option for most of them. Democrats, on the other hand, may see political reputation and following their own beliefs as additional advantages to enrolling, on top of the actual subsidies. They are therefore in no danger of not enrolling. Since the concrete incentives already push young people in the direction of enrolling, Propositions 2 and 7 show that the respect-based social image of the choice should be as downplayed as possible, by downplaying its partisan nature or keeping enrollment decisions as low pressure and low visibility as possible. On the other hand, young people also have an interest in other aspects of their reputation, such as the approval-seeking motivation to act like a responsible adult and minimize the risk that their parents will have to step in to help with emergency medical care. Social pressure on that dimension (s_{as}) should be increased to promote enrollment.

Marketing: Applications to marketing are elaborated on in Appendix 10, but I'll briefly mention them here as well. One can take a different view of the results in sections 1.4 and 2 as indicating to companies or groups how they should price products that are used to express identity or beliefs or to support moral beliefs, such as group membership fees, "Save the rainforest" products, brand name clothing, etc. Respect seekers seek opportunities to prove their commitment to their beliefs and are thus willing to pay for those products as a credible signal. Approval seekers, on the other hand, wish to follow their beliefs without being shunned and so will pay less for these products the higher is social pressure. See Appendix 10 for further details.

1.6 Conclusion

In this paper, I developed two alternative models of social-image motivations that influence moral choices. When people disagree about the appropriate action, two natural possibilities arise for the meaning of social image: people may wish to signal their adherence to their personal ideal, or they may wish for others to admire their choices. These alternatives lead to substantially different predictions. Populations who wish to signal integrity are prone to multiplicity of equilibria and exploiting costly signals as credible signals. The possibility for conformity and compromise is limited. Populations who care about pleasing others, on the other hand, are highly prone to conformity, and can agree on imperfect compromises. Both groups are potentially able to sustain equilibria in which individuals sacrifice both their moral ideals and their consumption utility for the sake of social image, although the extent is limited for respect seekers. Both groups pool on similar ideals when moral beliefs are endogenous, although the mechanisms behind these processes are very different in the two models. And when the two types of image motivations occur simultaneously in the same population, they can interact in ways that preserve the qualitative nature of the equilibrium but potentially lead to social disapproval itself acting as a signal of integrity.

These models provide a platform for future work on social image in the presence of disagreement over norms in general settings. Natural extensions include models of hypocrisy across multiple social groups, the evolution and transmission of norms, the formation of norms, and social image when information or inferences about others' norms are biased. Work in these directions is ongoing, and certainly the possibilities do not end there.

This theoretical framework also provides a foundation for rigorously understanding social-image motivations in many real-world contexts that have previously been out of reach of the social preferences literature, such as partisan politics, contentious moral choices, customs and taboos. Prescriptively speaking, it provides a theoretical basis for wisely designing institutions and/or interventions that anticipate the effect on the social pressure dynamic and result in the desired behavioral response. It immediately reveals the risks in ignoring the distinction between types of social image, and points to better alternatives when an initial approach fails due to targeting the wrong motivation.

Rigorously studying how these models play out in practice will also require empirically determining the contexts in which each model is applicable. Surely, people are motivated both by approval and by respect in different relative amounts in different scenarios, as touched on in Section 1.4. A likely possibility is that approval seeking is a more salient motivation when externalities of choices are large. On the other hand, Thomas Jefferson seems to prescribe approval-seeking and respect-seeking motivations to different classes of decisions when he said "In matters of taste, swim with the current; in matters of principle, stand like a rock." Characterizing the domains in which each model is applicable is an open empirical question and left for future work, but these models form an analytical foundation for beginning this research agenda.

Chapter 2

Rethinking Reciprocity

2.1 Introduction

Reciprocal behavior is of increasing interest in many areas of economics. Labor economists argue that firms pay above-market wages in order to induce reciprocal behavior among their workers (Akerlof 1982; Bewley 2009; Fehr, Goette, and Zehnder 2009), which in turn might explain involuntary unemployment (Akerlof and Yellen 1988, 1990) and inflation (Okun 1981). Conversely, workers who feel mistreated by their employer exert negative reciprocity and lower effort or produce faulty products (Krueger and Mas 2004; Kube, Maréchal, and Puppe 2013). Public economists show that inducing reciprocity with small gifts helps to raise funds for public projects, such as national parks (Alpizar, Carlsson, and Johansson-Stenman 2008) or for charities (Falk 2007), and argue that reciprocity is important for tax compliance (Feld and Frey 2007). Political economists provide evidence that politicians target pre-election transfers towards reciprocal individuals to increase their vote share (Finan and Schechter 2012). And health economists show that pharmaceutical representatives exploit the reciprocity of doctors with small gifts (Brennan et al. 2006). This literature in economics builds on earlier work in psychology, including the seminal work by Cialdini on the use of small gifts in marketing and politics with the aim to trigger reciprocity (see, e.g., Cialdini (1993)).

The recognition of reciprocity as a determinant of economic behavior extends the relevance of social preferences beyond their explanatory power for charitable giving and other 'one-sided' acts of sharing and giving to a broader class of market interactions. It suggests that market features such as dynamic pricing strategies (e.g., low introductory prices), consumer loyalty, or employee turnover are affected by social preferences. Profit-maximizing firms need to anticipate and respond to the non-standard behavior of their consumers.

Despite the empirical importance of reciprocal behavior, economists still struggle to converge on the best model of reciprocity. Existing theories rely on the response to good intentions or the kindness of others as the trigger for reciprocal behavior (e.g., Rabin (1993) or Levine (1998)). Under these theories, people reciprocate because another person's kind act or benevolent nature increase the intrinsic utility of acting kindly toward this person. Thus, such preferences are 'internal' in that they arise from an individual's preference to act in a way that rewards good behavior by others.

In this article, we propose that the analysis of reciprocity needs to be re-thought. Our argument is motivated by recent evidence on the motives underlying a simpler, 'one-sided' type of pro-social behavior. Recent literature shows that many seemingly altruistic people are reluctant to share and avoid the opportunity to share if they can (Broberg, Ellingsen, and Johannesson 2007; Dana, Cain, and Dawes 2006; DellaVigna, List, and Malmendier 2012; Lazear, Malmendier, and Weber 2012). Such evidence has led to a revision of the conventional theoretical motives thought to underlie generous or charitable behavior. In addition to an 'internal' preference for equality or social welfare, 'external' factors such as social image, self image, or the response to social pressure or norms appear to be important determinants of seemingly altruistic behavior (Akerlof and Kranton 2000; Battigalli and Dufwenberg 2007; Bénabou and Tirole 2006; Harbaugh 1998b).[1] These advances in our understanding of 'one-sided' giving decisions have thus far had little spillover to our understanding of 'two-sided' pro-social behavior, where the decision to give is a function of the earlier generous behavior of the other person. Our review discusses the existing literature on reciprocity, including the leading theoretical models, and points to the common failure among these models to account for the possibility of external motives.

We then describe a novel experiment that explores the importance of internal versus external determinants of reciprocal behavior. We employ a double-dictator game, which allows us to compare giving in a positive- and a negative-reciprocity setting with giving in a neutral setting (the standard dictator game). We modify these situations by giving dictators the option to avoid the sharing decision. The 3x2 design (neutral/positive/negative x no-sorting/sorting) allows us disentangle internal and external factors determining giving in one-sided versus two-sided (reciprocity) giving environments.

The findings cast a new light on the motives underlying reciprocity. First, when the other party induces positive reciprocity in the double-dictator game (with prior sharing), giving increases relative to the baseline (standard) dictator game, consistent with prior findings. The increase persists after we introduce sorting, relative to a baseline dictator game with sorting. However, comparing double-dictator games with and without sorting, we also continue to find a significant decrease in giving under sorting, very similar in size to the sorting effect in the single-dictator game. Thus, positive reciprocity does not eliminate reluctance to share and we can conclude that external factors are important determinants of sharing in reciprocity environments as well. We also find that negative reciprocity, induced by receiving $0 out of $2 in the first stage of the double-dictator game, virtually eliminates giving in the setting with sorting. It even induces some people to sort in and then share zero. The reduced-form results suggest that failure to account for external motives leads to a significant

[1] We use the terms 'internal' and 'external' to refer to the distinct sources of motivation in order to reflect the primary source of the motivation as one coming from inside the individual (preferences) or from outside (social considerations). We chose this simple terminology over alternative terms reflecting similar distinctions: e.g., 'intrinsic' vs. 'extrinsic,' 'presentational' vs. 'non-presentational.'

overestimation of internal motives such as fairness and altruism in positive-reciprocity settings, and to a significant underestimation of spite (negative altruism) in negative-reciprocity settings.

We confirm these conclusions by employing a simple structural model that distinguishes between internal and external motives. Even after the inducement of positive reciprocity, we estimate a significant influence of external factors, similar in size to the strength of external factors in one-sided (standard) dictator games.

We also compare our estimates to those based on a naive structural model that does not account for external motives. This omission could bias our understanding of social preferences in two ways. Consider the following example. Suppose that, in a simple dictator game, 65 percent of dictators share. Now allow recipients to send a small gift to the dictators, prior to the dictator game, and suppose that the percentage of dictators who share increases to 80 percent. A simple model of internally motivated reciprocity would attribute the decision of all 80 percent of sharers to welfare-enhancing motivations. This interpretation might be wrong on two counts. First, some of the 65 percent who share in both the simple one-sided DG and in the two-sided reciprocity setting might be motivated by extrinsic factors, such as social pressure or social image concerns. These people would prefer to avoid the sharing situation, as found in Lazear, Malmendier, and Weber (2012). Second, the additional 15% who share only in the two-sided game may be changing their behavior due to changes in external motivations (such as an increase in social pressure), rather than changes in internal motivations (such as an increase in altruism induced by the recipient's prior act of kindness).

Consistent with the reduced-form results, failure to account for external motives leads to a sizable overestimation of internal motives, both under positive and under negative reciprocity. At the same time, however, the estimation results also show that the amount of "additional" intrinsic motivation, in the sense of "reciprocity-induced" altruism or similar intrinsic motives under positive reciprocity is estimated approximately correctly in the naive model. In other words, it is correct to attribute the additional giving in a positive-reciprocity environment, compared to a neutral setting, to internal preferences, and even a naive model delivers the correct estimate. This robustness reflects the fact that external motives are remarkably stable across both environments. This conclusion is somewhat less true for the negative-reciprocity environment: The decrease in internal willingness to share after an unkind treatment, relative to a neutral treatment, is slightly underestimated when neglecting external factors and is strongly overestimated when allowing only for a reciprocity-insensitive external motive. In other words, failure to account for external motives and their context-dependence may be particularly detrimental when estimating the motives for and extent of negative-reciprocal behavior. Instead, it seems that individuals care less about external motivations, such as social image, regarding their actions towards unkind strangers than towards neutral strangers; this could also reflect an interaction between internal and external motivations, in which individuals care more about their social image regarding choices over which they have strong internal social preferences.

In summary, the experimental data confirms that external motives are a significant determinant of pro-social behavior in reciprocity environments, similar in strength to their role in 'one-sided' environments without reciprocity. This insight is relevant to the economic analysis of reciprocal behavior in markets, as it suggests that market participants aiming to 'trigger' positive reciprocity need to account for observability and avoidance options. For example, a charity might benefit from giving a small gift to the potential donor, as shown by Falk (2007); but the gift will be most effective if the recipient cannot avoid the subsequent request for a donation.

A second main insight from the experimental data relates to the value of combining reduced-form and structural analyses in disentangling determinants of pro-social behavior when the outcome (sharing) could be explained by rather different underlying psychological motives. The structural estimation allows us to decompose the share of giving that is due to internal versus external determinants, both in one-sided and in two-sided giving contexts.[2]

We organize the remainder of the paper as follows. In Section 2, we review the literature on social preferences, both in 'one-sided' and in 'two-sided' giving contexts. We suggest that recent insights on external determinants of giving in one-sided situations need to be applied to the study of reciprocity and other social preferences in two-sided giving contexts. We then describe the results of a novel experiment that allows us to better understand the motives underlying reciprocal behavior (Section 3). We then discuss a simple model integrating internal and external factors in the analysis of reciprocity and structurally estimate their relative importance (Section 4). Section 5 concludes.

2.2 Understanding Social Preferences

The relevance of social preferences has been a source of debate in the profession for several decades. How do we explain sharing behavior in standard economic settings when it is inconsistent with a classical model of self-interested preferences? Within this broad area of inquiry, most attention has been paid to what one might call 'one-sided' acts of kindness or generosity, where one party shares with another party without consideration of the other party's sharing behavior. This article focuses on 'two-sided' pro-social behavior, where sharing is a response to the other person's pro-social behavior. However, since our motivation for rethinking the reciprocity literature stems from advances in the literature on one-sided sharing decisions, we first review the two waves of literature focusing on one-sided sharing situations.

[2] It is also worth noting that internal motives for reciprocity need re-thinking beyond the confounds with external factors. Even motives that are cleanly identified as internal—i.e., the giver seeks an (anonymous, invisible) giving opportunity rather than avoids it—are not necessarily 'altruism' or 'fairness.' For example, rather than caring about the intentions or kindness of the other person (Levine 1998; Rabin 1993), a giver might feel an internal *obligation* to reciprocate (Cialdini 1993; Sugden 1984). For example, Malmendier and Schmidt (2012) observe that reciprocal behavior is triggered even if the preceding "gift" was solely motivated by the selfish goal of inducing a reciprocal response, which is inconsistent with an intention-based or type-based view of reciprocity, but consistent with an obligation- or norm-based understanding. See also Postlewaite (2011), Krupka, Leider, and Jiang (2012), and Bicchieri (2006).

One-Sided Giving

A large volume of laboratory and field experiments documents voluntary sharing behavior, and studies the individual characteristics and contextual factors that influence such sharing. These studies show that generosity varies based on determinants such as the gender of decision makers or other personal characteristics, the framing of the decision, the source of the surplus, as well as the social context in which an altruistic choice is embedded (Andreoni and Miller 2002; Andreoni and Vesterlund 2001; Brock, Lange, and Ozbay 2013; Cherry, Frykblom, and Shogren 2002; DellaVigna et al. 2013a; Fong and Luttmer 2009; Henrich et al. 2010; Hoffman, McCabe, and Smith 1996). While sharing can vary substantially based on these factors, the existing research provides robust evidence of a significant willingness to share with others that is inconsistent with purely self-interested preferences.

In response to these findings, economists have developed simple models of preferences that can account for such behavior. Leading models include altruism in the form of utilitarianism (Andreoni and Miller 2002) or in the form of maximin preferences (Charness and Rabin 2002), as well as different specifications of inequality aversion (Bolton and Ockenfels 2000; Fehr and Schmidt 1999). These models assume people derive utility from implementing equal or fair outcomes.

More recent research has demonstrated that the motives underlying a sharing decision are more complex. Several studies document a puzzling phenomenon: people share voluntarily when asked to decide between sharing and not sharing, but most of them prefer to avoid the decision altogether and keep their endowment (Broberg, Ellingsen, and Johannesson 2007; Dana, Cain, and Dawes 2006). The novel design feature in these studies is to give the decision maker the option of avoiding the decision to give altogether. For example, in the laboratory study of Lazear, Malmendier, and Weber (2012), subjects are allowed to “opt out” of playing the dictator game. Or, in the field study of DellaVigna, List, and Malmendier (2012), the exact timing of a door-to-door fund-raiser is pre-announced so that people can choose not to open the door of their home. In both cases, people can avoid the socially unpleasant act of having to say “no” when presented with a sharing opportunity.

Along similar lines, contextual features that allow decision makers to obscure the relation between their behavior and unfair outcomes also decrease the willingness to act pro-socially. For example, Dana, Weber, and Kuang (2007) show that giving decreases when the recipient does not know for sure how a choice was made, or when the dictator does not know how his choice will affect the recipient.[3] In fact, dictators prefer not to find out about the effect of their choice. Andreoni and Bernheim (2009) manipulate both anonymity and the probability that a computer secretly overrides the dictator’s choice with a known default, showing both that anonymity reduces sharing and that dictators will take the opportunity to hide their

[3] A more extreme way to remove transparency is having the recipient unaware of a game entirely. Dana, Cain, and Dawes (2006) show that dictators are less generous in this case. Koch and Normann (2008) and Johannesson (2000), instead, find similar generosity in a standard dictator game and a variant with uninformed recipients. It remains to be shown whether differences in experimental designs (e.g., double-blind anonymity, the precise kinds of recipients) account for the difference in findings.

selfish actions when possible. Grossman (2012) uses a similar probabilistic dictator game to demonstrate the importance of social signaling relative to self-signaling. Moreover, several of the experiments mentioned above have shown that people prefer to avoid letting the recipient know about the game, even at a cost (Broberg, Ellingsen, and Johannesson 2007; Dana, Cain, and Dawes 2006; Lazear, Malmendier, and Weber 2012).

The importance of avoidance options can also be seen in earlier experiments that manipulate anonymity and observability. Hoffman et al. (1994) and Hoffman, McCabe, and Smith (1996) show that dictator-game giving drops when using a double-blind framework. Conversely, Bohnet and Frey (1999b) and Bohnet and Frey (1999a) show that giving increases when dictator and recipient face each other. Relatedly, Franzen and Pointner (2012) show that dictator-game sharing is reduced when choices are concealed from the experimenter using a randomized response technique.

Anonymity and observability appear to play a similar role outside of the lab. For example, Mas and Moretti (2009) show that store employees work harder when observed by other employees who work harder. Gerber, Green, and Larimer (2008) find that revealing voting behavior to neighbors increases voter turnout. And DellaVigna, List, and Malmendier (2012) show that face-to-face interaction with a solicitor creates social pressure to donate to charity. In other words, generosity decreases significantly when individuals can avoid "saying no" directly to or being observed by someone.

These studies had a significant impact on the way economists think about social preferences. They revealed that people who share do not necessarily derive utility from sharing in the manner assumed by early social preference models. Instead, the motive driving much voluntary sharing appears to be a concern to not *seem* selfish or greedy. Rather than valuing the opportunity to act generously or altruistically, people are often motivated by an aversion to the guilt or shame that comes from disappointing the expectations of others, violating norms of sharing, or giving others the impression that one is selfish (Andreoni and Bernheim 2009; Battigalli and Dufwenberg 2007; Bénabou and Tirole 2006).

To sum up, the above stream of social-preference research can be organized as follows: first, empirical evidence that people act in a manner consistent with a preference for altruism or fairness and theoretical models of such a motivation; second, evidence that such behavior is susceptible to variations in observability and manipulations that allow the maintenance of a positive social image while acting selfishly, along with a new class of social-preference theories that account for such extrinsic concerns.

Two-Sided Giving

So far, the new insights have largely failed to spill over to a broader set of social preferences beyond the 'one-sided' individual decision-making. Here, we focus on one widely-discussed type of social preference, reciprocity. Reciprocity is the tendency to reciprocate kind acts with kindness and unkind acts with spite (Charness and Rabin 2002; Falk and Fischbacher 2006; Rabin 1993). A preference for acting reciprocally has been advanced as underlying many puzzling phenomena, both in the field, as discussed in the introduction, and in the laboratory. In what follows, we provide a brief review of some of the extensive literature on reciprocity, including a description of theoretical accounts for the phenomenon.

Evidence of reciprocity in the laboratory is widespread and robust (see also Fehr and Gächter (2000b) for a review). The pattern of reciprocal behavior survives in one-shot, anonymous scenarios (Hoffman, McCabe, and Smith 1998), persists over time and through learning opportunities (Keser and Winden 2000), and people even reciprocate on behalf of others (Carpenter and Matthews 2004; Carpenter, Matthews, and Ongonga 2004; Fehr and Fischbacher 2004), a phenomenon referred to as "indirect" or "social" reciprocity.

Positive reciprocity is observed as the response to costly investments by others in trust games, often strongly enough to yield a non-negative return to kindness (Berg, Dickhaut, and McCabe 1995). Pillutla, Malhotra, and Murnighan (2003) show that this reciprocal impulse is strengthened when the trustor takes a very unselfish, risky action to begin with, indicating that outcomes are not the only driving force behind reciprocity.

Negative reciprocity is observed as costly punishment in public-good games (Fehr and Gächter 2000a; Ostrom, Walker, and Gardner 1992). Croson (2007) uses variants of the public-good game to show that reciprocity is a stronger motivation than either altruism or the moral commitment to act in the manner one would prefer everyone to act.

Evidence from ultimatum game experiments, using populations from many countries, shows that reciprocity is a cross-cultural phenomenon that persists even when the stakes are as high as several months' wages (Camerer and Thaler 1995; Henrich et al. 2001; Roth and Erev 1995). A critical element driving reciprocity in such games seems to be the intention and agency behind the initial act (Blount 1995; Brandts and Solà 2001). Thus, in the ultimatum game, reciprocity manifests itself as costly punishment for "bad" behavior, which is evaluated based on more than just the resulting outcomes.

Closely related to trust and ultimatum games are findings on gift exchange, both in the laboratory and in the field. For example, a laboratory experiment by Fehr, Gächter, and Kirchsteiger (1997) finds that employers who offer high wages are rewarded with higher effort, even when wages are fixed at a flat rate and effort is non-contractible. Using a similar design, Fehr, Kirchsteiger, and Riedl (1993) show that buyers who offer a high price for a good are rewarded with higher quality.

Field experiments often find weaker evidence of gift exchange than the laboratory experiments, but confirm its existence and help to identify the conditions under which gift-exchange reliably occurs. For example, Gneezy and List (2006) find that workers reciprocate surprise gifts with higher effort, though the effect does not persist over time (see also List (2006)). Kube, Maréchal, and Puppe (2012) show that the type of gift is important: small non-monetary gifts and monetary gifts presented thoughtfully are better at inducing positive reciprocity.

The field data also confirms both the positive and the negative side of reciprocal behavior. On the side of positive reciprocity, Falk (2007) shows that donation requests are returned more frequently when they include a token gift (a postcard). On the side of negative reciprocity, Greenberg (1990) documents a rise in employee theft after wage cuts, and Krueger and Mas (2004) show that employee sabotage during a period of disputes with management was responsible for a dramatic increase in defective tires. Kube, Maréchal, and Puppe (2013) argue that negative reciprocity—induced by a surprise wage cut—is stronger and more robust than positive reciprocity in field settings.[4]

Models of reciprocity have naturally developed alongside these experimental results. Most fall in three broad classes: outcome-based, type-based (a.k.a. interdependent preferences), and intentions-based models.[5] We discuss these models here with an eye toward our main argument—the relevance of external determinants of reciprocal behavior—and the particular experimental analysis of a one-shot, non-strategic, anonymous game, to be discussed in the next two sections.[6]

In the category of *outcome-based models*, one type of model that can account for reciprocity is pure altruism. Altruists incorporate others' material outcomes into their utility just like any other good to which they wish to allocate wealth. This approach has been employed by many authors since Becker (1961). Bergstrom, Blume, and Varian (1986), for example, use this type of model to explain the private provision of public goods. Andreoni and Miller (2002) formulate a model of utilitarianism that has been widely used in the social-preferences literature. Impure altruism, also known as "warm glow" (Andreoni 1989, 1990), works similarly in that a giver obtains direct utility from improving another person's material outcome. In this case the utility stems from the act of giving, even though the giver does not care directly about the outcome of the recipient (e.g., the total donations obtained, including from other sources).

Another class of outcome-based models that can generate reciprocal behavior assumes distributional preferences. One example are maximin preferences as in Charness and Rabin (2002), where a giver would like the minimum payoff of any person to be as large as possible.[7] Another example is inequality aversion, as modeled by Fehr and Schmidt (1999) or Bolton and Ockenfels (2000). All of these models assert that we care about others' outcomes insofar as we care about our own outcomes relative to those of others. Related models incorporate absolute norms of fairness, as suggested by Ledyard (1997) and used by Cappelen et al. (2007).

[4] See also Baumeister et al. (2001).

[5] There are a number of additional theoretical approaches that do not fit neatly into the three categories. For example, some models account for behavior that might be considered reciprocal using bounded rationality (Gale, Binmore, and Samuelson 1995; Roth and Erev 1995). Cox, Friedman, and Gjerstad (2007) present a model of reciprocity in which the weight placed on someone else's material outcome is a function of one's "emotional state," which is in turn a function of the reciprocity motive, judged relative to a context-dependent neutral outcome. Sugden (1984) models reciprocity as an obligation: individuals feel obligated to contribute to a public good based on how much others are contributing. Yet other authors emphasize the role of evolution in supporting reciprocity as a motive, with a primary focus on the conditions under which reciprocity can be supported as an evolutionarily stable behavior (see Hoffman, McCabe, and Smith (1998), Gintis (2000) and Bowles and Gintis (2011)).

[6] Sobel (2005) and Cooper and Kagel ("Other-Regarding Preferences: A Selective Survey of Experimental Results") provide excellent additional discussions of reciprocity.

[7] Charness and Rabin (2002) also allow an agent to place a lower weight on others' payoffs when they have "misbehaved," thus incorporating an element of the type-based models discussed next.

Outcome-based models of distributional preferences predict reciprocal behavior as a byproduct of individuals trading off their personal material outcomes and fair outcomes. For example, a kind act by one party (e.g., a firm paying a high fixed wage) may create an inequality that the other party then seeks to mitigate through a reciprocal kind act. That is, reciprocity is implied by agents trying to re-balance their allocations in response to kind or unkind transfer from the others.

In the category of *type-based models*, or interdependent preferences, the seminal model of Levine (1998) allows the weight someone places on another's material outcome to depend both on a personal altruism parameter and the other person's altruism parameter. That is, nice people are nicer to others in general, and everyone is nicer to nice people. On the opposite end of the spectrum, the range of altruism parameters also allows for spite. Rotemberg (2008) develops a variant of this model, with the twist that decision makers treat others according to some default altruism parameter as long as their actions meet a "minimally acceptable altruism" threshold, but switch to a more spiteful altruism parameter if the other person's actions fall below that threshold.

These models treat reciprocity as the product of a signaling game. Actions signal a person's kindness, and this affects others' altruistic preferences toward that person. The type-based approach also accounts for indirect reciprocity, i.e., treating a person according to how they have previously treated others.

In the category of *intention-based models*, the foundational example is Rabin's (1993) psychological equilibrium model, in which beliefs enter directly into utility.[8] Players maximize a utility function that has both material utility and reciprocal kindness (or spite). The kindness (or spite) of an action is defined not only by what other actions are available to an agent, but also by the agent's beliefs about what the other player will do, insofar as these beliefs reveal the agent's intentions. If a player behaves neutrally, kindness and spite are absent, and the second player simply maximizes his material outcome. However, if the first player is kind—by taking an action that, based on what she believes the second player will do, makes the second player better off—then the second player may prefer to reciprocate those intentions. In equilibrium, each player's beliefs about the others' actions (and second- and third-order beliefs) must be correct, and players must optimally respond to those actions and to the "kindness" that is implied by those actions.

A key result of this approach is the existence of additional equilibria, relative to a game in which players consider only their own material payoffs. Here, players may act mutually kindly or mutually spitefully toward one another. This model diverges from models such as Fehr and Schmidt (1999) and Levine (1998) in that people care directly about psychological phenomena, rather than their preferences over everyone's consumption being changed by psychological phenomena. More concretely, people directly care about treating others the way they expect these others to treat them, and not solely about the resulting material outcomes. For example, in contrast with inequality aversion, the second mover is inclined to reciprocate a kind act of the first mover even if the first mover is much richer initially.

[8] See also Geanakoplos, Pearce, and Stacchetti (1989) and Battigalli and Dufwenberg (2009).

Rabin's (1993) model of fairness only applies to two-person normal-form games with pure strategies. Others have developed more general versions of similar models (Dufwenberg and Kirchsteiger 2004; Falk and Ichino 2006). Such intentions-based models have been corroborated experimentally by, for example, Dhaene and Bouckaert (2010), who show that Dufwenberg and Kirchsteiger's (2004) model is consistent with a large majority of people's behavior in an experimental setting, and by Falk, Fehr, and Fischbacher (2008) and Blount (1995), who show that both intentions and outcomes are necessary to explain behavior.

The Missing Piece

One way to frame the limitation of the three broad classes of reciprocity models is that they assume that the opportunity to invest in others' outcomes is always a welfare-enhancing expansion of the choice set. That is, the mere option to share with others cannot hurt the decision maker, but may increase utility.

As we reviewed above, the same limitation applied to the earlier models in 'one-sided' giving contexts. The key assumption used to be that individuals share because they like to share. More recently, however, the introduction of avoidance options into laboratory and field settings revealed that a majority of people share reluctantly. They share if asked, but prefer to avoid the sharing request. In response to these findings, the new wave of modeling approaches has focused on external factors, such as social pressure and social image.

Research on reciprocity has largely not yet incorporated these more recent approaches.[9] The neglect of external factors could bias our understanding of reciprocity, as demonstrated with the following experimental test.

2.3 External Motives in Reciprocity: An Experimental Test

The question of whether or not sharing in reciprocity environments is fully explained by internal motives is an empirical question. One way to explore this question is to mirror the approach from one-sided giving experiments and introduce avoidance options into reciprocity settings. Here, we describe a novel experiment intended to test how reciprocity affects avoidance. In particular, we apply the experimental approach of Lazear, Malmendier, and Weber (2012), which allows agents to avoid making a decision and keep all of their wealth, to a reciprocity setting.

[9] For a rare recent exception, see Weele et al. (2014). They study providing second-movers in reciprocity games with a means for plausibly denying resposibility for failing to act reciprocally. They find that this manipulation has little effect on reciprocal behavior, suggesting that the relative importance of internal and external motivations is different between one-sided and two-sided acts of giving, which is in contrast with the evidence we discuss in the next section. This conflicting evidence highlights the importance of further work for identifying the relative importance of external motives for pro-social behavior in the context of reciprocity.

Experimental Design

In a standard dictator game the decision maker has no prior interaction with the (typically anonymous) recipient before making a decision. Hence, reciprocity is irrelevant, and the dictator's choice only involves trade-offs between personal material payoffs, internal motivations to share (e.g. altruism), and external motivations to share (e.g. social pressure). To identify the importance of external motives, Dana, Weber, and Kuang (2007), Broberg, Ellingsen, and Johannesson (2007), and Lazear, Malmendier, and Weber (2012) gave dictators the option to avoid the decision to share and keep the endowment without the potential recipient learning about the sharing opportunity. They found that a large fraction of "sharers" decided to opt out, indicating the importance of external factors.

We conducted an experiment using variants of a "double dictator game" (DDG) to test whether external determinants also affect sharing in a reciprocity environment. Here, dictator and recipient play a mini-dictator game over $2[10] prior to the main dictator game, with the dictator and recipient roles switched. The parties learn about the second stage (the main DG) only after the mini-game has been played and the results revealed to the dictators of the main DG.[11] The purpose of the initial reversed mini-dictator game is to induce reciprocity as a motive for sharing in the second-stage main dictator game. Positive reciprocity is induced if the mini-dictator decides to share $1 out of a $2 endowment. Negative reciprocity is induced if the mini-dictator decides not to share. This binary choice is the only option available, allowing clean assignment to reciprocity treatments. The standard dictator game (DG) is the neutral "no reciprocity" benchmark for comparison.[12]

We cross these reciprocity treatments with a sorting option—as in Lazear, Malmendier, and Weber (2012)—which allows the second-stage dictator to avoid the sharing decision and keep the endowment. In the no-sorting conditions, the dictator is forced to choose an allocation, of which the recipient is then informed; in the sorting conditions, the dictator has the option to costlessly opt out of the game, thus receiving the full endowment but leaving the potential recipient uninformed about the game.

These six conditions allow us to evaluate the role of internal and external determinants of sharing. As described above, under leading models of reciprocity, giving is driven by internal motivations, and the sorting option should be irrelevant after positive reciprocity is introduced. Thus, while we know that about half of the dictators who share in a simple dictator game prefer to sort out if possible, the existing models of reciprocity predict no such sorting behavior after reciprocity has been induced. Similarly, negative reciprocity induced by not sharing in the prior mini-game should induce spite and a willingness to punish the other party, which makes people happy to share nothing regardless of whether they can opt

[10] These small stakes allow us to distinguish reciprocity from distributional preferences such as inequity aversion.

[11] Initial voluntary sharing can thus be interpreted as an act of kindness, rather than an attempt to induce reciprocal behavior, which avoids concerns about interpreting the dictators' reactions as reciprocity.

[12] With this comparison, the experiment avoids confounds with other social preferences, as cautioned by Cox (2004).

out.

Alternatively, external motives may still play a role after reciprocity has been induced. For example, it is possible that positive reciprocity increases the internal motivation, but does not eliminate the external motivation. In this case, we would expect to see sorting out in the DDG after the mini-dictator has shared, but less than in the standard DG with sorting. We would also expect the average amount shared to increase. Another possibility is that external motives affect reciprocal behavior directly. For example, positive reciprocity could increase the image cost of being ungenerous. In this case the predictions regarding opt-out frequency and average amount shared are ambiguous. With both internal and external determinants becoming stronger, both quantities could increase or decrease since sorting in is now more costly due to increased social pressure, and sorting out is more costly due to increased internal motivation. However, we do have a clear prediction for average sharing conditional on not opting out: those deciding to play the (main) dictator game will share more. We explore these possibilities in the context of a formal model further in section 2.5.

Full implementation details for the experiment are provided in Appendix C.1.

Experimental Results

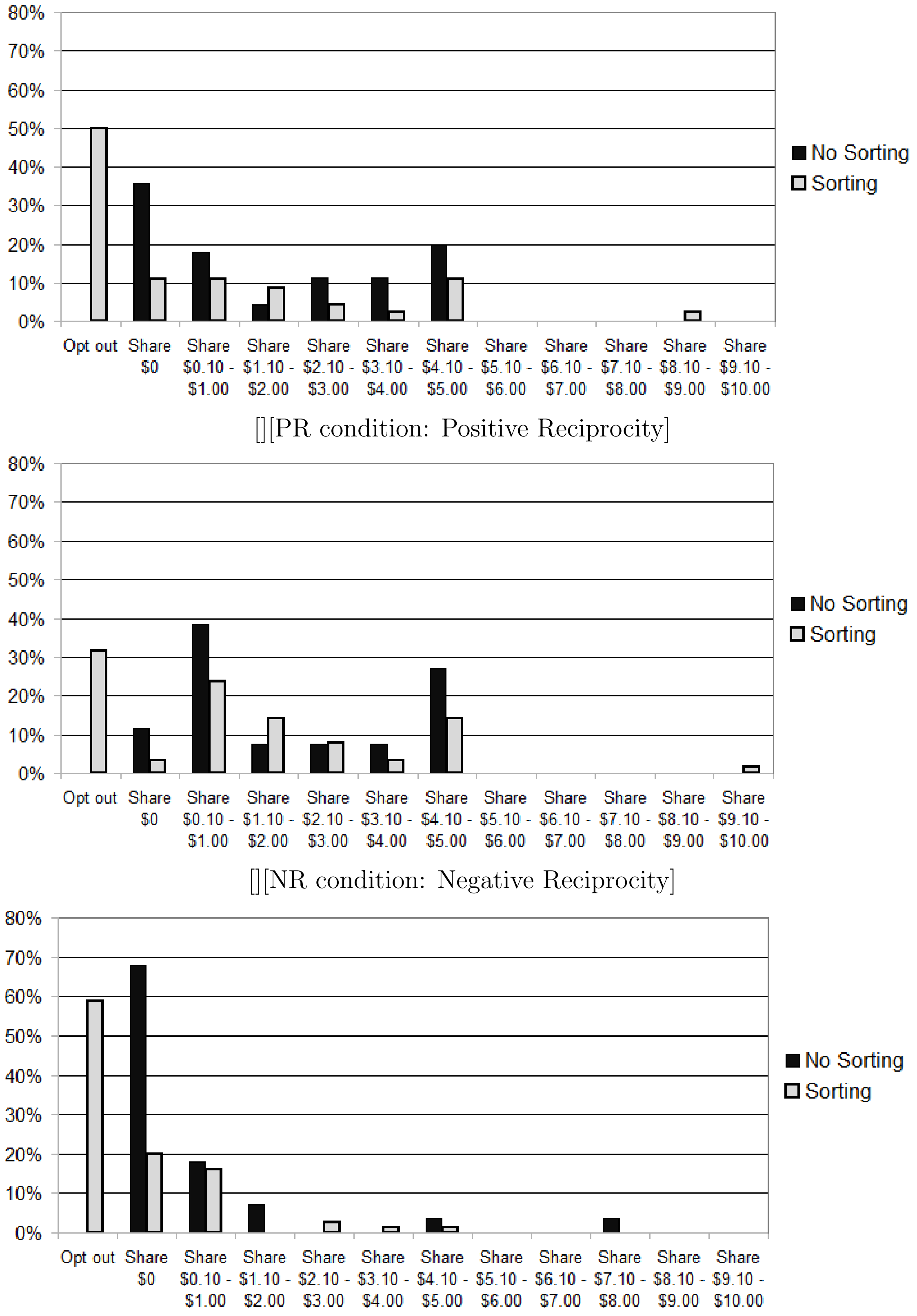
Detailed description of the reduced form results and supporting analyses are described exhaustively in Appendix C.2. We briefly summarize the results here. For example, Figure 2.1 shows the distribution of amounts shared in each condition, while Table 2.1 presents a summary of reduced-form statistical tests, from which one can also infer the mean amount shared in each condition.

We first consider the effect of reciprocity treatments on sharing when sorting opportunities are not available. The average amount shared, out of $10, in the standard (no-sorting) DG treatment is $2.00, which rises to $2.39 in the positive-reciprocity (PR) condition and falls to $0.70 in the negative-reciprocity (NR) treatment. Column 1 shows that only the negative-reciprocity effect is significant. In other words, there is a significant negative-reciprocity effect (−.130) and an insignificant positive-reciprocity effect (.039),[13] which is consistent with previous experimental evidence (e.g., weak positive reciprocity but strong "concern withdrawal" in Charness and Rabin (2002)).

The effect of sorting on sharing: In the standard DG, most subjects share a positive amount (64 percent) and the mean amount shared is $2.00. However, the introduction of sorting strongly decreases the average amount shared, to $1.21. The sorting opportunity also decreases the frequency of sharing dramatically, to only 39% sharing a positive amount. In the PR treatment, sorting again causes a large drop in average amounts shared, to $1.71, while sorting causes sharing in the NR treatment to drop from $0.70 to $0.31. Columns 2 through 6 confirm the statistical significance of these findings.

[13] Standard errors are robust to heteroskedasticity and adjusted for small-sample bias, using the residual-variance estimator HC3, which approximates a jackknife estimator (MacKinnon and White 1985). If we cluster by session, standard errors in this and in all other estimations are very similar and typically slightly smaller, though unlikely to be reliable given the few clusters.

Figure 2.1: Distributions of Amounts Shared



[][DG condition: No Reciprocity]

[][PR condition: Positive Reciprocity]

[][NR condition: Negative Reciprocity]

Table 2.1: Effect of Sorting on Sharing with Reciprocity

Model:	**OLS**			**Tobit**		**Probit**
Dependent Variable:	Proportion Shared					Shared Something
Sample	**Baseline (No Sorting)**	**All Data**				
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.200***	0.193***	0.200***	0.119***	0.131***	
	(0.030)	(0.024)	(0.030)	(0.041)	(0.048)	
Negative Reciprocity	−0.130***	−0.105***	−0.130***	−0.234***	−0.257***	−0.319***
	(0.045)	(0.026)	(0.045)	(0.056)	(0.090)	(0.111)
Positive Reciprocity	0.039	0.438	0.039	0.112**	0.089	0.319**
	(0.049)	(0.031)	(0.049)	(0.049)	(0.066)	(0.13)
Sorting		−0.063***	−0.079*	−0.135***	−0.161**	−0.253**
		(0.024)	(0.043)	(0.042)	(0.074)	(0.101)
Sorting ×			0.040		0.042	0.125
Negative Reciprocity			(0.055)		(0.119)	(0.157)
Sorting ×			0.011		0.041	−0.065
Positive Reciprocity			(0.063)		(0.096)	(0.177)
Observations	99	283	283	283	283	283
(pseudo) R^2	0.113	0.145	0.147	0.209	0.209	0.150

Independent variables are condition dummies. The tobit model accounts for 147 observations being left-censored at zero. The probit model shows marginal effects. Robust standard errors are in parentheses (with bias-correction (HC3) in the linear case, see MacKinnon and White (1985)) and are calculated using jackknife estimation for the tobit model.

* - $p < 0.1$; ** - $p < 0.05$; *** - $p < 0.01$

Figure 2.1 presents the distribuitons of individual behavior across conditions. (We display the frequencies of subjects who opt out separately at the left.) We observe a sharp shift of the distribution to the left when sorting becomes possible, regardless of the reciprocity conditions. By categorizing choices as sharing nothing, a small amount, or a large amount, we also find a statistically significant drop in generous sharing and a significant increase in sharing nothing when sorting is introduced.[14] This is true in all three reciprocity treatments, and the magnitude of the effect is not significantly different between reciprocity treatments. Hence, both the simple comparison of means and the distributional evidence suggest that sorting has a large impact on sharing even after reciprocity has been induced, inconsistent with a solely intrinsic motivation for reciprocal behavior.

We conclude that sorting retains its impact even after we induced either positive or negative reciprocity. The option to sort significantly decreases sharing, and shifts the distribution toward zero sharing, in both reciprocity conditions. In other words, givers who respond to a previous kind or unkind act are affected by the option to avoid the opportunity to give. This evidence suggest that the dominant approach to modeling sharing under reciprocity, which relies on internal factors as the primary motive, is incomplete. External, presentational, factors affect individuals' giving in reciprocity environments as well, reflecting in the change in sharing behavior in environments with sorting opportunities.

But, how strong are external factors after the inducement of positive or negative reciprocity? Are they identical to the degree found in neutral settings, or does reciprocity lead to differential sorting effects? The simple comparison of means suggests that the effect is smaller under positive reciprocity and larger under negative reciprocity than in the neutral DG setting, though the differences do not seem large. Sharing drops by 40% in the standard DG when sorting is introduced, by 29 percent in the PR condition, and by 56 percent in the NR condition. Columns 3 and 5 of Table 2.1 reveal that the significant decrease due to sorting does not differ significantly, from the DG condition, in either the NR or PR conditions. The same picture emerges if we consider the frequency of sharing, i.e., the fraction of subjects who share any positive amount. The probit regression in the final column of Table 2.1 shows that 25 percent of sharers opt out, but the interactions of Sorting with either reciprocity condition, PR or NR, are statistically insignificant.

In other words, the impact of sorting on average amounts shared is large and significant, and *approximately* invariant to the reciprocity setting. We will investigate these reduced-form findings in more detail in our structural estimation in Section 2.5, where we provide quantification of the role of internal and external factors.

The Effect of Reciprocity on Sorting: The above analysis of amounts shared suggests that reciprocal behavior is affected by external factors. Otherwise, giving would have been unaffected by the opportunity to avoid the giving request. We can go one step further and

[14] See Appendix C.2 for details. Mid-level sharing doesn't change significantly. Without within-subject data, of course, it's hard to know whether this reflects and overall shift to the left, or whether high sharers are converting to zero sharers when sorting is an option. However, the structural analysis of section 2.5 sheds light on this issue.

analyze to what extent the reduction in giving comes via the channel of opting out. In other words, compared to the simple DG, how does reciprocity impact the choice to opt out?

The leftmost bars in each of the three graphs in Figure 2.1 reveal that the sorting option is used by a significant fraction of subjects in all treatments. In the DG condition, 50 percent of all subjects choose to opt out. In the PR condition, many still sort out, but fewer than in the DG baseline (32 percent). The NR condition incites more sorting behavior than the PR condition, but only slightly more than the baseline DG (59 percent). The differences between sorting under reciprocity and sorting in the neutral DG setting are either insignificant or only marginally significant, as described in Appendix C.2.

The sorting evidence speaks to the presence of both internal and external factors affecting reciprocal sharing decisions. If sharing was fully explained by internal determinants then subjects should not make use of the option to sort out. If anything, subjects who do not share would be indifferent. Instead, we find that a significant fraction of subjects use the sorting option even after reciprocity has been induced. At the same time, we do see that the use of the sorting option decreases after positive reciprocity has been induced, suggesting that a kind initial treatment strengthens the internal motivation. But, by the same logic, negative reciprocity should make people intrinsically less willing to share, and therefore also less likely to avoid the sharing environment. In other words, we would expect that non-sharers are not afraid to sort in and reveal their non-sharing decision (anonymously) to a partner who did not share himself. The small and insignificant increase in opting out in the NR treatment, relative to the DG, along with the large observed drop in actual sharing discussed above, hints at problems with this interpretation.

To be clear, moving from DG to NR, we do observe variation in "spiteful non-sharing," i.e., in the fraction of subjects who sort in but share nothing. As can be seen in Figure 2.1, only a small number of subjects sort in and share zero in the DG (11 percent). Positive reciprocity reduces this rate to 3 percent (a 72 percent reduction). Meanwhile, in the NR condition, 20 percent of dictators sort in to share nothing (a 46 percent increase over DG). We infer that reciprocity has an impact on spiteful non-sharing, with further details deferred to Appendix C.2.

2.4 Discussion

What are the implications of these results for our understanding of reciprocity and for theories that describe it? Existing theories have to wrestle with two key findings:

1. In all three conditions, the sorting option has a significant impact on the distribution of sharing amounts and the average amount shared, and the impact is approximately invariant to reciprocity.

2. Moving from DG to PR, we observe less opting out and less spiteful non-sharing. Moving from DG to NR, we observe more opting out and more spiteful non-sharing. The increase in spiteful non-sharing under NR is economically large.

We briefly recap why existing models fail to predict these findings. To preview the gist of the discussion, existing reciprocity models predict that the sorting option has zero impact on the average or the distribution of gifts, since the only people who sort out are those who are indifferent between sorting out and sharing zero. This contradicts the first point. Additionally, without richer assumptions regarding indifference, the fraction who sort out and the fraction who sort in and share zero (spiteful non-sharers) are predicted to increase and decrease in proportion to each other, invariant to the reciprocity setting. This contradicts the second point. Additionally, the results comparing the impact of sorting across reciprocity settings are not spoken to by these models, due to their common prediction of zero impact of sorting in all cases. We propose several alternative explanations that might fill this gap.

In discussing the predictions of the outcome-, type-, and intentions-based reciprocity models from Section 2.2, we will need to make an additional assumption to solve the cases of indifference. All models imply that non-sharers are indifferent between sorting out of the dictator game and sorting in (and keeping all of the money). We assume that dictators choose to sort out with some fixed probability that is not context-dependent. This assumption allows, in particular, for “sorting out as default” and “randomize” as possible heuristics.[15]

First, consider outcome-based models, such as Fehr and Schmidt (1999). In the DG, types who put sufficiently high weight on fairness or others’ consumption will share a positive portion of the $10. Everyone else will be indifferent between sorting out or sorting in but sharing nothing. In the DDG, however, the initial distribution of an additional $2 might induce the dictator to share more or less to achieve his most preferred distribution of wealth.[16] In the negative reciprocity condition, the recipient has a head start of $2. Hence, a higher threshold for weight on fairness is required for the dictator to share a positive amount, and positive amounts shared will be smaller. In the positive reciprocity condition, dictators and recipients both have a head start of $1. Depending on the exact outcome-based model formulation, this could cause the threshold level of fairness for positive sharing to go up (if dictators no longer feel the need to ensure a small minimum payment for the recipient) or down (if the marginal utility of 11th dollars is particularly small) or stay the same (if dictators are merely inequity-averse), relative to the standard DG. However, the mere introduction of a sorting option has no impact on the distribution of positive gifts. Instead, any change in sharing behavior between reciprocity conditions is driven by a shift in the distribution of choices upwards or downwards, with the portion of the distribution censored at 0.

With regard to our data, these models therefore successfully predict that the distribution of giving shifts downward in the negative reciprocity case, compared to the baseline DG. They do not necessarily capture the increase in sorting in and in average gifts in the positive reciprocity conditions. With our additional assumption that non-sharers who are indifferent

[15] More complicated rules may improve compatibility of theory and data, such as allowing the probability of sorting out to depend on type or context, but would require the model to formulate these dependencies more explicitly, so we do not discuss these possibilities in this section.

[16] In this section, labels of players always refer to their role in the primary (second-stage) dictator game. For example, the decision-maker in the first-stage mini-DG is referred to as the recipient. Throughout, male pronouns are used for dictators, and female for recipients.

between sorting options choose to sort out some with probability p, they also predict that any change in actual sharing should be accompanied by a change in sorting choices. This is indeed the case in the positive reciprocity game (more people sort in to share a positive amount). But it is not the case in the negative reciprocity treatment. Here, we find an increase in spiteful non-sharing without a proportional increase in sorting out.[17]

Next, consider type-based models, such as Levine (1998). In these models, the dictator's altruism toward the recipient depends on his information about the recipient's type. In the standard game, then, since the dictator has no information about his partner, he simply chooses his optimal wealth allocation given his knowledge about the population distribution of types. In the reciprocity conditions, however, the dictator can infer something about his partner based on whether he received \$1 or \$0.[18]

After receiving \$1 in the PR condition, then, the dictator has strictly more favorable beliefs about the recipient than in the DG condition, and should place a higher weight on her consumption. In the NR condition, conversely, the dictator places strictly less weight on the recipient. Hence, type-based models can explain the increase in sorting in and giving that we observe in the PR conditions. As in the outcome-based models, however, this directional prediction is ambiguous: dictators may be less likely to share in the PR than in the DG condition, despite putting a higher weight on altruism, because the recipient is already starting out with \$1. Moreover, as in the outcome-based models, any changes in giving are driven by a shifting distribution of ideal allocations, and are not affected by a sorting option. Hence, the results about differential effects of the sorting option along with the rates of spiteful non-sharing cannot be explained by type-based reciprocity.

Lastly, consider intentions-based models. Since the recipient does not know about the later dictator game when she makes her mini-DG choice, we can analyze behavior as two normal-form games and apply the version of the model in Rabin (1993) (even though our game is technically an extensive form game). Also, note that, in our context, both players know their partner's actions (within the game they believe they are playing) so that there is no concern about distinguishing beliefs from actions.

In the DG, the recipient has taken no action and is thus neutral. The dictator should therefore maximize his material outcome. Likewise, in the initial mini-dictator game in the DDG, the recipient believes that the dictator cannot respond and is thus neutral, and should also maximize her material outcome by keeping the \$2.[19] The recipient's "kindness" in the

[17] Additionally, note that outcome-based models of reciprocity trivially predict that the difference-in-difference effects of a change in the impact of the sorting option across reciprocity settings is zero, since they predict zero impact of a sorting option in any context. We, in fact, find a small (insignificant) diff-in-diff in the level of average gifts, but it results from large, statistically similarly sized impacts across settings (see Table 2.1).

[18] Since the recipient was not informed about the main dictator game prior to making her choice in the mini game, this choice cannot be interpreted as a strategic signal.

[19] A more realistic version of the material outcome utility might incorporate some other form of social preferences such as inequity aversion or pure altruism; but to avoid misattributing implications of these other models to intentions-based preferences, we accept for now the strict material self-interest interpretation.

mini-dictator game measures the impact of her choice on the dictator's outcome relative to the neutral average. Hence, it is negative if she shares nothing and positive if she shares $1. In the negative case, the dictator simply maximizes his material outcome as there is no avenue for costly punishment of the recipient. In the positive case, however, any dictators with a high enough weight on reciprocating intentions shares a positive amount. (Note that since players do not have utility over their partners' outcomes directly, but only over each others' reciprocal intentions, this model makes unambiguous predictions in the PR game.) However, this model predicts, once again, that the sorting option should have no impact on the distribution of positive gifts in any condition. Hence, the other conflicts with the data are not addressed by intentions-based reciprocity any better than with outcome- or type-based models.

So what can explain the experimental findings? The sorting effect in the simple DG shows us where to look for an explanation for the sorting effect in the DDG as well. Between the standard dictator game and the same game with a sorting option, the only change is that the information acquired by the recipient becomes manipulable by the dictator. If the dictator sorts in, the recipient will know how her monetary outcome was determined, but if he sorts out, she will not. Anyone who modifies his behavior when given a sorting option must do so in response to this additional choice over information. The lack of such a behavioral sensitivity to observability provides an interpretation for the inadequacy of most reciprocity models to account for sorting behavior.

Earlier, we argue that the "missing piece" in existing models is external motivation. Individuals may share not because they care about the other person's payoff, but because they feel obliged or pressured to do so, and would rather avoid being in this situation. There is a variety of terminology describing such external motivations, including social image, social pressure, social signaling, audience effects, prestige, shame, guilt, and reputation. We will sometimes refer to "social image" or "social pressure," but our analysis does not pin down the exact form of external motivation. Rather, it illustrates that a model of reciprocity that incorporates external factors can predict and provide an interpretation for the above results.

Before we turn to the model and structural estimation in the next section, we briefly provide the intuition for our approach. Suppose that a dictator trades off monetary payments, internal factors such as fairness or distributional preferences, and external factors such as social image and social pressure in deciding whether to share. In the standard DG, a dictator who shares nothing would prefer to sort out if a sorting option becomes available, as this leads to identical outcomes but a better social image. A dictator who shares a positive amount in the standard DG might also prefer to sort out if possible. This would be the case, for example, if the material benefit from keeping the additional money and the benefit from not feeling social pressure to give more than intrinsically desired outweigh the costs of not obtaining social image benefits or of experience internal guilt. In either case, extrinsic determinants are needed to predict sorting out. Additionally, sorting could impact the distribution of gifts. For example, if generous givers in the standard DG are mostly individuals who are particularly susceptible to social pressure, they will make use of the sorting option if available and the distribution of amounts given will change significantly if the option to sort becomes available.

In the positive reciprocity scenario of the DDG, external factors might become less important as positive reciprocity strengthens the internal motivations for giving. Alternatively, positive reciprocity might strengthen external motives and make it less acceptable to share little. If this manifests as a larger image reward for sharing, people might share more. If it manifests instead as a larger image punishment for imperfect sharing, it might drive more people to sort out. In the negative reciprocity scenario, less sharing might reflect not only weaker internal motives to share, but also lower external motives. For example, the pressure to share may become lower and a dictator who would share a positive amount in the no-sorting setting might now believe it to be fair to share nothing. Alternatively, a dictator may share less simply because of a weaker internal motivation to share. However, only external factors such as social image concerns predict an increase in sorting in to share zero, namely as a boost to social image resulting from punishing unfair peers or as a relief from external pressures to share.

Hence, a model that allows external determinants of giving may provide an interpretation for the reduced form findings, including changes in donations and sorting-in rates that are not positively correlated. The structural estimations will provide a clearer picture of the relative strength of the underlying motives and their interactions with reciprocity than we can glean from the reduced form analysis.

2.5 Model and Estimation of Extrinsic Motives

We now present a simple model of reciprocity that incorporates both internal and external determinants of sharing. Many alternative models of social image and other extrinsic determinants of reciprocity are possible, and further experimentation is needed to pin down the exact form of preferences. Nevertheless, following the logic above, even a simplistic model of image-based reciprocity is sufficient to convey the underlying mechanism. We confirm this using a structural estimation, described in the next section.

Model

The outcome-, type-, and intentions-based models of reciprocity described above can be thought of as describing internal determinants of sharing whose weights vary with the reciprocity environment. That is, the dictator keeps an amount $x \in [0, 10]$ such that

$$U_r(x) = x - \alpha_r G_r(x),$$

where $\alpha_r G_r$ captures internal motives to sharing as disutility from keeping too much. G is an increasing function of x, up to some maximal level from which the dictactor derives utility from sharing, $\bar{x}$, and may vary depending on the reciprocity environment $r \in \{\text{DG,NR,PR}\}$, and $\alpha_r \in \mathbb{R}$ is the weight assigned to sharing the “right amount” and may also vary with the reciprocity environment. As before, we will assume that non-sharers, who are indifferent between sorting out and sorting in (and keeping all of the money) sort out with some fixed probability that is not context-dependent.

The simplest way to incorporate external factors into this framework is to add a parallel weight $\beta_r \in \mathbb{R}$, applied to an increasing function H, which kicks in only if the dictator's actions are observable to the recipient, i.e., if he sorts in:

$$U_r(x) = x - \alpha_r G_r(x) - \beta_r \mathbb{1}(\text{sort in}) H_r(x)$$

To minimize the degrees of freedom and allow model identification, we impose the following specification:

$$U_r(x) = x - (\alpha_r + \beta_r \mathbb{1}(\text{sort in}))(x-5)^2 \tag{2.1}$$

A few details deserve comment. First, the quadratic loss function around the (narrowly-bracketed) 50-50 split implies that people want to be generous but not too generous.[20] This functional form predicts that no one will give more than \$5, and indeed almost no one does. It also has the benefit of predicting the second mode of giving, at the 50-50 split: anyone above a certain threshold of intrinsic motivation will share exactly \$5. We will verify, in the robustness checks in Section 2.5, that the specification is not critical to our findings.

Second, the fact that the loss function is symmetric around \$5 in all reciprocity treatments means we assume narrow bracketing, rather than incorporating the initial \$2 mini-DG into the utility function. However, a global-bracketing 50-50 split would again prescribe sharing 5 in the PR condition, and sharing 4 in the NR condition. This appears to be irrelevant to our findings; we continue to observe a few people sharing exactly \$5 in the NR condition, and no comparable bump in the distribution at \$4 (see Figure 2.1).

Third, the fact that the external motivation is framed as a loss ($-\beta_r \mathbb{1}(\text{sort in}) H_r(x)$) implies that a giver cannot gain utility (as long as β is positive). At best, he obtains no loss in utility, namely when sharing the "fair" amount of \$5. For example, in our specification in (2.1), sharing less than \$5 can never lead to an increase in utility.[21]

Fourth, treating α and β additively implies that internal and external determinants of sharing operate in parallel, with one exception: an agent may feel spiteful despite societal pressure to share, or may feel compelled to share despite pressure to punish. That is, the two parameters may have opposite sign.

[20] This is also the view supported by, e.g., Andreoni and Bernheim (2009). The debate over whether over-generosity induces guilt and a bad social image is still active (see Krupka and Weber (2013)), but is not relevant to our estimations.

[21] And, vice versa, if $\beta < 0$, i.e., if there is a reward for punishing the other person, then sharing any amount leads to an increase in utility. This is almost certainly not true. It seems plausible that being "sufficiently fair" or giving "sufficient punishment" would suffice to break even. The arguments apply also to the internal motivation for sharing. However, this simplification merely limits our ability to estimate the absolute utility impact of presenting someone with an opportunity to share. (Our treatments are not designed to identify the cutoff for giving to be welfare increasing.) We relax this assumption in several robustness checks and verify that this assumption is not critical for our results (see Section 2.5.)

The sharing implications of this model are straightforward. In the absence of social pressure or other external factors, dictators would like to keep the amount $x^* = 5 + 1/(2\alpha_r)$ (rounded to the nearest available discrete choice).[22] But in the presence of external factors– e.g., when dictators sort in—they will face social pressure and hence keep the adjusted amount $x^s = 5 + 1/(2(\alpha_r + \beta_r))$.[23] If the utility obtained with this adjusted choice is smaller than the utility when selfishly opting out and sharing zero, they will sort out if possible. Higher internal motivation to share increases the cost of sorting out, and higher social pressure increases the cost of sharing moderately.

Figure 2.2 illustrates how the two type parameters, α and β, determine choices when sorting is allowed, for a given reciprocity environment. When $\beta < 0$, no one chooses to opt out. That is, if selfishness induces a good social image, there is no benefit to sorting out rather than visibly sharing nothing. Furthermore, more negative levels of β must be met by increasingly high levels of α in order for people to be persuaded to share anything at all. When $\beta > 0$, instead, higher levels of α still lead to more generosity among those who sort in, and higher levels of β force relatively ungenerous people to share more generously. But now, people who want to give small amounts would rather sort out and experience internally motivated disutility from being selfish than sort in and be seen as greedy.

The figure shows how the availability of a sorting option influences the average level of giving and the distribution of gifts (among those who share) in different ways. At low levels of β, introducing a sorting option will most affect the lowest givers, so the impact on average generosity should be small. When β is higher, many generous givers (with low α's) will also opt out. Conversely, at low levels of α, a sorting option will affect givers, and the opting out of the most generous givers (with high β) will have a large impact, while givers with higher levels of α are more immune to the sorting option overall.

The figure also illustrates why existing models of reciprocity fail to explain the aspects of our data discussed in Section 2.3. Looking at Figure 2.2 along the $\beta = 0$ axis, the only individuals who care to sort out are those who are also willing to sort in and share nothing ($\alpha = 0$). Hence, the systematic decrease after PR and increase after NR is not predicted. Additionally, a mere shift in intrinsic motivation cannot explain a change in the distribution of gifts since, again, the only people willing to sort out are those who would share nothing in any case.

Finally, another way to look at Figure 2.2 is to consider the possibility that the distributions of β_{DG}, β_{PR}, and β_{NR} are the same, while the distributions of the α_r's may vary, for example, $\alpha_{NR} < \alpha_{DG} < \alpha_{PR}$. The figure shows that, as the distribution of α moves to the right, people share more and opt out less. In other words, the impact of sorting on the unconditional average level of sharing and rate of sorting should be smaller when the recipient has acted more kindly.

[22] Note that x^* is a maximum if and only if $\alpha_r > 0$. If $\alpha_r < 0$ then x^* is a minimum. Therefore, in this case, we would have a corner solution.

[23] Note that x^s is a maximum if and only if $\alpha_r + \beta_r > 0$. Otherwise, we will have a corner solution. This becomes pertinent in the negative reciprocity case, in which we estimate that $\alpha_r + \beta_r < 0$.

Table 2.2 summarizes the predictions of three possible variants of the model in terms of the patterns we observed in the data. All three variants allow for internal motivations, such as altruism or fairness, and assume that their strength is increasing in the prior kindness of the recipient. (For the purpose of the table, we consider reciprocity, r, to be a continuous variable, where a more positive value means a more positive reciprocity context, i.e, kinder prior treatment.) The first variant, shown in the first row, does not allow for external motivation ($\beta = 0$). This variant captures the predictions of the outcome-, type-, and intentions-based models discussed earlier. The second variant, shown in the second row, allows for external motivation, but assumes that it is unaffected by prior kind treatment. The third variant, shown in the third row, also allows external motivation to increase in the degree of prior kindness. The last row shows the data.

While all models are consistent with reciprocity increasing the average amount given and decreasing the frequency of opting out, a model without external motives cannot explain why the use of the sorting option and the distribution of gifts vary systematically with reciprocity, as indicated in columns 4 and 5. (Note that, as discussed, the data is not entirely clear on whether this prediction holds.) Moreover, column 3 shows that the increase in spiteful non-sharers in the negative-reciprocity setting is the main reason why an external factor that is constant across reciprocity treatments seems insufficient.

The table also clarifies that the reduced form results are inconclusive about the relationship between the reciprocity environment and the relative magnitudes of α and β. (See the "possible" entries in row 3, and "maybe" in row 4.) Our structural estimation will paint a clearer picture of the impact of reciprocity on internal and external motivations for sharing. The variable rates of opting out, across treatments, and the change in the distribution of shared amounts conditional on sorting in are the main sources of variation that allow us to identify α and β from the data.

Estimation

Exhaustive details on the estimation procedure and results are provided in Appendix C.3, but we summarize the main findings here.

We estimate the parameters μ_{α_r}, σ_{α_r}, μ_{β_r} and σ_{β_r} for each $r \in \{\text{DG,NR,PR}\}$, using minimum distance estimation. For the vector of moments in the baseline estimation, we break down the choices of giving into bins: exactly 0, from 25 cents to \$2.50, from \$2.75 to \$4.75, exactly \$5, and more than \$5. In the sorting conditions, an additional moment specifies the fraction who sort out. Altogether, we have 11 moments in each reciprocity environment, or 33 total. We also assume that α_r is normally distributed according to $N(\mu_{\alpha_r}, \sigma_{\alpha_r})$ and β_r is similarly distributed according to $N(\mu_{\beta_r}, \sigma_{\beta_r})$.

An individual, i, with type parameters, α_i and β_i, in a particular reciprocity environment will share $x^s = 5 + 1/(2(\alpha_i + \beta_i))$ (or the closest element of the discrete choice set) if he cannot opt out; he will sort in and share x^s even if he can opt out if $U(x^s) > 10 - 25\alpha$ (and otherwise will sort out). This threshold allows us to simply integrate over the distribution

Table 2.2: Model Predictions and Data

	1) Gift distributions sorting dependent	**2)** Sorting decreasing in r	**3)** Spiteful non-sharing decreasing in r	**4)** Impact of sorting on giving decreasing in r	**5)** Average giving increasing in r
No extrinsic motives: $\partial\alpha_i/\partial r > 0$, $\beta_i = 0$	no	yes	no	no	yes
Constant extrinsic motives: $\partial\alpha_i/\partial r > 0$, $\beta \sim\ F_\beta\ \forall r$	yes	yes	no	yes	yes
Variable extrinsic motives: $\partial\alpha_i/\partial r > 0$, $\partial\beta_i/\partial r > 0$	yes	possible	possible	possible	possible
Data	yes	yes	yes	maybe	yes

Notes: Comparison of three restrictions on the distributions of α and β in the model of Section 2.5.

1) Gift distributions are dependent on sorting: If $x_{r,S} \sim F_{x,r,S}$ and $x_{r,NS} \sim F_{x,r,NS}$, then $F_{x,r,s} \neq F_{x,r,NS}$.

2) Sorting is decreasing in r: $P[U(10|\text{sort out}, \text{NR}) > U(x^s|\text{sort in}, \text{NR})] < P[U(10|\text{sort out}, \text{PR}) > U(x^s|\text{sort in}, \text{PR})]$.

3) Spiteful non-sharing is decreasing in r: $P[U(10|\text{sort in}, \text{NR}) > U(10|\text{sort out}, \text{NR})] < P[U(10|\text{sort in}, \text{PR}) > U(10|\text{sort out}, \text{PR})]$. Note that "spiteful" indicates a *strict* preference for sorting in and sharing nothing; equivalently, the decrease with reciprocity in sorting in and sharing nothing is disproportionately more than the decrease in sharing nothing overall.

4) Impact of sorting on giving is decreasing in r: $\overline{x}_{r,S,NR} - \overline{x}_{r,NS,NR} < \overline{x}_{r,S,PR} - \overline{x}_{r,NS,PR}$.

5) Average giving is increasing in r: $\overline{x}_{NR} < \overline{x}_{PR}$.

"Yes" indicates that a prediction is required by a given model, "possible" indicates that the prediction is consistent with but not required by a model, and "no" means the model is not able to predict that aspect of the data. "Maybe" indicates that the data produces a small but statistically insignificant effect consistent with the prediction. We assume here that α and β have full support, so that even if with some distribution of types most people don't change behavior when r changes, as long as someone does, the model is still said to predict a change overall.

of types within the respective intervals to calculate the total fraction that fall within each choice category.

Column 1 of Table 2.3 shows the results. We estimate μ_α in the dictator game to be significantly negative[24]. The estimate indicates that a majority of people does not like to share at all. While our model is simplistic, this finding generally accords with our and others' experimental findings that altruism towards strangers is widespread but far from universal. This baseline estimate implies that 31% of people would share a positive amount in a completely anonymous, dictator game with zero social pressure, which is not far from the observed rates of giving in experiments that have attempted to create this kind of setting (Koch and Normann (2008), for example).

Table 2.3: Structural Model Comparison

		Variable Extr. Mot.	Constant Extr. Mot.	No Extr. Mot.
Dictator Game	μ_α	-1.732 (0.486)	-1.643 (0.244)	-0.294 (0.452)
	σ_α	3.569 (0.757)	3.395 (0.347)	6.031 (0.796)
	μ_β	2.560 (0.489)	2.435 (0.193)	
	σ_β	3.159 (0.725)	3.372 (0.239)	
Negative Reciprocity	μ_α	-5.723 (2.073)	-7.561 (0.453)	-3.938 (1.064)
	σ_α	5.789 (2.124)	6.922 (0.695)	6.920 (1.423)
	μ_β	1.010 (0.657)	2.435 (0.193)	
	σ_β	1.627 (1.071)	3.372 (0.239)	
Positive Reciprocity	μ_α	0.153 (0.140)	0.123 (0.211)	1.389 (0.352)
	σ_α	1.247 (1.005)	1.961 (0.276)	5.220 (0.557)
	μ_β	2.893 (0.450)	2.435 (0.193)	
	σ_β	3.485 (0.489)	3.372 (0.239)	
Weighted SSE		288.671	291.839	380.243

Notes: GMM estimation results for baseline model specification (33 moments, α and β normally distributed) and comparison models requiring that social pressure be invariant to reciprocity, or zero social pressure. α refers to the weight on internal altruism, and β the weight on external social image.

What is striking is the magnitude of β in the baseline DG. With $\mu_\beta = 2.56$ and $\sigma_\beta = 3.159$, fully 73% of people feel pressure to share, dwarfing the fraction who truly want to share, and the magnitudes of these estimates are significantly greater than zero. This adds to the

[24] Since we do not investigate outright spite, we can't distinguish this result from an alternative specification in which altruism is censored at zero; the key prediction is that a majority of people do not put positive weight on others' outcomes.

mounting evidence that non-reciprocal (one-sided) giving is perhaps more driven by external considerations such as social image than by internal social preferences.

The picture changes with the introduciton of reciprocity is involved. When the recipient has previously behaved selfishly (NR treatment), internal altruism parameters plummet to the point where only 16% of second-stage dictators feel any drive to share. At the same time, the weight of external factors seems to drop. The decrease might reflect reduced pressure to share, or increased pressure (or at least a license) to punish the recipient by answering selfishness with selfishness.[25]

In the positive reciprocity (PR) environment, however, the weight of external factors returns to a very similar levels as in the baseline game. At the same time, the distribution of altruism moves upwards and a much smaller amount of the mass falls below zero. A majority, 52%, of people are predicted to return a favor to a kind partner, even if the partner would never find out.

In the Appendix C.3, we report a series of robustness checks for the above results. The results are qualitatively unchanged.

Next, we contrast our estimates with those based on models that does not allow the external motivation to depend on the reciprocity environment ("Constant extrinsic motivation") and with estimates based on standard reciprocity models ("No extrinsic motivation"). Column 2 of Table 2.3 shows the estimates when external motivation is held constant across reciprocity environments, and column 3 shows the estimates when external motivation is required to be 0. As such, column 3 illustrates how well previous outcome-, type-, and intentions-based models of reciprocity fit the data.

In this estimation, to break ties, we assume that anyone indifferent between sorting out and sorting in but sharing nothing chooses to sort out. Within our general model assumption (indifferent non-sharers sort out with some fixed probability), this specification fits the data best and creates the most conservative comparison to the model with external factors.

Nonetheless, a model that does not allow for social pressure or other external factors does significantly worse at fitting the data—as shown in column 3, the weighted SSE is much higher. The estimates are also rather implausible when applied to games with no social pressure. They imply that about half of all people will share in the DG (with sorting), and still 28% in a negative reciprocity environment (with sorting).

A model that allows for external motives but restricts them to be reciprocity-invariant, instead, yields estimates that are rather similar to those we obtain when we allow external motivation to vary. The estimates for the constant-beta model, shown in column 2, are rather close to those in column 1, both in the neutral (DG) and in the positive reciprocity settings. The estimates for the negative reciprocity case differ a bit more, with internal motivation being estimated to be even more negative ($mu_\alpha = -7.561$ rather than $mu_\alpha = -5.723$) in

[25] External determinants, such as social pressure and social norms, are usually thought of as attributes of a situation that apply equally to everyone; but since individuals might have different beliefs about how their actions are judged in this case, there is room for a mixed interpretation in which social pressure is reduced for some and inverted for others.

order to counterbalance the constant and hence relatively high β. At the same time, standard errors are about halved, reflecting the gain in power from estimating fewer parameters.

Overall, the table reveals that allowing reciprocity to influence not only internal but also external factors does not significantly improve the fit of the model, relative to assuming reciprocity-invariant external factors. This result parallels the reduced-form results that the impact of sorting on giving is approximately invariant across reciprocity environments.

We conclude that while reciprocity may slightly influence external motivations to share, it primarily acts through the channel of internal motives. Incorporating reciprocity-invariant social pressure into theories of reciprocity greatly improves their predictive power, but an additional interaction between external factors and reciprocity has a relatively small impact.

At the same time, the estimation results also show that the amount of "additional" internal motivation—in the sense of "reciprocity-induced" altruism or similar heightened internal motives under positive reciprocity—is estimated approximately correctly under any of the models, including the naive model that does not allow for social pressure or other extrinsic motives. Specifically, if one asks how much "additional" internal willingness to share is induced by a kind treatment of the recipient, the naive model implies that the average α_{NR} exceeds the average α_{DG} by 1.683. Under the models that allow for external motivation the corresponding differences between the μ_α's equal 1.885 (variable beta) and 1.766 (constant beta), and are hence quite similar. This robustness reflects the fact that external motives are remarkably stable across both environments.

This conclusion is somewhat less true for the negative-reciprocity environment: The decrease in internal willingness to share after unkind treatment, relative to a neutral treatment, is slightly underestimated when neglecting external factors (the difference in μ_α's amounts to -0.347) and is strongly overestimated when allowing only for a reciprocity-insensitive external motive (with a difference of 1.927), relative to the model with reciprocity-dependent betas. In other words, failure to account for extrinsic motives and their context-dependence is particularly detrimental when estimating the motives for and extent of negative-reciprocal behavior.

2.6 Conclusion

This paper questions whether reciprocal behavior primarily reflects internal motivations to share, such as altruism or fairness. Recent research highlights the importance of incorporating external factors, such as a concern for social image, into models of pro-social behavior. However, this recent research has, so far, had little impact on how economists view and model reciprocity.

To address this potential gap in the literature, we use a novel set of experimental conditions to show that reciprocal behavior responds to "avoidance opportunities" in ways unaccounted for by existing models of reciprocity. The experimental conditions compare sharing behavior in social environments where positive (or negative) reciprocity is in play, while varying whether the potential giver has the option to exit and avoid the sharing decision. We find that, regardless of the nature of the reciprocity environment, introducing an avoidance option causes generosity to drop significantly. The effect is comparable in size to the drop in a no-reciprocity (dictator game) environment. We also find that the rate of spitefully sorting in without sharing anything is decreasing in the inferred kindness of the recipient, i.e., out of proportion to the general rate of sharing zero. Traditional models of internally motivated reciprocity—whether outcome-, type-, or intentions-based—are not able to account for these findings. That is, regardless of whether the choice context involves reciprocity or simple, one-sided sharing, external motivations are required to explain behavior.

We also show how one might incorporate external determinants (e.g., social pressure) into a model that is in the spirit of previous reciprocity models, applicable to our experimental setting. By estimating this model, we demonstrate that even this simple extension of models of reciprocity is substantially better able to explain our results: Allowing for non-zero external motivations strongly improves the predictive power of the model. We also find, however, that reciprocity primarily changes behavior by changing people's internal motivation to share, consistent with existing models.

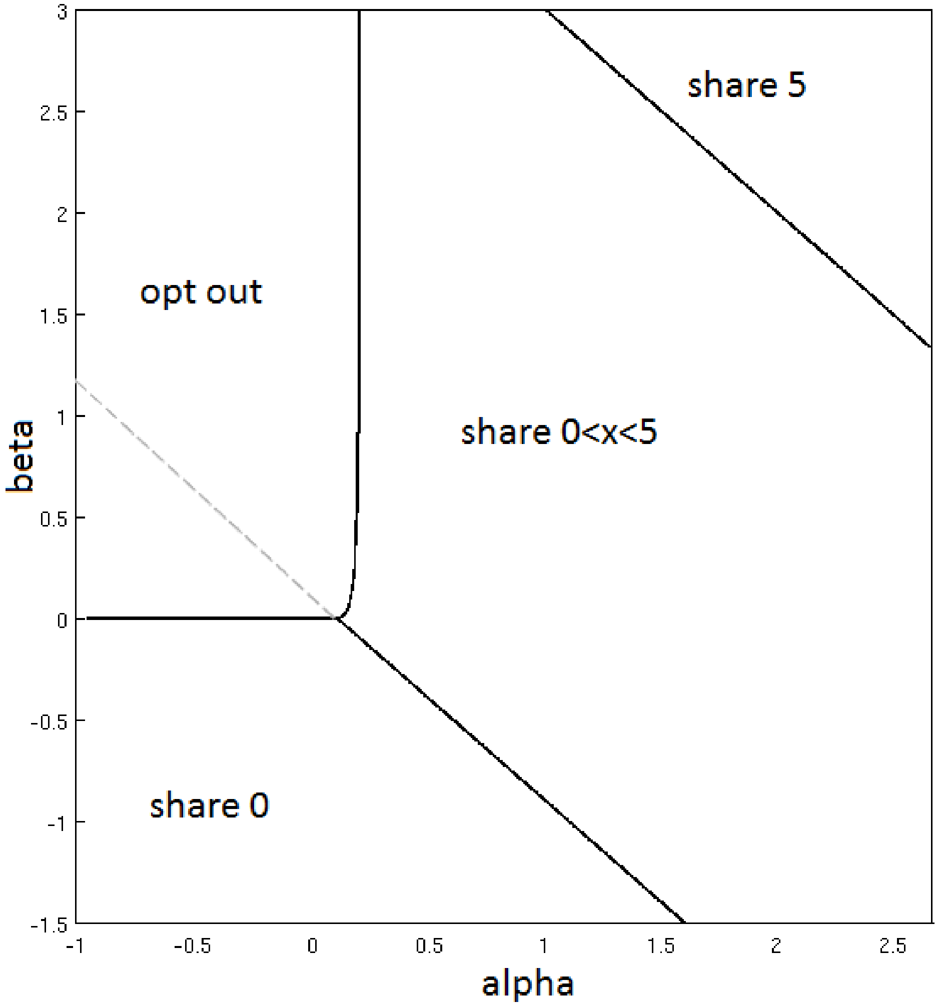
As a next step in this research agenda, it is valuable to conduct further experimental treatments along the lines of the ones we use here, to improve our understanding of reciprocity. For example, further variations on standard reciprocity games, such as ultimatum games, that introduce features similar to our sorting manipulation, could help identify preferences in a clean way and avoid mis-attributing behavior to easily confounded motivations. Weele et al. (2014) provide a valuable example of this kind of work, they reach different conclusions than we do here regarding the role of external motives in reciprocal behavior. Hence, further research is highly valuable for understanding the degree to which reciprocity—like one-sided acts of generosity—is partly motivated by external factors.

Similarly, field experiments are valuable for exploring whether reciprocity in real-world economic settings, such as contracts and labor markets, may in fact reflect external motives. For example, do above-market wages induce more effort due to internal motives, such as increased altruism towards the employer, or might they also reflect social pressure to work hard when paid well?

As another important and bigger step, it is worthwhile to pin down the type of preferences underlying our general term "external motivation." For example, to what extent are external determinants of sharing social-image concerns, and to what extent do they reflect direct social pressure? And, in the case of social image, is reciprocal giving welfare reducing, or is there an image reward for reciprocity? Or do such motivations reflect self-image concerns, which can act similarly to social image motivations (as in, e.g., Bénabou and Tirole (2011) or Rabin (1995)), but might have different welfare implications? Moreover, if there is a confluence of factors, what is the net effect? DellaVigna, List, and Malmendier (2012) find that sharing decisions under social pressure are welfare-reducing, but it is possible that reciprocity of either variety counteracts this effect. It also remains to be discovered what factors trigger external motivation.

Finally, we suggest that even the interpretation of the internal determinant of reciprocal giving might need to be re-thought. Rather than reflecting concern about the intentions or type of the other person, the “internal” element could reflect a sense of obligation in the sense of Cialdini (1993) and as proposed in the anthropology and sociology literature. This distinction is particularly relevant for our economic understanding of the role of reciprocity in markets. When a firm gives gifts with the intention to maximize profits, by inducing reciprocity among customers, will consumers respond even if the firm’s intentions are clear? The literature addressing this question remains scarce and would benefit from further theoretical and empirical research.

Figure 2.2: Parameterization of Reciprocity Model



Chapter 3

You've Earned It

Combining Field and Lab Experiments to Estimate the Impact of Human Capital on Social Preferences

3.1 Introduction

Social scientists have long sought to disentangle the relationship between formal education, cultural modernization, and economic development. In the African context, sociologists have argued that "Western" education is associated with the adoption of "modern" values including "independence from family and other traditional authority, belief in science and in man's ability to control his fate, and orientation toward the future" (Armer and Youtz 1976, p. 605). Inkeles (1969) constructs an index of individual modernity which aggregates independence from traditional sources of authority, openness to new experiences, belief in science and modern medicine, ambition, punctuality, and civic participation; he finds that educational attainment is the single most powerful predictor of a modern orientation in all six countries he studies.[1] More recently, Barro (1996) has shown that female education is the strongest long-term predictor of democracy, while Mattes and Bratton (2007, p. 199) claim that education builds support for democratic institutions by "diffusing values of freedom, equality, and competition throughout the population." Glaeser et al. (2004) argue that human capital gains are critical drivers of institutional change. Whether schooling *causes*

[1] See also Inkeles and Smith (1974). More generally, Easterlin (1981) argues that the introduction of mass primary education has preceded industrialization in most developed economies. Goldin and Katz (2008) trace out how the expansion of public education contributed to the economic and social transformation of U.S. society.

such changes in cultural values is an open question; it is also possible that those with an innately modern outlook choose to obtain more schooling, and the observed correlations result from sample selection. More broadly, though researchers have identified a robust correlation between modern cultural values and industrialization (Inglehart and Baker 2000), the mechanisms through which such changes occur remain obscure.

In this paper, we provide evidence that academic achievement alters individual values, specifically social preferences governing the appropriation of others' income, as captured in an economic experiment. Our novel research design combines a randomized evaluation specifically, the introduction of a scholarship program for girls in a random sample of Kenyan primary schools with a lab experiment designed to measure respect for earned property rights. From a methodological perspective, ours is among the first studies to use lab experimental methods to measure the impacts of a development intervention.[2] We argue that this setting provides cleaner identification of the link between education and social preferences than has previously been possible.

In 2001, the Dutch NGO ICS Africa introduced a scholarship competition for sixth grade girls (called the Girls Scholarship Program or GSP) in a random sample of primary schools in Busia District, in western Kenya; the program led to improvements of 0.2 to 0.3 standard deviations on standardized academic tests, relative to schools in the control group (Kremer, Miguel, and Thornton 2009). Our experimental subject pool comprises girls from the treatment and control schools in the scholarship program. The design allows us to identify the causal impact of academic achievement on social preferences using an instrumental variables approach, since assignment to a school in the scholarship program (treatment group) is unrelated to baseline characteristics such as cognitive ability and family background that might themselves affect social preferences.[3]

We measure the impact of academic achievement on social preferences in an experimental lab setting which allows us to turn off strategic considerations such as the fear of social sanctions. Economic experiments are a widely used tool for measuring cross-cultural differences in values, norms, and beliefs that are difficult to capture in survey data. In particular, dictator, ultimatum, and trust games have been conducted on every inhabited continent, with subject populations ranging from university students in the United States to hunter-gatherers in Tanzania (Henrich et al. 2004; Roth, Prasnikar, and Okuno-Fujiwara 1991, cf.).[4] Dictator games — in which one player (the "dictator") is provisionally allocated an amount of money, and decides how to divide it between *self* and another subject, *other* — mea-

[2] **Barr2012** use public goods games to measure the impact of a school committee monitoring intervention in Uganda; while Fearon, Humphreys, and Weinstein (2009) use similar experiments to measure the impact of a post-conflict community development initiative in Liberia. Paluck and Green (2009) demonstrate that randomized experiments can be used to demonstrate the efficacy of policies explicitly intended to change cultural norms.

[3] Friedman et al. (2011) use a similar identification strategy to explore the impact of the GSP on political attitudes, knowledge, and behavior.

[4] See Henrich, Heine, and Norenzayan (2010) for an overview of the ways in which subjects in western university experimental labs are not representative of humanity in general.

sure the willingness to share in non-strategic settings, and have been used to measure the strength of egalitarian (or other) ideals underlying perceptions of what constitutes a "fair" distribution of income (Barr et al. 2009; Cappelen et al. 2007; Forsythe et al. 1994, cf.).

We employ a variant of the dictator game designed to measure preferences governing the distribution of earned income — specifically, the willingness to appropriate *other*'s earnings. Hoffman et al. (1994) first used earned, rather than windfall, income in dictator games to generate an informal "property right"; they find that enhancing dictators' sense of entitlement via the earnings manipulation decreases generosity.[5] In contrast, our design increases the extent to which the *other* has property rights over the budget: dictators in our experiment decide how to divide money that *other* was paid for completing a real effort task. Thus, our design intentionally separates the right to determine the final allocation — i.e. control rights, which Grossman and Perry (1986) define as property rights — from the "natural" but informal property rights proposed by Locke (1690), which result from generating something through one's own labor.[6] Our specific design measures generosity toward those who have increased social surplus through their own effort.[7] The experiment was first proposed by Jakiela (2009), who reports that more educated Kenyan adults are significantly more generous than the rest of the population when deciding how to divide income earned by others, though not in other situations. The novel research design in the current paper, exploiting the random assignment of schools to the GSP treatment and control groups, allows us to determine whether this association is driven by the causal impacts of schooling on social preferences and beliefs about hard work.

We find that subjects drawn from the GSP treatment group have higher levels of academic performance (measured in terms of the primary school exit exam), and that they allocate significantly more to other in our modified dictator game. Point estimates suggest that a one standard deviation increase in academic test scores is associated with a 10 percentage point increase in the share of the budget allocated to other. Using data on subjects expectations about the amount that dictators were likely to allocate to them, we show that our results are not driven by changes in beliefs: subjects drawn from the GSP treatment group do not expect that dictators will allocate them more. Hence, our findings can not be explained by changes in the beliefs of individuals holding identical (reciprocal) social preferences. We also report suggestive evidence from pilot experiments that girls in the GSP treatment group do not allocate more than control girls in a standard dictator game (involving unearned income). This suggests that academic success impacts the respect for earned property rights but not generalized altruism, a finding which is consistent

[5] Cherry (2001), Cherry, Frykblom, and Shogren (2002), and List and Cherry (2008) conduct similar earnings treatments. **FahrIrlenbusch00** Konow (2000), and Cappelen et al. (2007) also explore distributional preferences governing earned income.

[6] Building on Locke (1690), Gintis (2007) models "preinstitutional" property rights as the equilibrium result of the interaction between the endowment effect and possession. Following **FahrIrlenbusch00** we refer to the entitlement effect generated by our design as an "earned property right."

[7] The design is quite similar to a trust game involving real effort rather than investment, except that receivers can only generate payoffs for themselves by "trusting" their labor income to the dictator.

with Jakiela (2009).

Our findings relate to recent evidence suggesting that the level of allocation to *other* observed in dictator games is strongly associated with the extent of market integration within a community (Henrich, Heine, and Norenzayan 2010; Henrich et al. 2004), though the underlying causal mechanism is not well understood. At the individual level, Almå s et al. (2010) report that the tendency to reward others for hard work emerges during adolescence among Norwegian subjects: fifth graders participating in a dictator game preceded by a period of team production tended to favor egalitarian allocations, while older subjects were more inclined to base their allocation decisions on relative contributions to total output. Both Henrich, Heine, and Norenzayan (2010) and Almå s et al. (2010) suggest that the fairness norms invoked in dictator games are not innate, but emerge over time through cognitive development and socialization. However, neither is able to identify a causal mechanism to explain how and why disparate cultural norms of fairness emerge where and when they do.

The project is also related to recent studies exploiting natural experiments to show how cultural values and norms evolve. Di Tella, Galiani, and Schargrodsky (2007) demonstrate that the acquisition of formal land titles by squatters leads to the adoption of more market-oriented beliefs. Employing a methodology similar to ours, Fisman, Kariv, and Markovits (2009) combine a lab experiment with a natural experiment to show that random assignment of Yale law students to first year instructors trained in economics, rather than in law or humanities fields, leads to the adoption of distributional preferences which are both more selfish and more concerned with efficiency. In highlighting the extent to which life experiences shape individual preferences regarding property rights, our results are broadly consistent with both studies.

3.2 Experimental Design

Primary Education in Kenya

Since 1985, Kenya has had an educational system involving 8 years of primary school (standards 1 through 8) and 4 years of secondary school (forms 1 through 4). Admission to secondary school is contingent on the successful completion of a government exit exam, the Kenya Certificate of Primary Education (KCPE), at the end of standard 8. The KCPE is the equivalent of a primary school diploma, and the vast majority of students who complete standard 8 take the KCPE exam, whether or not they intend to continue on to secondary school.

Like many African countries, Kenya experienced large increases in educational attainment in the post-independence period. Between 1970 and the present, the adult literacy rate increased from 32 percent to 87 percent (UNDP 1993, 2013). Kenya instituted a policy of free primary education in 2003, and the gross primary enrollment ratio is now above 100 percent.[8] However, grade repetition is common, and more than a quarter of those who start

[8] Prior to the introduction of free primary education, the gross primary enrollment rate was approxi-

primary school drop out before the end of standard 8 (UNDP 2013). Women have tended to lag behind men, particularly at higher levels of education: only 25 percent of Kenyan women over 25 completed secondary school, as compared with 52 percent of men (UNDP 2013). Since Kenyan children typically enter primary school at age 6 or 7 and frequently repeat grades, women are nearing marriageable age by the end of primary school; it is at this point that gender disparities in education begin to emerge.

Prior to the introduction of free primary education, parents of children in primary school had to pay school fees which averaged about 6.40 USD per year Kremer, Miguel, and Thornton 2009. The revenue raised from school fees was used to pay for a range of educational inputs for example, classroom maintenance and school supplies which were not covered by the central government. These fees discouraged those not planning to attend secondary school from remaining in primary school and completing the KCPE exam.

The Girls' Scholarship Program (GSP)

The Girls Scholarship Program (GSP) was an education initiative targeting adolescent girls who were enrolled in primary schools near Busia, Kenya, in 2000. The GSP was implemented by the Dutch NGO International Christian Support Africa (ICS) in 34 primary schools in Busia District. The aim of the program was twofold: to improve girls academic performance by incentivizing effort, and to encourage girls to remain in school by defraying the costs (for those who won the scholarships). To that end, ICS organized a scholarship competition for girls enrolled in standard 6 in participating schools.

The program took advantage of the fact that most children in Kenyan primary schools take practice KCPE exams at the end of standards 4 through 7.[9] Like the KCPE, the practice exams are proctored by representatives of the District Education Offices (rather than the teachers themselves), and it is consequently very difficult to cheat. The GSP offered girls in program schools a performance incentive: in each year of the program, ICS awarded scholarships to all girls who scored in the top 15 percent of females in standard 6 in Busia District on the KCPE practice exam. For the two years after they won the competition, scholarship recipients were given an annual cash grant of approximately 12.80 USD (1000 Kenyan shillings) and had their school fees paid, for a total award amount of approximately 38 USD per winner. Thus, the total amount of the award package was large relative to the income of the typical Kenyan household, which averaged about 400 USD at the time of the intervention. Winners were also recognized at a public awards ceremony. ICS administered the competition in both 2001 and 2002, so two cohorts of girls received awards.

In order to assess the overall impact of the GSP, ICS conducted a randomized evaluation of the program. 69 primary schools in Busia District were randomly assigned to either the GSP treatment group or a control group which did not participate in the scholarship

mately 90 percent. See Lucas and Mbiti (2012) for an extended discussion of the abolition of school fees in Kenya.

[9] The practice exams are not required, and students must pay a fee of between 1 and 2 USD to participate in each practice exam.

competition.[10] The program was announced in treatment schools in March of 2001, at which point school headmasters were asked to pass information about the GSP competition on to the parents of eligible girls.[11] To make sure that parents of children in GSP treatment schools were fully informed about the program, ICS also organized community meetings in September and October of 2001. A first cohort of program participants took practice KCPE exams in November of 2001, and scholarships were subsequently awarded to 110 girls. A second cohort of girls participated in the program the following year.

Kremer, Miguel, and Thornton (2009) discuss the impacts of the GSP intervention. In the year that they were eligible for the scholarship, girls in GSP treatment schools had practice exam scores that were 0.27 standard deviations higher than those in control schools. Though only girls scoring near the top of the distribution were eligible for scholarships, the GSP program led to test score improvements at all performance levels, and among boys (who were not eligible for scholarships). When program impacts are disaggregated by baseline test score (for the sub-sample of girls for whom baseline test scores are available), the results suggest that test scores increased by at least 0.19 standard deviations for the top three baseline test score quartiles, even though only 5 percent of girls in the next-to-lowest quartile of baseline test scores ended up winning a scholarship (Kremer, Miguel, and Thornton 2009). Kremer, Miguel, and Thornton (2009) also report that the program led to a 0.10 standard deviation increase in test scores among sixth grade boys in treatment schools, and to increases in teacher attendance, which may partially explain the apparent spillover effects.

Data Sources

We combine our experimental data with information from three additional sources. The first is administrative data on individual test scores in 2000 (prior to the intervention), 2001, and 2002. Because students have to pay a fee (approximately 1 to 2 USD) to take the KCPE practice exams, not all enrolled students participate. In 2001, for example, approximately 78 percent of girls in standard 6 in the control schools chose to take the practice exam. Test score data is available for the majority of students in GSP treatment and control schools.

Student surveys, which were administered in treatment and control schools in 2002, constitute a second source of data on our subjects. Because of financial constraints, only a limited amount of individual-level data was collected at the time of the intervention. Baseline data on individual characteristics (e.g. parents names and education levels) was collected

[10] A parallel randomized experiment was simultaneously conducted in neighboring Teso district (Kremer, Miguel, and Thornton 2009), but since it is unclear whether the scholarship increased human capital in this district, in part due to program implementation difficulties there, follow-up surveys were only conducted in the Busia district. For that reason, we only have actual KCPE scores for Busia students, and we focus only on that experiment in this paper.

[11] Only those girls who were enrolled in standards 5 and 6 in treatment schools in January 2001 were eligible for scholarships. This restriction was imposed to avoid creating incentives for girls to transfer from control to treatment schools.

during school visits in early 2002, but only those students who were present in class on the day of the survey could be interviewed.

Finally, between 2005 and 2008, an extensive follow-up survey was administered to 1,862 women from both treatment and control schools all girls in the GSP cohorts who could be located at the time of the follow-up survey. The effective tracking rate is 80 percent, and attrition from the survey does not differ substantially between the GSP treatment and control groups (Friedman et al. 2011). This follow-up survey provides information about educational attainment after the GSP competition, including self-reported KCPE scores for those who took the exam.[12]

Experimental Subjects

To estimate the impact of the GSP intervention on individual social preferences, it was necessary to recruit experimental subjects who were enrolled in standards 5 and 6 in the GSP treatment and control schools in 2001. This presented two challenges. First, eligible young women, many of whom had moved out of their family homes to marry or continue their schooling, had to be located and contacted. Second, they needed to be brought together to conduct our lab-in-the field experimental sessions.

Analysis of data from the GSP follow-up survey indicates that the program did not increase the probability of migrating out of Busia District (Friedman et al. 2011), so we felt that it was reasonable to focus on recruiting those individuals still residing there. Members of the research team met with local officials throughout the district to compile a list of such potential participants. We conducted our experimental sessions during the August break in the academic year so that girls who were in boarding school much of the time would be able to participate (while they were home visiting their families). The list was organized by sublocation, the second most disaggregated level of local government in Kenya. We then identified clusters of sublocations which contained enough girls from the GSP treatment and control groups to warrant organizing an experimental session. Once experimental sessions were scheduled and target participants identified, the same local officials were tasked with delivering invitation letters to each of the girls explaining the project and inviting them to attend a specific experimental session.

We expected the GSP intervention to impact academic performance and other educational outcomes directly, and to influence preferences and values primarily through the education channel. It is therefore important to focus on a population for whom comparable education-related outcomes are available for the treatment and the control group. Because more than half of the control group was still in school at the time of the GSP follow-up survey (and estimates of the programs impact on educational attainment were consequently

[12] Administrative data on past KCPE scores is not publicly available from any centralized source. Because many girls took the KCPE exam in different years and might have changed schools, it was not feasible to collect hard copy records of KCPE scores for all the schools that GSP respondents might have attended.

biased toward zero), we chose to focus on the KCPE score.[13] As discussed above, KCPE scores provide a measure of academic success for all those who complete primary school; moreover, the GSP follow-up survey does not suggest that girls in treatment schools were more likely to take the KCPE exam. Performance on the KCPE is a particularly salient measure of academic success, since it determines whether or not a student will be admit- ted to a government secondary school.[14] Because KCPE scores are such an important determinant of future academic success, it is not uncommon for Kenyan students to repeat standard 8 in order to retake the test. 14 of our 101 subjects report taking the KCPE exam twice. To avoid conflating academic performance with the likelihood of success, we focus on the first reported KCPE score. The majority of those in our sample (93 percent) took the KCPE between 2003 (the first year that a girl who was in standard 6 in 2001 could be eligible) and 2005.[15]

The 101 young women in our sample (45 treatment vs. 56 control) were enrolled in 23 different schools in 2001, 10 treatment and 13 control. Thus, each session contained relatively small numbers of girls from the same primary school, though, since all subjects were from Busia District, they could easily be socially connected to girls who attended different primary schools. Subjects ranged in age from 17 to 23. 71 percent of them were still in school at the time of the experiment, while 12 of them were married. Subjects in the control group had completed an average of 8.3 years of schooling, while those in the treatment group had completed 8.6 years (though this difference is not significant).

Though they are not a random sample of girls in the GSP, our subjects are broadly representative of GSP Survey respondents who took the KCPE exam. Table 3.1 compares the two groups. Our subjects are similar to other GSP respondents who took the KCPE in terms of educational attainment, KCPE score, cognitive ability, English and Swahili vocabulary, household size, parents education, and work experience. Our subjects are somewhat less likely to come from a GSP treatment school, though the difference is only marginally significant (p-value 0.056). They are also slightly (approximately 3 months) younger, but again this difference is only marginally significant (p-value 0.1). Thus, we expect that our findings would generalize to the population of GSP respondents who took the KCPE exam.

Table 3.2 compares the GSP treatment and control groups within our sample in terms of baseline (pre-GSP) characteristics. Those in the GSP treatment group are not significantly different from the control group in terms of age or parents education. There is a small

[13] In our sample, GSP treatment is associated with an 8.3 percentage point increase in the likelihood of being in school, but the effect is not significant. This estimated impact is very similar to the 7.9 percentage point effect reported in Friedman et al. (2011).

[14] Ozier (2010) reports that scoring above the mean on the KCPE increases the probability of completing secondary school by 20 percentage points. Test scores are arguably more relevant as an indicator of quality, rather than quantity, of education: Barro (2001) and Hanushek and Kimko (2000) both find that test scores on internationally comparable exams are more predictive of future income growth rates than years of schooling.

[15] Unfortunately, Kenyan secondary schools do not conduct regular standardized tests that could be used to provide a more recent measure of academic achievement. Because those subjects attend secondary schools which vary in quality, grade point averages and class ranks would not be comparable.

Table 3.1: Summary Statistics: Subjects vs. Rest of GSP Sample

Lab Experimental Subjects? $(S = 0, 1)$	$S = 0$	$S = 1$	Difference
N	1024	101	
First KCPE score (among those who took exam)	258.276	259.604	-1.328
	(1.392)	(4.430)	(4.643)
Change in test scores during GSP	-0.011	-0.001	-0.011
	(0.026)	(0.076)	(0.081)
Highest grade completed	8.602	8.426	0.176
	(0.028)	(0.127)	(0.130)
Age	20.161	19.901	0.260*
	(0.045)	(0.145)	(0.152)
Ravens matrices score	20.727	21.538	-0.810
	(0.169)	(0.622)	(0.644)
English vocabulary score	9.939	10.089	-0.151
	(0.080)	(0.245)	(0.258)
Swahili vocabulary score	9.478	9.812	-0.334
	(0.081)	(0.254)	(0.258)
Respondent held job in last 12 months	1.881	1.871	0.010
	(0.010)	(0.033)	(0.035)
GSP Treatment Group	0.546	0.446	0.100*
	(0.016)	(0.050)	(0.052)
Father's education	9.786	10.420	-0.634
	(0.133)	(0.395)	(0.417)
Mother's education	7.301	7.263	0.038
	(0.132)	(0.415)	(0.435)
Household size	6.951	6.812	0.139
	(0.088)	(0.283)	(0.297)
Household Assets (1000s of KSh)	27.727	30.095	-2.369
	(0.545)	(1.718)	(1.802)

Note: standard deviations in parentheses in columns 1 and 2, and standard errors in parentheses in column 3. *** indicates significance at the 99 percent level; ** indicates significance at the 95 percent level; and * indicates significance at the 90 percent level. The number of observations contributing to each number may differ from the pool sizes shown when particular variables are unavailable for some people.

but insignificant different in baseline practice test scores (for those subjects who took the practice KCPE in 2000, prior to the GSP intervention). Given the randomized design and the absence of differences between the treatment and control groups at baseline, it is reasonable to attribute differences in behavior within the experiment to the impact of the GSP program, and the gains in academic performance it generated, on individual social preferences.

Table 3.2: Summary Statistics: GSP Treatment vs. Control

GSP Treatment Group? $(T = 0, 1)$	Both	$T = 0$	$T = 1$	Difference
N	101	56	45	
Age	19.901	19.696	20.156	0.459
	(0.145)	(0.185)	(0.227)	(0.293)
Baseline father's education	11.631	11.469	11.788	0.319
	(0.404)	(0.596)	(0.555)	(0.814)
Baseline mother's education	9.574	9.733	9.419	-0.314
	(0.487)	(0.733)	(0.655)	(0.984)
Baseline practice KCPE score	0.077	-0.003	0.219	0.223
	(0.098)	(0.117)	(0.175)	(0.210)

Note: standard deviations in parentheses in columns 1, 2 and 3, and standard errors in parentheses in column 4. *** indicates significance at the 99 percent level; ** indicates significance at the 95 percent level; and * indicates significance at the 90 percent level. The number of observations contributing to each number may differ from the subject pool sizes shown when particular variables are unavailable for some people. Data on father's education, mother's education, and baseline (2000) KCPE practice test score is available for (respectively) 65, 61, and 64 subjects.

Experiment Design and Procedures

Our experiment is a modified dictator game designed to to measure respect for the earned property rights of others (Fahr and Irlenbusch 2000). As in all dictator games, one subject (the dictator) divides a budget between *self* and an anonymous *other*, another subject attending the same experimental session (Camerer 2003; Forsythe et al. 1994; Kahneman, Knetsch, and Thaler 1986). Our variant is a real effort dictator game in which each subject divides money that was earned by *other*.[16]

Our study is motivated by previous evidence suggesting a link between educational attainment and social preferences, particularly respect for earned property rights. Jakiela (2009) conducts four different versions modified dictator game treatments in Kenyan villages. In her experiments, the dictator divides either her own or *other*s earned or unearned

[16] Our design is identical to that used in Jakiela (2009), which was motivated by Ruffle (1998) and Greig (2010). Hoffman et al. (1994), Cherry (2001), Cherry, Frykblom, and Shogren (2002), and List and Cherry (2008) conduct dictator games in which subjects divide their own earned income between *self* and *other*; they find that the amount allocated to *other* is lower when the dictators endowment is earned. Bardsley (2007), List (2007), and Fisman, Jakiela, and Kariv (2013) conduct modified dictator games which allow for both giving and taking.

income between *self* and *other*. She finds that villagers with more than a primary school education allocate more to *other* than less educated subjects in one of her four experimental treatment, the one in which subjects divide income earned by other. Thus, more educated subjects appear more inclined to respect the earned property rights others, but not more altruistic or generous overall. That result motivates the present study.

We replicate the experimental treatment in which Jakiela (2009) finds an association between education and allocation decisions: dictators divide money earned by *other* between *self* and *other*.[17] In our experiment, each subject was matched with an anonymous other who was seated in another room, and whose identity was not revealed during or after the experimental session. Subjects first learned about the structure of the experiment, and then about the nature of the real effort task (which determined earnings). We selected an activity which could be easily understood by all subjects, regardless of educational attainment, and which would allow players to increase their output by exerting greater effort up to some maximum feasible level: subjects earned money by clicking a handheld tally counter, and were paid based on the number of times they clicked within ten minutes.[18] Subjects were given a two-minute practice period during which they tried out the real effort task before they made their allocation decisions. After the practice period, subjects decided how they wished to divide *other*s earnings between *self* and *other*. We used the strategy method: for each of the 20 possible earnings levels, subjects recorded the allocation that they wished to implement by circling the amount (presented as images of Kenyan currency) that they wished to allocate to self. We chose this pictorial approach to choice elicitation so that subjects who were relatively uncomfortable with entering numbers into tables could record their own allocation decisions. After individual decisions were recorded, subjects performed the real effort task for ten minutes, and were informed how much money they had earned (based on the piece rate and their level of production); they earned 30 Kenyan shillings (approximately $0.375) for every 200 times they clicked the tally counter.[19] These activities took place in parallel in the two separate rooms. At the end of the experiment, one room was chosen at random, and the decisions of dictators in that room were combined with the earnings information about the matched subjects in the other room to determine final payoffs.[20]

[17] We also piloted the 3 other variants of the dictator game proposed in Jakiela (2009). However, we did not locate large enough numbers of potential participants to be able to carry out all 4 treatments. (Each session lasted approximately 3 hours, and each subject participated in only one treatment.) We chose to focus on the treatment described here because it is in that treatment that Jakiela (2009) finds an association between education and allocation decisions. In any potential analysis of the pilot data from the other 3 treatments, we face a weak instrument problem in the first stage regression because of the limited sample size.

[18] We opted for a non-cognitive task so that output would reveal minimal information about education or innate intelligence. The task was inspired by Ariely, Bracha, and Meier (2009), but adapted to a non-computerized environment. Other non-cognitive tasks which have been used in experimental settings include stuffing envelopes (Falk and Ichino 2006; Konow 2000) and cracking walnuts (Fahr and Irlenbusch 2000).

[19] Interestingly, Jakiela (2009) finds no evidence that subjects exert less effort when they expect that another may appropriate a portion of their earnings.

[20] Thus, all subjects make allocation decisions which might determine final payoffs this was necessary

Complete experimental instructions, which were presented orally during the sessions, are included in Appendix D.

We conducted 4 experimental sessions in August of 2008, each of which was held at a different primary school in Busia District. August is a school vacation in Kenya, and empty primary school classrooms provide a sheltered location for conducting experiments. Primary schools are also easy for subjects to locate because they are well-known within the community. Because most schools in the area have one or two classrooms per grade level, it is also feasible to split subjects into separate rooms. Experimental sessions took approximately 3 hours. Final payouts averaged 1.80 USD (144 Kenyan shillings) plus a 0.25 USD (20 shilling) show-up fee.

3.3 Results

The main sample includes data from 101 subjects, each of whom made allocation decisions over all twenty potential budget sets. On average, subjects allocated 67.1 percent of the budget to *self* and 32.9 percent to *other* (Table 3.3). Thus, our subjects allocate more to *other* than is typical in dictator games involving students (Camerer 2003), though not more than has been previously observed in African populations (Henrich et al. 2010). The distribution has modes at 0 and 50 percent. 5 percent of subjects allocated the entire budget to *self*, while 13.7 percent split the money evenly and an additional 14.9 percent allocated more than half the money to *other*. Subjects who had some secondary schooling allocated their partners slightly more than those who did not (33.6 versus 31.4 percent of the budget, p-value 0.0226, results not shown). More interestingly, there are clear differences between the GSP treatment and control groups in terms of behavior within the experiment. The two groups are equally likely to allocate the entire budget to *self*, but subjects drawn from the GSP treatment group are substantially more likely to divide the budget evenly (19.2 percent of subjects versus 9.3 percent, p-value < 0.001) or to allocate more than half the budget to *other* (16.8 percent versus 13.3 percent, p-value 0.031).

Our main analysis estimates the causal impact of academic performance on social preferences, as measured by allocation to *other* within the dictator game, using random assignment to the GSP treatment group as an instrument for the KCPE score (Table 3.4). The key outcome variable is the share of the total budget that the dictator allocates to *other*. We first report linear IV specifications (Panel A, Columns 1–3), then reduced form OLS specifications (Panel B, Columns 1–3), and the IV first stage (Panel C, Columns 1–3). The IV estimates indicate that a one standard deviation increase in a student's KCPE score causes

because of our small sample size. In contrast to Andreoni and Miller (2002) and Fisman, Kariv, and Markovits (2007), subjects in our experiment do not receive two sets of tokens (one based on their own decision and one based on the decision of another subject). Instead, each subject within a matched pair makes an allocation decision, and one of the two decisions is randomly chosen to determine payouts, as in Cappelen et al. (2007). The amount of money being allocated is determined by the effort level of the subject whose decision is not chosen to determine payouts.

a large and statistically significant increase in partner share. Without any regression controls, the coefficient on instrumented KCPE score is 10.6, and is significant at the 90 percent confidence level. After adding controls for individual age, ethnicity, and session-room fixed effects, the coefficient remains almost unchanged at to 10.3 and the confidence level increases to 95 percent (Table 3.4, Panel A, Columns 1–3).[21] Compared to an average partner share of 32.9 percent of the budget, this is a large effect. This corresponds to the approximately 6 percentage point average GSP treatment effect shown in the reduced form specifications (Panel B, Columns 1–3).

[21] Age controls include both age in 2008 (normalized) and an indicator for being in the first GSP cohort. Studies by Fehr, Bernhard, and Rockenbach (2008), **AlmasEtAl10** Bekkers (2007), and Fowler (2006) suggest that age is an important predictor of altruistic behaviors. Ethnicity controls are indicators for being a member of a minority ethnic group (Teso or Luo) and for belonging to a minority subgroup of the locally dominant Luhya ethnic group.

Table 3.3: Reduced Form GSP Impacts

GSP Treatment Group? $(T = 0, 1)$	Both	$T = 0$	$T = 1$	Difference
Partner share	32.865	30.029	36.394	6.365***
	(0.462)	(0.606)	(0.695)	(0.922)
Gave nothing	0.050	0.050	0.050	0.000
	(0.005)	(0.007)	(0.007)	(0.010)
Gave exactly half of budget	0.137	0.093	0.192	0.099***
	(0.008)	(0.009)	(0.013)	(0.016)
Gave more than half of budget	0.149	0.133	0.168	0.035**
	(0.008)	(0.010)	(0.012)	(0.016)

Note: standard deviations in parentheses in columns 1, 2 and 3, and standard errors in parentheses in column 4. *** indicates significance at the 99 percent level; ** indicates significance at the 95 percent level; and * indicates significance at the 90 percent level.

Table 3.4: Instrumental Variable Results for Test Scores

	Dependent Variable: Partner Share			*Dependent Variable: Gave Half*		
	(1)	(2)	(3)	(4)	(5)	(6)
PANEL A: IV REGRESSION						
KCPE Score	10.552*	9.290*	10.322**	0.628***	0.623***	0.505**
	(5.910)	(4.940)	(4.732)	(0.211)	(0.195)	(0.230)
Budget	0.028	0.028	0.028	-0.004*	-0.004*	-0.004**
	(0.026)	(0.026)	(0.026)	(0.002)	(0.002)	(0.002)
Constant	31.973***	33.382***	44.756***	-0.827***	-0.647**	-0.668*
	(1.715)	(3.427)	(3.853)	(0.155)	(0.209)	(0.301)
Observations	2020	2020	2020	2020	2020	2020
R^2	-0.106	0.013	0.086	.	.	.
PANEL B: REDUCED FORM						
GSP Treatment	6.365*	6.504*	7.189***	0.456***	0.496***	0.387**
	(3.420)	(3.493)	(2.404)	(0.163)	(0.155)	(0.173)
Budget	0.028	0.028	0.028	-0.004**	-0.004**	-0.005**
	(0.027)	(0.027)	(0.027)	(0.002)	(0.002)	(0.002)
Constant	29.137***	27.223***	39.934***	-1.188***	-1.181***	-0.968***
	(2.417)	(4.685)	(3.421)	(0.119)	(0.156)	(0.201)
Observations	2020	2020	2020	2020	2020	2020
R^2	0.024	0.049	0.179	.	.	.
Pseudo R^2	.	.	.	0.029	0.039	0.082
PANEL C: FIRST STAGE						
GSP Treatment	0.603**	0.700**	0.696**	0.603**	0.700**	0.696**
	(0.259)	(0.280)	(0.300)	(0.259)	(0.280)	(0.300)
Constant	-0.269*	-0.663***	-0.467*	-0.269*	-0.663***	-0.467*
	(0.147)	(0.222)	(0.262)	(0.147)	(0.222)	(0.262)
First Stage F-stat	5.423	6.267	5.396	5.423	6.267	5.396
Observations	101	101	101	101	101	101
R^2	0.090	0.262	0.300	0.090	0.262	0.300
Age Controls	No	Yes	Yes	No	Yes	Yes
Ethnicity Controls	No	Yes	Yes	No	Yes	Yes
Session-Rooms FEs	No	No	Yes	No	No	Yes

Note: all errors are robust and clustered by school × GSP cohort (the unit of randomization in the GSP). Reduced form regressions of partner share (Panel B) are estimated using OLS, and IV regressions (Panel A) with GMM. Reduced form regressions of the indicator variable for giving half of the budget (Panel B) are estimated with probit analysis, and IV regressions (Panel A) use a conditional maximum-likelihood IV probit estimator. The dependent variable in the first stage OLS regressions in Panel C is the KCPE score. *** indicates significance at the 99 percent level; ** indicates significance at the 95 percent level; and * indicates significance at the 90 percent level.

Panel C shows that the F-statistic in the first stage is between 5.3 and 6.3 depending on the controls, and that random assignment to the GSP program increases subsequent KCPE scores by an average of at least 0.6 standard deviations within our sample.[22] Though our first stage F-statistics are below the rule of thumb proposed in Staiger and Stock (1997), the coefficient of interest is median-unbiased in the just-identified case (Angrist and Pischke 2009); nonetheless, hypothesis tests may be incorrectly sized (Dufour 1997; Stock and Yogo 2002). Anderson and Rubin (1949) provides a statistic that produces confidence intervals of the correct size in the presence of weak instruments. These confidence regions are asymmetric and potentially disjoint or unbounded, but the AR statistic allows us to verify that our results are not dependent on inappropriately small Wald standard errors. With no controls or with age and ethnicity controls, the coefficient on the endogenous regressor KCPE score is marginally significant under the AR χ^2 test with p-values of 0.064 and 0.063, respectively, and with additional room fixed effects, it is highly significant with a p-value of 0.003. The 95 percent AR confidence intervals are, respectively, (-0.90,48.45), (-0.71,31.40), and (3.56,42.83). Although these barely include zero in the first two cases, overall the AR test merely shows that we can't reject even larger effects, as the asymmetric confidence intervals are skewed upwards compared to the standard confidence intervals. This strongly suggests that our result is not a spurious consequence of a weak instrument.

Figure 3.1 shows our main result graphically via non-parametric, locally-weighted Fan regressions: the partner share function for participants in the GSP treatment group lies almost entirely above the partner share function for those in the control group.[23]

We further explore the impact of academic achievement on social preferences by estimated IV probit specifications where the outcome variable is an indicator for splitting the budget exactly evenly (Table 3.4, Panel A, columns 4-6). In all specifications, instrumented test scores are positively and statistically significantly associated with a tendency to divide the budget evenly. Thus, the impact of academic achievement is not simply greater generosity, but a clear tenancy to shift toward an exactly equal distribution of the budget. This pattern is consistent with the desire, documented in Charness and Rabin (2002), to avoid receiving a lower payoff than another subject.

3.4 Discussion

At this point, we have established the relationship between the GSP intervention and behavior in our experiment, and explored a one potential causal mechanism linking the scholarship program to respect for earned property rights: academic achievement as mea- sured by KCPE exam scores. We now discuss the channels through which human capital might impact be-

[22] This GSP treatment effect on test scores is larger than the roughly 0.2 to 0.3 standard deviations effect reported in Friedman et al. (2011) for the full GSP Follow-up Survey sample. Sampling variation is a likely explanation for the discrepancy, given our limited subsample of 101 lab subjects.

[23] Following Deaton (1997), we choose a reasonable bandwith by trial and error, since the figure is for illustrative purposes only.

havior in our experiment in more detail, and consider several alternative explanations of our empirical findings.

One possibility is that, as we have argued, human capital directly alters social preferences by increasing respect for earned property rights. In an educational environment where effort is rewarded and the benefits from effort are privately held, one might learn to embrace the values that lead to success in that environment. A related possibility is that success in school is a signal for success later in life, and after observing this signal, students choose self-serving moral codes: those who are capable of high productivity adopt norms that reward high productivity. Either pathway might explain a causal impact of academic achievement on individual beliefs about what constitutes a fair allocation, particularly in settings where individual effort determines income.

An alternative explanation is that winning the scholarship contest impacted individual preferences via a channel other than academic achievement, for example, through a wealth effect. To explore this possibility, we estimated our main regression specifications omitting the 15 subjects who won the scholarship contest (results not shown). Though sample sizes, and consequently significance levels, are reduced somewhat, estimated coefficients are essentially unchanged.

Another possibility is that people choose allocations based on their beliefs about the types of individuals they are likely to be matched with in the experiment: those who believe that other is likely to be kind or altruistic may put more weight on the payoff to other, along the lines proposed in Levine (1998). Thus, individuals with different beliefs about the average level of altruism and respect for property rights in the population (or the experimental subject pool) might behave differently in our experiment even if their underlying preferences were the same. If GSP-induced improvements in test scores caused girls to attend higher quality secondary schools with smarter, kinder peers, academic achievement may be associated with increases in the amount allocated to other in our experiment because beliefs are different, even if social preferences (conditional on beliefs) are the same.

To explore the hypothesis that beliefs, rather than preferences, change with academic experience, we asked participants to report how much they thought *other* would allocated to them at four of the twenty possible budget sizes.[24] Table 3.5 reports OLS regressions of the average amount a subject believed her partner would allocate her on the GSP treatment indicator (Panel A) and the KCPE score (Panel B), both with and without controls. Neither treatment nor academic achievement is significantly associated with beliefs in any specification, and all estimated coefficients are quite small in magnitude. The point estimates suggest a negative relationship between KCPE scores and expectations, instead of the positive relationship required if our results were explained by academic achievers reciprocating a higher

[24] Beliefs were elicited through survey questions and not in an incentive-compatible manner. However, the average belief reported in the survey is not significantly associated with the average amount a subject allocated to her partner. Moreover, beliefs are substantially higher, on average, than actual allocations, despite potential self- and social- image motivations to underestimate others' generosity. Thus we believe the beliefs data are reliable.

perceived level of altruism among their peers. We are consequently able to rule out the possibility that academic achievement mainly impacts beliefs rather than social preferences.

Another alternative explanation for our main results is that the GSP treatment had a positive impact on generalized altruism. Prior to conducting our main experiments, we conducted pilots of standard dictator games (in which dictators divided their own unearned income) with a small sample of 40 subjects, 19 from GSP treatment schools and 21 from control schools. In these small-scale pilots, girls in the GSP control group allocated *other* 19.0 percent of the budget, on average, while girls in the treatment group allocated *other* an average of 16.6 percent of the budget (p-value 0.0229). Thus, the evidence suggests that, if anything, the GSP treatment is associated with lower levels of generalized altruism.

Finally, Table 3.6 shows that un-instrumented academic achievement on the KCPE exam is associated with an increase in the amount allocated to *other* in our main experimental treatment. However, the coefficient on KCPE score is substantially smaller than in the IV regressions reported earlier.[25] It is not surprising that the coefficients are different, since academic outcomes depend on factors such as parental influence, socioeconomic status, and innate individual personality traits which may also shape norms and preferences, as discussed in Malmendier and Nagel (2011).

The fact that the OLS coefficient is smaller suggests that some factors which explain better academic performance are associated with lower levels of respect for earned property rights, or possibly that the IV approach is helping to address attenuation bias caused by noise in the KCPE achievement test score. A further possibility that we cannot rule out is that the GSP experiment affects social preferences through educational channels other than the test score, with schooling attainment being the leading potential channel, and that the IV estimates are in part capturing effects through these other channels. While this possibility somewhat alters the interpretation of the KCPE coefficient estimates, the hypothesized schooling attainment channel is still consistent with the overall thrust of our argument that boosting human capital affects social preferences. Those readers who believe that schooling attainment — or some other outcome — is a major channel through which the scholarship program affects social preferences thus might prefer to focus on the reduced form results in Panel B of Table 3.4 rather than the IV results in Panel A. More generally, the GSP intervention may have changed the likelihood that a girl marries young, or expected lifetime wealth, or the level of social capital in treatment communities. Nonetheless, our reduced form results provide an estimate of the program on behavior in our experiment, and respect for earned property rights, regardless of the channel mediating these impacts.

3.5 Conclusion

We provide evidence that increases in human capital, as captured in academic achievement tests, alter individual values, generating greater respect for earned property rights. This

[25] A Hausman test rejects the equality of the IV and OLS coefficients with 90 percent confidence (p-value 0.065) when the full set of controls is included in the regressions, as in column 3.

Table 3.5: OLS Regressions of Expected Partner Share

	Dep. Var.: Expected Partner Share		
	(1)	(2)	(3)
PANEL A: IMPACTS OF GSP TREATMENT			
GSP Treatment	-0.326	0.508	0.106
	(2.449)	(3.123)	(2.932)
Constant	47.159***	45.334***	46.496***
	(1.739)	(2.092)	(4.085)
Observations	98	98	98
R^2	0.0002	0.033	0.090
PANEL B: ASSOCIATION WITH KCPE SCORE			
KCPE Score	-0.744	-1.283	-1.705
	(1.659)	(1.726)	(1.780)
Constant	47.024***	45.235***	46.087***
	(1.239)	(2.011)	(3.802)
Observations	98	98	98
R^2	0.003	0.041	0.103
Age Controls	No	Yes	Yes
Ethnicity Controls	No	Yes	Yes
Rooms FEs	No	No	Yes

All specifications estimated using OLS and robust standard errors clustered by school × GSP cohort, the unit of randomization in the GSP. *** indicates significance at the 99 percent level; ** indicates significance at the 95 percent level; and * indicates significance at the 90 percent level.

Table 3.6: OLS Regressions of Partner Share on KCPE Scores

	Dep. Var.: Partner Share		
	(1)	(2)	(3)
KCPE Score	3.100**	4.283***	3.380***
	(1.416)	(1.532)	(1.248)
Budget	0.028	0.028	0.028
	(0.027)	(0.027)	(0.027)
Constant	31.973***	31.737***	42.790***
	(1.599)	(3.773)	(3.744)
Observations	2020	2020	2020
R^2	0.023	0.063	0.175
Age Controls	No	Yes	Yes
Ethnicity Controls	No	Yes	Yes
Rooms FEs	No	No	Yes

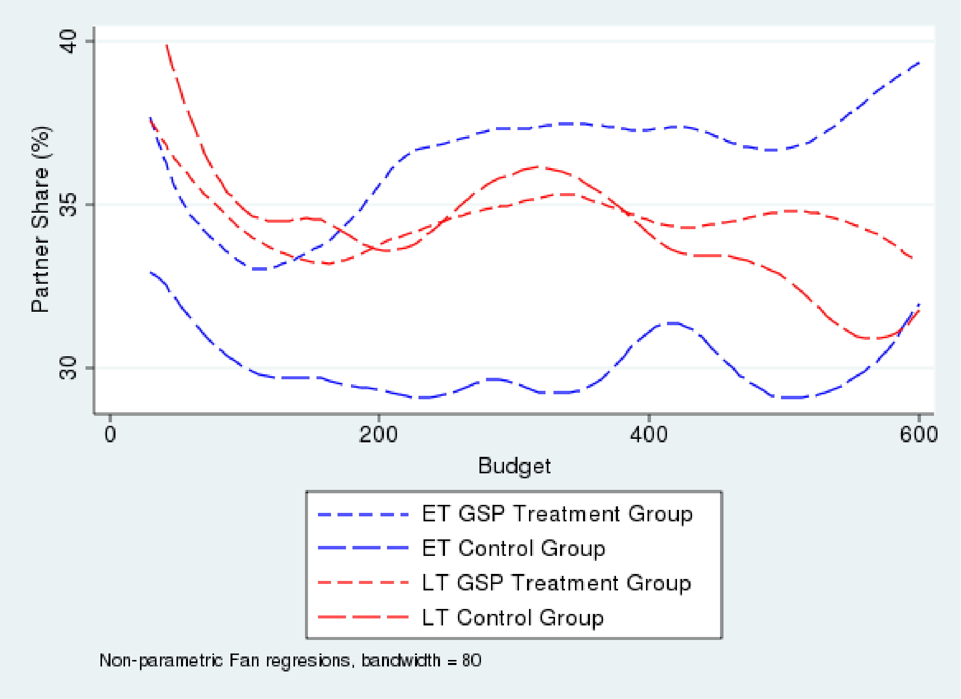
All specifications estimated using OLS and robust standard errors clustered by school × GSP cohort, the unit of randomization in the GSP. Coefficients significantly nonzero at .99 (***), .95 (**) and .90 (*) confidence levels.

finding demonstrates that formal education can have cultural impacts beyond the direct production of human capital, and may have social returns beyond whatever wage gains the human capital generates.

Though there is an extensive empirical literature exploring the labor market returns to education in less developed countries (Duflo 2001, cf.), relatively few empirical studies have directly tested the claims of modernization theory — that formal education leads to changes in individual values — with convincing research designs. Such cultural change could benefit society in several ways. First, as individuals become more respectful of property rights and more permissive of earned wealth accumulation, the private returns to entrepreneurship may increase. This may be particularly important in rural villages in Africa, where strong egalitarian traditions often lead to the social sanctioning of households that accumulate wealth (Barr and Stein 2008; Platteau 2000). More speculatively, the expansion of educational opportunities may generate positive spillovers if changes in values eventually facilitate the emergence of market-oriented institutions (Bernard, De Janvry, and Sadoulet 2010; Glaeser et al. 2004). At the same time, education may have impacts on individual values and beliefs other than those documented here; for example, academic success may change later individual aspirations, and these in turn may influence long-run outcomes (**Ray2006**). Our work complements recent cross-cultural comparisons documenting the correlation between market integration and generosity within dictator games (Henrich, Heine, and Norenzayan 2010; Henrich et al. 2001), and contributes to the emerging literature documenting the causal mechanisms underlying changes in individual values (Di Tella, Galiani, and Schargrodsky 2007; Fisman, Kariv, and Markovits 2009).

Our work is one of one of several recent studies which demonstrate that lab experiments can be combined with randomized controlled trials to measure the direct impact of programs on individual preferences and, more broadly, on social norms and cultural values. In response to recent calls for a greater focus on understanding why and how (rather than just whether) anti-poverty programs "work", we demonstrate that progress in understanding the underlying mechanisms, which is so often the focus of lab experiments, can fit naturally together with the clean econometric identification generated by randomized trials.

Figure 3.1: Fan regressions of Partner Share on Budget



Non-parametric Fan regresions, bandwidth = 80

Appendices

Appendix A

Chapter 1 Extensions

The modeling framework developed in section 1.3 can be used to analyze many situations beyond the primary results on the affect of social pressure discussed section 1.4. I demonstrate the flexibility and power of the model in two such settings in this appendix. This of course merely scratches the surface of the possibility set; future work will also explore social pressure when individuals interact with multiple audiences, dynamic models of norms under various mechanisms of norm transmission/formation/change, the influence of pseudo-rational reasoning on social pressure and norm transmission, the endogenous determination of the relative strength of respect- and approval-seeking motivations, alternative versions of social-image motivations that are likely to arise in specialized environments, and social welfare issues.

A.1 Endogenous norms

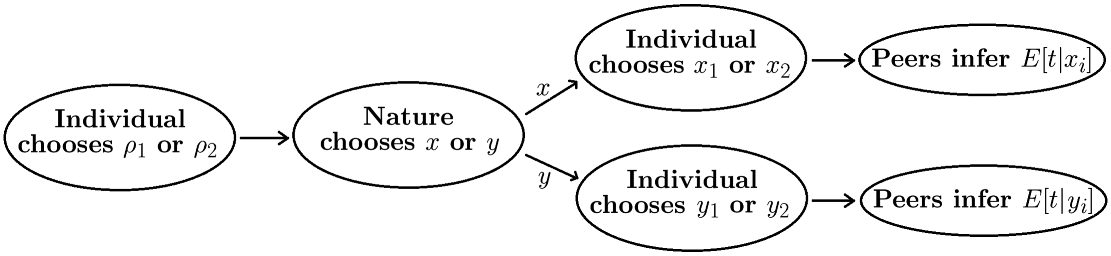
Consider the behavior of respect seekers and approval seekers when ρ is no longer assigned exogenously. Let's see how social-image motivations might influence the equilibrium outcome when individuals have to choose both their moral beliefs and their actions. This enables us to think about the influence of social-image motivations in two new classes of situations: first, if people can change their beliefs over time, but not immediately before each new decision, the long-run equilibrium should be described by a model in which people choose both their beliefs and then their actions. This might describe malleable norms such as fashions and cultural customs. Second, if people are put in a new situation and choose a new norm, the resulting stated beliefs and actions can be described by this model. For example, if a new company is deciding between flexible or inflexible hours, or a society has to decide new rules of etiquette for a new technology such as Google Glass, anticipated social pressure and inferences can impact which side people will choose from the beginning.

The previously considered single decision setting is not quite rich enough to examine this problem satisfyingly. Clearly, people would prefer to choose the ideal that allows them to make selfish choices guilt free. I will introduce an element of uncertainty to make the model substantive.

As an intuitive motivation, imagine that people are born with a certain level of integrity or activist tendencies. In their youth, they choose a personal norm: whether to be egalitarian or utilitarian, for example. As an adult, they face choices that force them to trade off personal gain, their chosen personal norms, and social image, in different ways in different situations.

Formally, suppose that with equal probability, an individual will face either the choice between x_1 and x_2, or the choice between y_1 and y_2, where WLOG, I assume that $v(x_2) > v(x_1)$ and $v(y_1) > v(y_2)$ and $v(x_2) - v(x_1) > v(y_1) - v(y_2)$. That is, action 1 is more profitable in the y choice set while action 2 is more profitable in the x choice set, but in expected utility, action 2 is more profitable. Two possible norms, denoted ρ_1 and ρ_2, respectively indicate that x_1 and y_1 should be chosen, and that x_2 and y_2 should be chosen. Each person is born with their type t but chooses which norm that level of integrity will apply to, prior to realizing which choice set they will be faced with. The distribution of t is again assumed to have full support on $\mathbb{R}^+$ to simplify the analysis; ρ may of course endogenously end up correlated with t. Figure A.1 diagrams the game.

Figure A.1: Endogenous norm game


Individual chooses ρ_1 or ρ_2
Nature chooses x or y
x
y
Individual chooses x_1 or x_2
Individual chooses y_1 or y_2
Peers infer $E[t|x_i]$
Peers infer $E[t|y_i]$

Concretely, imagine that x_2 corresponds to an egalitarian policy. x_1 is efficient but disadvantageous to the decisionmaker. Conversely, y_1 is efficient and advantageous, while y_2 is egalitarian and disadvantageous. This particular example corresponds to a veil of ignorance scenario in which someone has to commit to either egalitarian or utilitarian principles before knowing whether he will be poor or rich.

Once individuals have chosen their ideals, analysis reduces to the section 1.4 case. But, the fact that individuals must optimally choose their ideals when foreseeing that outcome complicates matters substantially.

First I define some additional notation. Let $\underline{\Delta H}$ be the smallest difference in image utilities possible between choosing one or the other option within a given choice set. That is, $\underline{\Delta H} = \min_c H\left(\frac{\int_0^c t\phi(t)dt}{\Phi(c)}\right) - H\left(\frac{\int_c^\infty t\phi(t)dt}{1-\Phi(c)}\right)$. Likewise, $\overline{\Delta H}$ is the largest possible difference in image utilities: $\overline{\Delta H} = \max_c H\left(\frac{\int_0^c t\phi(t)dt}{\Phi(c)}\right) - H\left(\frac{\int_c^\infty t\phi(t)dt}{1-\Phi(c)}\right)$. Notice that both quantities are finite, since H is bounded, and may not correspond to the largest or smallest difference in inferred types. Also note that since $\Delta H = H(\bar{t})$ when $c = 0$, $\underline{\Delta H} < H(\bar{t}) < \overline{\Delta H}$.

Proposition 8 describes the equilibria of respect seekers and approval seekers, which turn out to appear quite similar:

Proposition 8. *In the endogenous norm game, the following hold:*

1. *For respect seekers, equilibria take one of four forms: a pooling equilibrium in which everyone chooses x_2 and y_2 and ρ_2 always, a pooling equilibrium in which everyone chooses ρ_1, x_1 and y_1, an equilibrium in which sufficiently high types choose ρ_1, x_1 and y_1 while lower types choose x_2 over x_1, and an equilibrium in which sufficiently high types choose ρ_2, x_2 and y_2 always, while lower types choose y_1 over y_2.*
2. *For approval seekers, there exists an equilibrium of one of these same four types described in part 1.*

Proposition 9 goes on to describe which equilibria are possible for different levels of social pressure for respect and approval seekers. For both, at very low levels of social pressure, the only sustainable equilibrium has everyone choosing ρ_2, but some low types still defect to y_1 within that choice set. At high levels of social pressure, everyone can be motivated to comply with ρ_2 perfectly. And at even higher levels of social pressure, everyone may choose and comply with ρ_1 despite the fact this is a strictly worse outcome for everyone than everyone choosing and complying with ρ_2. At mid-level social pressure, there is also the potential for partial pooling on ρ_1, in which everyone chooses that norm but low types defect to x_2.

Proposition 9. *In the endogenous norm game, the following hold:*

1. *For respect seekers, at sufficiently small $s < \frac{v(y_1)-v(y_2)}{\underline{\Delta H}}$, there is an equilibrium in which all types choose ρ_2 and some individuals defect to y_1. For sufficiently large $s > \frac{v(y_1)-v(y_2)}{H(\bar{t})}$, there is an equilibrium in which all types choose and comply perfectly with ρ_2. One of these equilibria always exists. For higher $x > \frac{v(x_2)-v(x_1)}{H(\bar{t})}$, there is additionally an equilibrium in which all types choose and comply perfectly with ρ_1. For mid-range $s \in \left[\frac{v(x_2)-v(x_1)+v(y_2)-v(y_1)}{H(\bar{t})+\overline{\Delta H}}, \frac{v(x_2)-v(x_1)}{\underline{\Delta H}}\right]$, there is an equilibrium in which all types choose ρ_1 but some individuals defect to x_2.*
2. *For approval seekers, at sufficiently small s, there is a unique equilibrium in which everyone chooses ρ_2 and complies with it, perhaps imperfectly with low t individuals defecting to y_1. For higher s, there is an additional possible equilibrium in which everyone chooses ρ_1 and complies with it, perhaps imperfectly with low t individuals defecting to x_2.*

This result is fairly intuitive: since ρ_2 is more materially advantageous, on average, in absence of social-image motivations people would prefer to choose that ideal. In that case, with low social pressure, types without much moral guilt will deviate when profitable. But, when social pressure is higher, perfect compliance is attainable. On the other hand, the costly norm is also sustainable in a partial pooling or pooling equilibrium if social pressure is high enough. Proposition 9 suggests that these costly equilibria can be broken by either switching to a different available equilibrium (pooling on ρ_2 is always available as an option when pooling on ρ_1 is sustainable), or by reducing social pressure, whether individuals are respect seekers or approval seekers.

Taken at face value, Propositions 8 and 9 show the models of respect seekers and approval seekers are both consistent with the formation of new customs. For example, suppose every one in a new company prefers to dress casually most of the time. That is, most days choosing casual dress corresponds to choosing x_2. But sometimes it might be more convenient to dress formally in order to go straight to a formal event after work, or they might just be in the mood to dress nicely; in these cases choosing casual dress corresponds to choosing y_2. Absent social pressure, everyone would prefer a policy of casual dress. But if someone anticipates that everyone else will advocate a formal policy, and they fear that their advocacy of casual dress will be interpreted as insincere and selfish (by respect seekers) or simply frowned upon (by approval seekers), they might choose to support a formal policy from the beginning.

The differences between respect seekers and approval seekers in this setting are reduced to the kinds of differences we see in homogeneous norm decisions, as discussed in section 1.3. This is simply because all possible equilibria when ρ is endogenous involve pooling on a single belief system. It's in fact intuitive that approval seekers would be prone to pooling on a single moral code, because they don't want to thwart each other's beliefs. It's more surprising, however, that respect seekers do not have the option of contradicting the majority opinion and insisting that that be interpreted as true belief in an alternative ideal.

This example reveals that even in scenarios that appear to be morally uncontentious (at least, within a particular society), the models of respect seekers and approval seekers can help us understand the dynamics of social image. The respect- or approval-seeking motivation may be what is enforcing that unanimity, and while the outcome is qualitatively the same for respect seekers and approval seekers, the meaning of social image is different. Take, for example, violent conflict. If individuals are pressured into participating in such a dispute, and they participate in order to signal their commitment to the cause despite not believing in those destructive and costly tactics, then this situation clearly constitutes a Pareto inefficient equilibrium. But as the model suggests, whether or not people care about respect or approval, there must be another feasible equilibrium in which everyone refuses to use violent tactics. But it may be difficult to trigger a unilateral shift in beliefs about good forms of political participation. Another option, however, is to reduce social pressure to the point where the destructive equilibrium isn't sustainable. Making choices more anonymous might accomplish this, or respect seekers may be willing to signal their integrity in another way if provided an alternative, such as committing to participate in personally costly but welfare-increasing activities instead. Approval seekers, on the other hand, have to be convinced that they won't be harshly judged for deviating. Publicizing and praising other approaches, or making it easier for dissenters to anonymously advertise their presence with immunity, may serve this purpose.

A.2 Group Membership and Marketing

The analysis of sections 1.4 and 1.4 can be used to answer more prescriptive questions than the positive analysis of the impact of social pressure focused on above. This subsection addresses the *choice* of v_1.

Imagine that a political action group wants to choose a membership fee that will attract a certain number of people who are most in agreement with the group's mission. This group is dedicated to a minority interest in the overall population, such as the Libertarian party, or the NAACP. In some of these contexts, social pressure is obviously very low, but in other cases it could be significant: approval seekers wouldn't judge a rent control reform activist too harshly, but they would probably shun someone who joined the KKK. Likewise, respect seekers wouldn't care that someone's donations to the Human Society of America signals his true dedication to pet care, but they might admire someone who pays dues to the ACLU each month (whether or not they share the views of that organization.)

In an approval-seeking society, if social pressure is very high, the membership dues will need to be very low in order to draw in enough members. If social pressure is low, anyone with a passing interest in the group will happily join, so to keep numbers at a desirable level and restrict membership to the truest believers (to maintain group cohesion, say, or ration any provided benefits), the dues must be high. On the other hand, in a respect seeking society when social pressure is high, potential members will be anxious to signal their integrity by adhering to their ideal and will only be dissuaded from joining if the cost is high. If social pressure is low, on the other hand, interested members are happy to avoid the cost by defecting to the majority ideal, so the cost must be low to attract the same numbers.

This intuition is formalized in the following Proposition. All assumptions from section 1.4 apply.

Proposition 10. *If a group needs to set membership dues d to attract a fraction $q \leq p$ of the population, where p is the fraction of the population that is sympathetic to the group, then:*

1. *for approval seekers, d is strictly decreasing in s, and*

2. *for respect seekers, d is increasing in s over the range of s but* possibly *decreasing for small increases in s.*

This provides a clear test of whether people act like approval seekers or respect seekers in a particular context: If, after increasing social pressure (say by making member roles public, or after interest and publicity on a particular issue rises for any reason), membership in a political group increases, approval can't be the dominant type of social-image motivation that is in operation. This example also emphasizes the importance of understanding how social image and partisan beliefs operate: depending on what aspect of reputation individuals are concerned with, a political action group trying to recruit volunteers or drum up enthusiasm should use very different tactics, appealing to different kinds of social image.

An isomorphic interpretation of this example could also apply to marketing practices: Suppose a company sells a product that is associated with some group's identity. This could be goth clothing, club or hobby-related paraphernalia, or any version of a product that is sold at a premium when labeled or associated with a particular identity. Then respect seekers, who wish to signal integrity towards their personal identities (which, in this context, create personal norms directing their own behavior), will be willing to pay a high premium when social pressure is high. On other issues, people might act more as approval seekers, and avoid any explicit labeling of their minority interests. Products associated with those identities will command lower prices.

This analysis rests on an assumption that the company has a fixed stock they wish to sell at the highest possible price. In general, of course, companies have a degree of control over quantity in addition to price, but the profit maximizing price, as a function of s, isn't pinned down without additional simplifying assumptions. The key conceptual point, however, remains: demand for a good associated with a minority identity/group/viewpoint will decrease as social pressure rises if consumers are approval seekers, and will rise if consumers are respect seekers.

This analysis also raises a welfare question: does the creation of a signaling opportunity make people better or worse off? Are people with lawn signs excited to be participating in an election, or are they reluctantly helping out when asked? Do inactive or passive club members continue to pay dues because they want to support its continued existence, or would they prefer the club didn't exist so they wouldn't feel pressured to join? The sign of H is not committed to be positive or negative in either model, so anything is possible. This is an important question but not strongly relevant to the comparison of the two models of social image, and not one I wish to commit to *because* the framework is designed to be generally applicable, so I defer further discussion to future work.

Appendix B

Chapter 1 Proofs

Throughout these proofs, for notational convenience, define $m_i = m(x_i)$ (or $m_{i,as}, m_{i,rs}$), $H_i = H(m_i)$ (or $H_{i,rs}, H_{i,as}$) and $v_i = v(x_i)$.

Proposition 1 part 1. *If $X = \{x_1, x_2\}$, $v(x_2) > v(x_1)$, and ρ is independent from t, the following holds:*

1. *For approval seekers, there is exactly one equilibrium outcome, in which sufficiently high-t types will adhere to their personal ideals and low types will choose whichever option yields a better combination of material and image utility. Choosing x_1, the more costly option, will lead to a higher social image* iff *$p_1 > .5$.*

Proof: Type t with ρ_1 will choose x_1 *iff* $v_1 + sH_1 > v_2 - tG + sH_2 \Leftrightarrow$

$$t > \frac{v_2 - v_1 + s(H_2 - H_1)}{G} \equiv \tilde{t}_1.$$

Likewise, type t with ρ_2 will choose x_2 *iff*

$$t > \frac{v_1 - v_2 + s(H_1 - H_2)}{G} \equiv \tilde{t}_2 = -\tilde{t}_1.$$

Since one of these cutoff values is positive and one is negative, low t types with one ideal will defect to the other action, and all types with the other ideal will adhere to their ideal. For approval seekers, $H_1 = H_{as}(p_1)$ and $H_2 = H_{as}(1 - p_1)$ are exogenous, so all components $\tilde{t}_1$ and $\tilde{t}_2$ are exogenously fixed, so existence and uniqueness is trivial. The last statement is immediate from assumption 1.

Proposition 1 part 2. *If $X = \{x_1, x_2\}$, $v(x_2) > v(x_1)$, and ρ is independent from t, the following holds:*

2. *For respect seekers,*

a) There exists at least one pure strategy equilibrium, and all equilibria must take a form in which sufficiently high-t types will adhere to their personal ideals and low types will choose x_2. In these equilibria, choosing the costly action x_1 leads to a higher social image.

b) This equilibrium is unique if 1) G is sufficiently large, 2) s is sufficiently small, 3) p_1 is sufficiently small, or 4) $\max \phi(t)$ *is sufficiently small.*

Proof: As in the proof of Proposition 1 part 1, the cutoff values $\tilde{t}_1$ and $\tilde{t}_2$ are opposite sign, so there are two possibilities: either all first types choose x_1 while some low t second types also choose x_1, or vice versa. Now, however, $H_{1,rs}$ and $H_{2,rs}$ are endogenously determined.

Suppose that the former possibility is the case: all types with ρ_1 adhere to x_1 and low t types with ρ_2 defect. Then it must be that $\tilde{t}_1 < 0$ (ignoring knife-edge cases). But then, $s(H_1 - H_2) > v_2 - v_1$, which requires $H_{1,rs} > H_{2,rs}$. But this cannot be the case because low t types with ρ_2 are also choosing x_1, which makes the conditional expectation of t on choosing x_1 lower than on choosing x_2.

So we must have $\tilde{t}_1 > 0$, $\tilde{t}_2 < 0$. We must now only show that such an equilibrium exists.

Given the inference function and this cutoff value, we can calculate the image associated with each choice:

$$m_{2,rs}(\tilde{t}_1) = \frac{(1-p_1)\bar{t} + p_1 \int_0^{\tilde{t}_1} t\phi(t)dt}{1 - p_1 + p_1\Phi(\tilde{t}_1)}$$

and

$$m_{1,rs}(\tilde{t}_1) = \frac{\int_{\tilde{t}_1}^{\infty} t\phi(t)dt}{1 - \Phi(\tilde{t}_1)}.$$

These two equations, along with the one defining $\tilde{t}_1$ above, define the equilibria of the model. This system of equations must be shown to have a solution with $\tilde{t}_1 > 0$.

Define

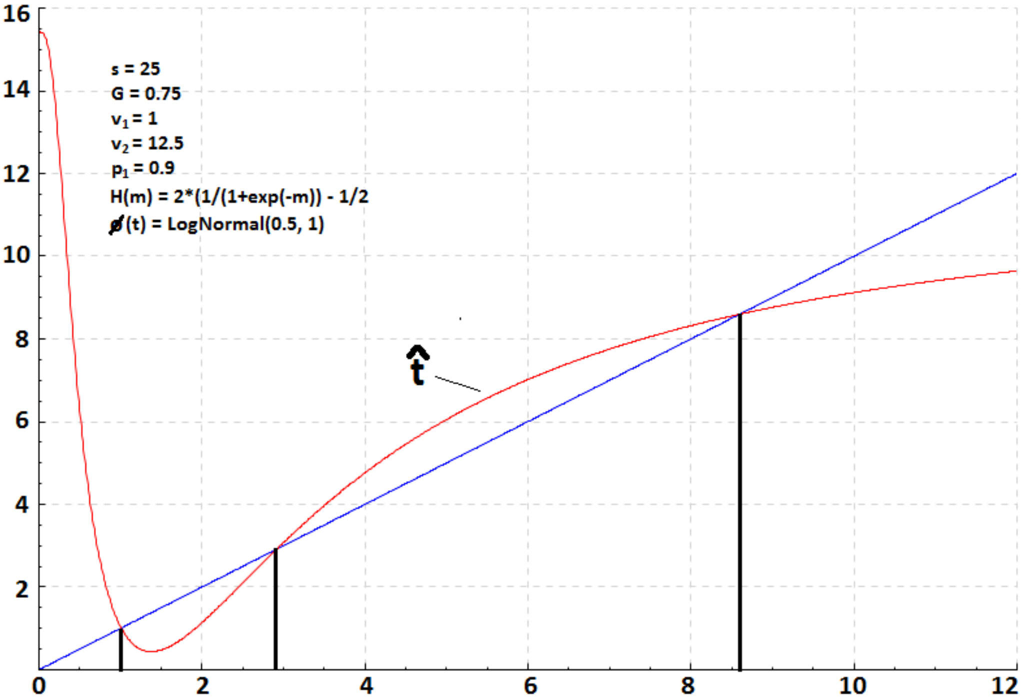
$$\hat{t}(t) = \frac{s(H_{rs}(m_{2,rs}(t)) - H_{rs}(m_{1,rs}(t))) + v_2 - v_1}{G}.$$

This is a continuous and finite valued function, by assumption 1. At $t = 0$, $\hat{t} = (s(H_{rs}(\bar{t}) - H_{rs}(\bar{t})) + v_2 - v_1)/G > 0 = t$. As $t \to \infty$, $\hat{t} < t$ necessarily. Therefore by the intermediate value theorem, there is some positive, finite t with $\hat{t}(t) = t$. This provides the desired equilibrium value of $\tilde{t}_1$ and determines the equilibrium outcome fully.

Figure B.1 shows an example graph of t and $\hat{t}$, with three intersections and therefore three possible equilibria.

Number of equilibria: Looking at Figure B.1 again for reference, note that multiple equilibria can only exist if $\hat{t}$ achieves a slope of larger than 1. (This is of course a necessary, not sufficient, condition.) Applying Liebniz's rule, we find that $\frac{\delta m_1(t)}{\delta t} = \frac{\phi(t)}{1-\Phi(t)}(m_1(t) - t)$, $\frac{\delta m_2(t)}{\delta t} = \frac{p\phi(t)}{1-p+p\Phi(t)}(t - m_2(t))$, and therefore we can (conservatively)guarantee that there is a

Figure B.1: Example with three equilibria



An example of the model with the specified parameters. The curve shows $\hat{t}$ as defined in the proof of Proposition 1, and any point where it crosses the 45° line marks an equilibrium.

unique equilibrium if $\frac{\delta \hat{t}}{\delta t} < 1$, or equivalently

$$H'(m_2(t))\left(\frac{\phi(t)}{1-\frac{1}{p}-\Phi(t)}\right)(m_2(t)-t) - H'(m_1(t))\left(\frac{\phi(t)}{1-\Phi(t)}\right)(m_1(t)-t) < G/s.$$

Note that $m_1(t) > t$ and H is increasing so the second term is positive. $1/p > 1$ so the first term can also be positive if $m_2(t) < t$. Since this is not always the case, we can't always guarantee uniqueness, as demonstrated by the example in Figure B.1. But, the right side will be always larger than the left if 1) G is sufficiently large, 2) $\max \phi(t)$ is sufficiently small, 3) p_1 is sufficiently small, guaranteeing that $1 - \frac{1}{p}$ is highly negative, or 4) s is sufficiently small.

Finite support for $\phi(t)$: As noted in the text, Proposition 1 holds strictly as stated, under the D1 criterion, even if the support of $\phi(t)$ is allowed to be finite. If $\max \operatorname{supp} \phi = T < \infty$, there is a discontinuous drop in m_1 from T to 0 when $\tilde{t}_1$ increases just past T, so the above equilibrium no longer applies. It's now possible that no equilibria exists in which both actions are chosen.

Note that the inference function as described does not apply to actions that are never taken in equilibrium. But, we can use the D1 criterion of Cho and Kreps (1987) to explore these non-separating equilibria.

Under the D1 criterion, in order to rule out type (t, ρ) from the inference function after a disequilibrium choice x is observed, it must be the case that for any mistaken belief about inferences off the equilibrium path that might induce (t, ρ) to deviate to x (that is, with indifference or strict preference), there is another type (t', ρ') (the same type for any potential mistaken belief) who would strictly prefer to deviate with that same mistaken belief.

First suppose no one chooses x_1 in equilibrium. Type (t, ρ_2) might deviate to x_1 under mistaken beliefs $\tilde{m}_1$, resulting in mistaken beliefs about image utility $\tilde{H}_1$, if $s\tilde{H}_1 \geq v_2 - v_1 + sH_2 + tG$. Type (t, ρ_1) might deviate if $s\tilde{H}_1 > v_2 - v_1 + sH_2 - tG$. Clearly, if any type is willing to deviate for a given $\tilde{H}_1$, then the type (T, ρ_1) strictly prefers to deviate. This is therefore the only type that can be inferred after observing x_1, and $m(x_1)$ is required to be T.

On the other hand, $m(x_2) = \bar{t} < T = m(x_1)$. If $v_2 - v_1$ is large enough to overcome the image benefit of defecting, then, this pooling equilibrium is sustainable. This occurs when $v_2 + sH(\bar{t}) - TG \geq v_1 + sH(T) \iff T \leq \frac{v_2 - v_1 + s(H(\bar{t}) - H(T))}{G}$. But this is exactly the opposite of the condition that guaranteed a separating equilibrium above. Therefore, if no separating equilibrium exists, there is a pooling equilibrium (and vice versa, guaranteeing existence of *some* equilibrium) in which all types choose x_2, in accordance with the statement of the proposition.

Note that if no one chooses x_2 in equilibrium, a similar argument shows that $m(x_2) = T$. But now $m(x_1) = \bar{t} < m(x_2)$, and type (T, x_2) would strictly prefer to deviate, so no pooling equilibrium on x_1 exists. This shows that pooling equilibria also satisfy the conditions of the theorem.

Proposition 2 part 1. *If $X = \{x_1, x_2\}$, $v(x_2) > v(x_1)$, and ρ is independent from t, the following hold:*

1. *For approval seekers, in the limit as social pressure increases, a consensus forms around x_1 (x_2) if $p_1 > .5$ ($p_1 < .5$), enforced by high social-image utility relative to x_2 (x_1).*

2. *For respect seekers, in the limit as social pressure increases, everyone strictly adheres to their personal ideal and $m_{rs}(x_1) > m_{rs}(x_2)$.*

Proof: Part 1 follows from the proof of Proposition 1 part 1: As $s \to \infty$, one of the cutoff values (corresponding to the ideal with the lower image) will approach infinity as well, so that all types with either ideal will choose the other action. The relative social-image utility is immediate from assumption 1.

As for part 2, at low levels of s, the relative cost of actions determines the relative numbers that choose those actions and the relative image consequences of them. If $s = 0$ exactly, the signaling game disappears and people simply choose action one unless their guilt from not choosing action two outweighs the cost. As $s \to \infty$, on the other hand, image motivations dominate all other concerns, so any difference between $H_{2,rs}$ and $H_{1,rs}$ is not sustainable in equilibrium. By Proposition 1 part 2, the only way for them to be equal is for $\tilde{t}_1 = \tilde{t}_2 = 0$.

Proposition 3: *If $X = \{x_1, x_2\}$, $v(x_2) > v(x_1)$, and ρ is independent from t, the following hold:*

1. *For approval seekers, if $p_1 > .5$, equilibrium is increasingly sacrificial when s is sufficiently high.*
2. *For respect seekers, equilibrium is never sacrificial.*

This follows directly from assumption 1 and Propositions 1 and 2.

Proposition 4 part 1: *If $X = \{x_1, x_2\}$, $v(x_2) > v(x_1)$, and t and ρ are correlated, then for respect seekers:*

1. *There exists an equilibrium in which all respect seekers with ρ_1 (ρ_2) choose x_1 (x_2), in addition to types with sufficiently low t and ρ_2 (ρ_1).*

Proof: Similarly to Proposition 1 part 2, a system of three equations for $m_{1,rs}(\tilde{t}_1)$, $m_{2,rs}(\tilde{t}_1)$, and $\tilde{t}_1$ define the equilibria. And as before, if either cutoff value $\tilde{t}_i$ is positive, *all* types with the other ideal choose their ideal action. The system of equations is:

$$m_{1,rs}(\tilde{t}_1) = \frac{p_1 \int_{\max(0,\tilde{t}_1)}^{\infty} t\phi_1(t)dt + (1-p_1)\int_0^{\max(0,-\tilde{t}_1)} t\phi_2(t)dt}{p_1(1-\Phi_1(\tilde{t}_1)) + (1-p_1)\Phi_2(-\tilde{t}_1)}$$

$$m_{2,rs}(\tilde{t}_1) = \frac{p_1 \int_0^{\max(0,\tilde{t}_1)} t\phi_1(t)dt + (1-p_1)\int_{\max(0,-\tilde{t}_1)}^{\infty} t\phi_2(t)dt}{p_1\Phi_1(\tilde{t}_1) + (1-p_1)(1-\Phi_2(-\tilde{t}_1))}$$

$$\tilde{t}_1 = \frac{s(H_{2,rs} - H_{1,rs}) + v_2 - v_1}{G}$$

The argument for existence of equilibrium follows similarly to Proposition 1 part 2, but is more directly implied by Brouwer's fixed point theorem. $\hat{t}(t)$, as defined above, is finite valued (bounded due to the upper bound on H) and continuous, so it maps a convex, compact subset of $\mathbb{R}^3$ to itself. Therefore $\hat{t} = t$ has a solution, which provides the equilibrium value of $\tilde{t}_1$. But, unlike before, we can't rule out either sign of $\tilde{t}_1$, so imitation in either direction can occur.

Proposition 4 part 2 and 3: *If $X = \{x_1, x_2\}$, $v(x_2) > v(x_1)$, and t and ρ are correlated, then for respect seekers:*

2. *As social pressure increases, in the limit, if $E[\phi_1] > E[\phi_2]$, everyone with ρ_1 will choose x_1 and those with ρ_2 and sufficiently low t will also choose x_1. The same applies if subscripts are reversed: the role of costliness is irrelevant in the limit.*

3. *If $E[t|\rho_1] > E[t|\rho_2]$, social pressure that is sufficiently high can sustain a sacrificial equilibrium.*

Proof: As before, a difference in the image outcome of each choice isn't sustainable in equilibrium as s becomes sufficiently large. Given the operation of the cutoff values $\tilde{t}_1$ and $\tilde{t}_2$, clearly the only way for the image outcome to be the same is for low t types with the less "prestigious" ρ ideal (i.e. the ideal of the sub-population with the higher average t) to seek a higher status by choosing against their ideal. The costliness of c influences the exact cutoff values but doesn't change which action is chosen by impostors seeking higher status.

Part 3 simply points out again what part 1 says when correlation implies this relationship: costliness doesn't prevent imitation, leading to "too much" sacrifice overall.

Proposition 5 part 1: *If $X = \{x_1, x_2, x_3\}$ with $G(x_2 - x_1) = G(x_2 - x_3) = G_1$ and $G(x_1 - x_3) = G(x_3 - x_1) = G_2 > G_1$, and ρ is uncorrelated with t, then the following holds:*

1. *For respect seekers, at least one equilibrium exists, in which high integrity types adhere to their norms. Either low types with either norm defect to the middle option, or low types with ρ_1 defect while ρ_3 is adhered to perfectly. In the latter case, low types with ρ_1 either all defect to x_3, or mid-level types defect to the middle and low types defect to the opposite ideal.*

Proof: Define $\tilde{t}_{i,j,k}$ to be the cutoff type t above which someone with ideal x_i will prefer x_j to x_k. In particular,

$$\tilde{t}_{1,1,2} = \frac{s(H_2 - H_1) + v_2 - v_1}{G_1},$$

$$\tilde{t}_{1,2,3} = \frac{s(H_3 - H_2) + v_3 - v_2}{G_2 - G_1},$$

$$\tilde{t}_{1,1,3} = \frac{s(H_3 - H_1) + v_3 - v_1}{G_2},$$

$$\tilde{t}_{3,3,2} = \frac{s(H_2 - H_3) + v_2 - v_3}{G_1} = -\tilde{t}_{1,2,3}\frac{G_2 - G_1}{G_1},$$

$$\tilde{t}_{3,2,1} = \frac{s(H_1 - H_2) + v_1 - v_2}{G_2 - G1} = -\tilde{t}_{1,1,2}\frac{G_1}{G_2 - G_1},$$

and

$$\tilde{t}_{3,3,1} = \frac{s(H_1 - H_3) + v_1 - v_3}{G_2} = -\tilde{t}_{1,1,3}.$$

Note that $\tilde{t}_{1,1,2}$ and $\tilde{t}_{3,2,1}$, $\tilde{t}_{1,1,3}$ and $\tilde{t}_{3,3,1}$, and $\tilde{t}_{1,2,3}$ and $\tilde{t}_{3,3,2}$, are respectively opposite sign, and that they are pairwise determined. These relationships, along with a requirement of transitivity for all types, restricts the possible relationships between the six cutoff values to one of 5 behaviorally distinct types of equilibria (the reader can check that any relationship not included in this list isn't feasible):

Type 1: $\tilde{t}_{1,1,2} > \tilde{t}_{1,1,3} > \tilde{t}_{1,2,3} > 0$ (while $\tilde{t}_{3,2,1}, \tilde{t}_{3,3,2}, \tilde{t}_{3,3,1} < 0$ necessarily). Types with ρ_1 differentiate between all three options: types with $t > \tilde{t}_{1,1,3}$ choose x_1, with $\tilde{t}_{1,2,3} < t < \tilde{t}_{1,1,3}$ choose x_2, and with $t < \tilde{t}_{1,2,3}$ choose x_3. All types with ρ_3 choose x_3.

Type 2: $\tilde{t}_{1,2,3} > \tilde{t}_{1,1,3} > 0, \tilde{t}_{1,1,2}$ ($\tilde{t}_{1,1,2}$ may have either sign). In this type, types with ρ_1 choose x_1 if $t > \tilde{t}_{1,1,3}$ and x_3 otherwise. All types with ρ_3 choose x_3.

Type 3: $\tilde{t}_{1,1,2} > 0, \tilde{t}_{1,1,3} > \tilde{t}_{1,2,3}$. In this type, types with ρ_1 choose x_1 if $t > \tilde{t}_{1,1,2}$ and choose x_2 otherwise, and types with ρ_2 choose x_3 if $t > \tilde{t}_{3,3,2} > 0$ and x_2 otherwise.

Type 4: $\tilde{t}_{3,2,1} > \tilde{t}_{3,3,1} > 0, \tilde{t}_{3,2,3}$. In this type, types with ρ_2 choose x_3 if $t > \tilde{t}_{3,3,1}$ and x_1 otherwise. All types with ρ_1 choose x_1.

Type 5: $\tilde{t}_{3,3,2} > \tilde{t}_{3,3,1} > \tilde{t}_{3,2,1} > 0$. Types with ρ_3 differentiate between all three options: types with $t > \tilde{t}_{3,3,1}$ choose x_3, with $\tilde{t}_{3,2,1} < t < \tilde{t}_{3,3,1}$ choose x_2, and with $t < \tilde{t}_{3,2,1}$ choose x_1. All types with ρ_1 choose x_1.

Additionally, the assumption that $v_3 > v_1$ eliminates the last two possibilities. In these equilibria, by definition of the image function, $H_1 < H3$, so since $v_3 > v_1$ as well, $\tilde{t}_{3,3,1} = \frac{s(H_1-H_3)+v_1-v_3}{G_2}$ must be negative. But equilibria of type 4 or 5 require that it be positive.

This establishes the described form of all equilibria. Next, I will show that only type 2 equilibria are permitted in the limit when $s \to \infty$.

1. By definition of m_{rs}, in a type 1 equilibrium, $m_1 > m_2, m_3$. A partial requirement for a type 1 equilibrium is that $\tilde{t}_{1,1,3} > \tilde{t}_{1,2,3} > 0 \leftrightarrow \frac{v_3-v_1+sH_3-sH_1}{G_2} > \frac{v_3-v_2+sH_3-sH_2}{G_2-G_1} > 0$. Therefore, $\tilde{t}_{1,1,3} > 0$ requires, as $s \to \infty$, that $m_1 \to m_3$ and $\tilde{t}_{1,1,3}$ remains finite. This occurs only when $\tilde{t}_{1,1,3} \to 0$, which implies that $m_2 = 0$, which implies that $\tilde{t}_{1,2,3}$ grows infinite. This contradicts the stated relationship, so no equilibrium of type 1 exists when $s \to \infty$.

2. Type 2 requires, in part, that $\tilde{t}_{1,2,3} > \tilde{t}_{1,1,3} > 0 \leftrightarrow \frac{v_3-v_2+sH_3-sH_2}{G_2-G_1} > \frac{v_3-v_1+sH_3-sH_1}{G_2} > 0$. By definition of m_{rs}, $m_1 > m_3$, and m_2 is undefined as x_2 is never chosen on the equilibrium path. We must resort to the D1 criterion to evaluate m_2.

   We must consider three types of deviations to x_2: A person with ρ_1 and $t < \tilde{t}_{1,1,3}$ would normally choose x_3, but would prefer x_2 if $sH_2 > v_3 - v_2 + sH_3 - t(G_2 - G_1)$. Since $G_2 > G_1$, then if type $t \in [0, \tilde{t}_{1,1,3})$ is tempted to deviate for some mistaken belief $\hat{H}_2$, then type $t = \tilde{t}_{1,1,3}$ is also tempted to deviate for the same mistaken belief. The D1 criterion therefore says that no weight can be placed on $t \in [0, \tilde{t}_{1,1,3})$ (along with an inferred ρ_1) when inferring a type after observing x_2. By a similar argument, someone with ρ_1 and $t > \tilde{t}_{1,1,3}$ would deviate from their normal choice of x_1 under a mistaken belief satisfying $s\hat{H}_2 > v_1 - v_2 + sH_1 + tG$, and similarly no weight can be placed on $t \in (\tilde{t}_{1,1,3}, \infty)$ (along with an inferred ρ_1) when inferring a type from a choice of x_2. Lastly, someone with ρ_3 might wish to deviate for a mistaken belief satisfying

$sH_2 > v_3 - v_2 + sH_3 + tG_1$, and no weight may be placed on $t \in (0, \infty)$ (along with an inferred ρ_3) when observing x_2. Altogether, all weight must be placed on $t = 0$ or $t = \tilde{t}_{1,1,3}$, which implies that $m_2 \in [0, \tilde{t}_{1,1,3}]$.

Referring back to the required relationship above, $\tilde{t}_{1,1,3} > 0$ requires that $H_1 \to H_3$ as $s \to \infty$, which can only occur when $\tilde{t}_{1,1,3} \to 0$. By the D1 criterion, as above, this means that $m_2 \to 0$. Therefore, $\tilde{t}_{1,2,3} \to \infty$, and $\tilde{t}_{1,1,3} \to \frac{v_3 - v_1}{G_2}$, and the relationship is satisfied *iff* $v_3 > v_1$, as we have assumed.

The final requirement is that $\tilde{t}_{1,1,3} > \tilde{t}_{1,1,2}$, which is also satisfied since $\tilde{t}_{1,1,2} \to -\infty$.

In sum, there exists an equilibrium of type 2 as $s \to \infty$.

3. Type 3 equilibria require, in part, that $\tilde{t}_{1,1,2} > 0 \leftrightarrow v_2 - v_1 + sH_2 - sH_1 > 0$ and $\tilde{t}_{1,2,3} > 0 \leftrightarrow v_3 - v_2 + sH_3 - sH_2 > 0$. And by definition of m_{rs}, $m_1, m_3 > m_2$. The latter inequality is therefore always satisfied as $s \to \infty$. The former inequality requires both that $v_2 > v_1$ and $H_1 \to H_2$. But by definition of m_{rs}, this can only occur if $\tilde{t}_{3,3,2} \to \infty$. But this can't be true, since $\tilde{t}_{3,3,2} = \frac{v_2 - v_3 + sH_2 - sH_3}{G_1} \to -\infty$ when $H_2 = H_1 = H(\bar{t})$ and $H_3 \to \infty$. So no type 3 equilibrium exists when $s \to \infty$.

It remains to be shown that some equilibrium of one of these three types always exists. I will again appeal to Brouwer's fixed point theorem, but a continuous function on a compact, convex space that defines equilibrium at its fixed points must be carefully constructed. In the following, the three parameters of interest are $t_{1,1,2}$, $t_{1,1,3}$ and $t_{1,2,3}$, but I will refer to $t_{3,j,k}$ where convenient rather than the equivalent values written in terms of $t_{1,j,k}$.

A type 1 equilibrium is defined by the relationship $\hat{t}_{1,1,2}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,1,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,2,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3})$ along with the following six equations that must be satisfied:

$$\hat{t}_{1,1,2}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{s(H(m_2(t_{1,1,2}, t_{1,1,3}, t_{1,2,3})) - H(m_1(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}))) + v_2 - v_1}{G_1}$$

$$\hat{t}_{1,1,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{s(H(m_3(t_{1,1,2}, t_{1,1,3}, t_{1,2,3})) - H(m_1(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}))) + v_2 - v_1}{G_2}$$

$$\hat{t}_{1,2,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{s(H(m_3(t_{1,1,2}, t_{1,1,3}, t_{1,2,3})) - H(m_2(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}))) + v_2 - v_1}{G_2 - G_1}$$

$$m_1(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{\int_{t_{1,1,2}}^{\infty} t\phi(t)dt}{1 - \Phi(t_{1,1,2})}$$

$$m_2(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{\int_{t_{1,2,3}}^{t_{1,1,2}} t\phi(t)dt}{\Phi(t_{1,1,2}) - \Phi(t_{1,2,3})}$$

and

$$m_3(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{(1 - p_1)\bar{t} + p_1 \int_0^{t_{1,2,3}} t\phi(t)dt}{1 - p_1 + p_1\Phi(t_{1,2,3})}.$$

And in a type two equilibrium, the first three equations remain the same, but we must have the relationship $\hat{t}_{1,1,2}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,1,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,2,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3})$ and the image functions

$$m_1(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{\int_{t_{1,1,3}}^{\infty} t\phi(t)dt}{1 - \Phi(t_{1,1,3})}$$

$$m_2(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = t_{1,1,3}$$

and

$$m_3(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{(1-p_1)\bar{t} + p_1 \int_0^{t_{1,1,3}} t\phi(t)dt}{1 - p_1 + p_1\Phi(t_{1,1,3})}.$$

where m_2 results from restricting attention to a subset of equilibria that satisfy the D1 criterion. As shown above, m_2 must fall in the interval $[0, t_{1,1,3}]$, and imposing $m_2 = t_{1,1,3}$ ensures continuity in $t_{1,1,2}$, $t_{1,2,3}$, and $t_{1,1,3}$.

In a type three equilibrium, the expressions for $\hat{t}_{i,j,k}$ remain the same but we must satisfy $\hat{t}_{1,1,2}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,1,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,2,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3})$ and the following image functions:

$$m_1(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{\int_{t_{1,1,2}}^{\infty} t\phi(t)dt}{1 - \Phi(t_{1,1,2})}$$

$$m_2(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{(1-p_1)\int_0^{t_{3,3,2}} + p_1 \int_0^{t_{1,1,2}} t\phi(t)dt}{(1-p_1)\Phi(t_{3,3,2}) + p_1\Phi(t_{1,1,2})}$$

and

$$m_3(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{\int_{t_{3,3,2}}^{\infty} t\phi(t)dt}{1 - \Phi(t_{3,3,2})}.$$

We can combine the conditions for all three types of equilibria as follows: The equations for $\hat{t}_{i,j,k}$ remain the same, and we must satisfy *either* $\hat{t}_{1,1,2}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,1,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,2,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3})$ *or* $\hat{t}_{1,2,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,1,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,2,3}(t_{1,1,2}, t_{1,1,3}, t_{1,1,2})$. And, we have the following image functions:

$$m_1(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{\int_{\max(t_{1,1,3}, t_{1,1,2})}^{\infty} t\phi(t)dt}{1 - \Phi(\max(t_{1,1,3}, t_{1,1,2}))}$$

$$m_2(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \begin{cases} \frac{p_1 \int_{\max(t_{1,2,3},0)}^{t_{1,1,2}} t\phi(t)dt + (1-p_1)\int_0^{\max(0,t_{3,3,2})} t\phi(t)dt}{p_1(\Phi(t_{1,1,2}) - \Phi(\max(0,t_{1,2,3})) + (1-p_1)\Phi(\max(0,t_{3,3,2})))} & \text{if } t_{1,2,3} < t_{1,1,3} \\ t_{1,1,3} & \text{otherwise} \end{cases}$$

(ensuring continuity again by imposing $m_2 = t_{1,1,3}$ when x_2 is never chosen), and

$$m_3(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{(1-p_1)\int_{\max(0,t_{3,3,2})}^{\infty} t\phi(t)dt + p_1 \int_0^{\max(0,\min(t_{1,2,3},t_{1,1,3}))} t\phi(t)dt}{(1-p_1)(1 - \Phi(\max(0, t_{3,3,2}))) + p_1\Phi(\max(0, \min(t_{1,2,3}, t_{1,1,3})))}.$$

Next define some convenient notation:

$$\underline{\underline{H}} \equiv \min_c \frac{(1-p_1)\bar{t} + p_1 \int_0^c t\phi(t)dt}{1 - p_1 + p_1\Phi(c)}.$$

Then, we can establish that each $\hat{t}_{1,j,k}$ must fall within a finite interval, using the maximum and minimum values of the image functions above. In particular,

$$\hat{t}_{1,1,2} \in \left[\frac{-s\overline{H} + v_2 - v_1}{G_1}, \frac{s(\overline{H} - H(\bar{t})) + v_2 - v_1}{G_1}\right],$$

$$\hat{t}_{1,1,3} \in \left[\frac{s(\overline{H} - H(\bar{t})) + v_3 - v_1}{G_2}, \frac{s(\underline{\underline{H}} - \overline{H}) + v_3 - v_1}{G_2}\right],$$

and

$$\hat{t}_{1,2,3} \in \left[\frac{s\overline{H} + v_3 - v_2}{G_2 - G_1}, \frac{s(\underline{\underline{H}} - \overline{H}) + v_3 - v_2}{G_2 - G_1}\right].$$

Since each of these intervals is finite, the range of $\hat{T} = (\hat{t}_{1,1,2}, \hat{t}_{1,1,3}, \hat{t}_{1,2,3})$ is a compact, convex subset of $\mathbb{R}^3$. Call this set D. Since $\hat{T}$ is also defined to be continuous, by Brouwer's fixed point theorem, we know that $\hat{T}$ has a fixed point within D.

This fixed point will satisfy the six equations necessary for either a type 1, type 2, or type 3 equilibrium, but is not guaranteed to satisfy the inequalities relating $t_{1,1,2}$, $t_{1,1,3}$, and $t_{1,2,3}$ which guarantee that these three parameters describe a state in which preferences are transitive. Restricting attention to the subset of D corresponding to feasible preferences prevents us from appealing to Brouwer's fixed point theorem, as this subset is not convex; for example, while $(\tilde{t}_{1,1,2}, \tilde{t}_{1,1,3}, \tilde{t}_{1,2,3}) = (5, 4, 1)$ falls in the category of type 1 equilibria, and $(-1, 4, 5)$ falls in case 2, the midpoint between these values, $(2, 4, 3)$, leads to intransitive preferences.

However, we can show that the image of *any* point in D under $\hat{T}$ leads to transitive preferences. Transitive preferences arise when either $t_{1,1,2} > t_{1,1,3} > t_{1,2,3}$, or when $t_{1,2,3} > t_{1,1,3} > t_{1,1,2}$. But notice that

$$\begin{aligned}
& t_{1,1,2} > t_{1,1,3} \\
\iff & \frac{s(H_2 - H_1) + v_2 - v_1}{G_1} > \frac{s(H_3 - H_1) + v_3 - v_1}{G_2} \\
\iff & s\left(\frac{H_2}{G_1} - \frac{H_3}{G_2} - \frac{(G_2 - G_1)H_1}{G_1 G_2}\right) + \frac{v_2}{G_1} - \frac{v_3}{G_2} - \frac{(G_2 - G_1)v_1}{G_1 G_2} \\
\iff & s\left(\frac{H_2}{G_2 - G_1} - \frac{H_3}{G_2(G_2 - G_1)} - \frac{H_1}{G_2}\right) + \frac{v_2}{G_2 - G_1} - \frac{v_3}{G_2(G_2 - G_1)} - \frac{v_1}{G_2} > 0 \\
\iff & \frac{s(H_3 - H_1) + v_3 - v_1}{G_2} > \frac{s(H_3 - H_2) + v_3 - v_2}{G_2 - G_1}
\end{aligned}$$

$$\iff t_{1,1,3} > t_{1,2,3}.$$

That is, no matter what relation two components of $\hat{T}$ take towards each other, the third is guaranteed to fall in the range required for rational preferences. In other words, while D is a convex, compact subset of $\mathbb{R}^3$, $\hat{T}(D) \subset D$ is the nonconvex subset containing only points that lead to rational preferences. Therefore, whatever the fixed point of $\hat{T}$ is on D, it describes a valid equilibrium of one of the three types described above. This completes the proof.

Proposition 5 part 2: *If $X = \{x_1, x_2, x_3\}$ with $G(x_2 - x_1) = G(x_2 - x_3) = G_1$ and $G(x_1 - x_3) = G(x_3 - x_1) = G_2 > G_1$, and ρ is uncorrelated with t, then the following holds:*

2. *For approval seekers, a unique equilibrium exists of the same form as described for respect seekers.*

Proof: Any equilibrium must be of the form described in the proof of Proposition 5 part 1, as that argument does not depend on the definition of H_{rs} compared to H_{as}. The unique equilibrium trivially exists as the response of each type to fixed, exogenous factors in their optimization problem.

Proposition 6: *If $X = \{x_1, x_2, x_3\}$ with $G(x_2 - x_1) = G(x_2 - x_3) = G_1$ and $G(x_1 - x_3) = G(x_3 - x_1) = G_2 > G_1$, and ρ is uncorrelated with t, then the following hold:*

1. *As social pressure rises, in the limit, all respect seekers adhere closer to their own norms, with only very low integrity types with one norm defecting to the other norm and no one choosing the compromise option.*

2. *As social pressure rises, in the limit, either all approval seekers pool on the majority's ideal action x_1 or x_3, or they all compromise by choosing x_2.*

Proof: Part 1 is a secondary conclusion of the proof of Proposition 5 part 1.

For part 2, note that the social image of each action is fixed: $m(x_1) = -(1 - p_1)G_2$, $m(x_2) = -G_1$, and $m(x_3) = -p_1 G_2$. Any of these quantities may be smallest (i.e. most negative), and as s increases, $H(m(x))$ becomes the overwhelming factor in each person's decision. Therefore, in the limit, everyone pools on the action with the least negative image. Note that this is a substantive difference from lower levels of social pressure since, as in Proposition 5, all five types of equilibria exist at low s.

Proposition 7: *If individuals are motivated by both approval and respect, $X = \{x_1, x_2\}$, and ρ and t are uncorrelated, the following hold:*

1. *Low-t types choose* $\arg\max v(x_i)$ *for sufficiently small s_{as} but choose* $\arg\max H_{as}(m_{as}(x_i))$ *for larger s.*

2. *At a fixed level of s_{as}, increasing s_{rs} in the limit leads to perfect adherence to personal norms.*

3. *At a fixed level of s_{rs}, increasing s_{as} in the limit leads to perfect conformity with majority opinion.*

4. *As s_{as} and s_{rs} simultaneously increase, in the limit, equilibrium can take* either *of the two forms above.*

Proof:

Part 1 is true by Propositions 1 and 2, since the quantity $v(x_i)$ in those results is replaced in this setting with $v(x_i) + s_{as}H_{as}(m_{as}(x_i))$. At small s_{as}, the first term dominates, and at higher s_{as}, the latter dominates.

Similarly to part 1, parts 2 and 3 follows from Propositions 1 and 2.

Following the same intermediate value theorem argument of the proof of Proposition 1 part 2, with an appropriately modified definition of $\hat{t}$, shows that when s_{as} is large enough that $v_2 - v_1 + s_{as}(H_{as}(m_2) - H_{as}(m_1)) > 0$, $\hat{t}$ must fall within $[0, \infty]$. Along with the previous two parts, this proves part 4.

Proposition 8 part 1: *In the endogenous norm game, the following holds:*

1. *For respect seekers, equilibria take one of four forms: a pooling equilibrium in which everyone chooses x_2 and y_2 and ρ_2 always, a pooling equilibrium in which everyone chooses ρ_1, x_1 and y_1, an equilibrium in which sufficiently high types choose ρ_1, x_1 and y_1 while lower types choose x_2 over x_1, and an equilibrium in which sufficiently high types choose ρ_2, x_2 and y_2 always, while lower types choose y_1 over y_2. 1.*

Proof: This proof examines eight cases in order to determine whether equilibria exist for those cases. In each choice set either $\tilde{t}_1$ or $\tilde{t}_2$ is positive, and the positive cutoff value for the x choice set might be larger or smaller than the positive cutoff value for the y choice set, so there are 8 possible cases corresponding to the relationships between cutoff values. Recall the definitions of the cutoff values:

$$\tilde{t}_{1,x} = \frac{s(H(m(x_2)) - H(m(x_1))) + v(x_2) - v(x_1)}{G}$$

$$\tilde{t}_{2,x} = \frac{s(H(m(x_1)) - H(m(x_2))) + v(x_1) - v(x_2)}{G} = -\tilde{t}_{1,x}$$

$$\tilde{t}_{1,y} = \frac{s(H(m(y_2)) - H(m(y_1))) + v(y_2) - v(y_1)}{G}$$

$$\tilde{t}_{2,y} = \frac{s(H(m(y_1)) - H(m(y_2))) + v(y_1) - v(y_2)}{G} = -\tilde{t}_{1,y}$$

Given $\tilde{t}_{i,j}$, type ρ_i chooses j_i *iff* $t > \tilde{t}_{i,j}$ $\forall$ $i \in \{1, 2\}$ and $j \in \{x, y\}$.

In all cases, I determine what equilibria are possible, and find the conditions on the magnitudes of the cutoff values and beliefs about off-equilibrium path outcomes that must hold (relative to the model parameters specifying consumption utility, etc, of course) for such an equilibrium to exist.

First consider low types. Their integrity is low enough that their choices are not affected by their choice of ideal. If their choices are aligned along a particular ideal, they will strictly prefer to choose the corresponding ideal. If not, they are indifferent between choices of ideal, because either way, half of their choices will induce guilt. Therefore, simply based on the cutoff values, we immediately know the choices and ideals of low types.

Now consider high types. High types will always choose in accordance with their ideal. They will prefer to choose $\rho = \rho_1$ *iff*, in expectation, those choices lead to a higher utility. That is, *iff*

$$sH(m(x_1)) + v(x_1) + sH(m(y_1)) + v(y_1) > sH(m(x_2)) + v(x_2) + sH(m(y_2)) + v(y_2). \quad \text{(B.1)}$$

Rearranging reveals that this condition is equivalent to $\tilde{t}_{2,x} > \tilde{t}_{1,y} \Leftrightarrow \tilde{t}_{1,x} < \tilde{t}_{2,y}$.

Lastly, consider middle types. They will always behave like high types, which can be seen by going through the algebra of two key examples and generalizing that to all cases. Suppose, first, that $t_{1,x} > t_{1,y} > 0$. Then middle types will choose x_2 and y_2 if they choose ρ_2, but will choose x_2 and y_1 if they choose ρ_1. They will prefer to choose ρ_1 *iff* $sH(m(y_1)) + v(y_1) - tG > sH(m(y_2)) + v(y_2) \Leftrightarrow t \leq \tilde{t}_{2,y}$. This is false by hypothesis, since $\tilde{t}_{2,y} = -\tilde{t}_{1,y} < 0$. They will therefore choose the ideal that leads to them comply perfectly with it, and therefore will act like high types.

On the other hand, suppose that $t_{1,x} > t_{2,y} > 0$. The middle types with ρ_1 will choose y_1 and x_2 and with ρ_2 will choose x_2 and y_2. They prefer ρ_1 *iff* $sH(m(y_1)) + v(y_1) - tG > sH(m(y_2)) + v(y_2) \Leftrightarrow t \leq \tilde{t}_{2,y}$, which is again false by hypothesis. Middle types will act like high types.

These two cases generalize to all possibilities by relabeling the cutoff values in the argument above, so altogether we have that high and middle types choose ρ_1 *iff* $\tilde{t}_{2,x} > \tilde{t}_{1,y} \Leftrightarrow \tilde{t}_{1,x} < \tilde{t}_{2,y}$ and low types will either choose the ideal that allows them to be consistent but selfish, if it exists, or will be indifferent between ideals.

I will rely on these insights about low and high types in each of the cases below.

1. $\tilde{t}_{1,y} > \tilde{t}_{1,x} > 0$: In this case, by the above reasoning, low types will choose ρ_2, x_2, and y_2. High and middle types must choose the same, since $\tilde{t}_{2,y} = -\tilde{t}_{1,y} < 0 < \tilde{t}_{1,x}$. If all types choose x_2 and y_2, $m(x_2) = m(y_2) = \bar{t}$, then $m(x_1)$ and $m(y_1)$ represent off-equilibrium-path beliefs that must support equilibrium. The conditions for this equilibrium to hold are that the cutoff values satisfy the hypothesized relationship. If they do, optimality of choices and beliefs for all types follow immediately by definition of the cutoff values, as described above. The two conditions are, therefore: 1) $\tilde{t}_{1,y} > \tilde{t}_{1,x} \Leftrightarrow H(m(x_1)) > H(m(y_1)) + \frac{v(x_2)-v(x_1)+v(y_1)-v(y_2)}{s}$, and 2) $\tilde{t}_{1,x} > 0 \Leftrightarrow H(m(x_1)) < H(\bar{t}) + \frac{v(x_2)-v(x_1)}{s}$.

2. $\tilde{t}_{1,x} > \tilde{t}_{1,y} > 0$: In this case, all types types will still choose ρ_2, x_2, and y_2. In order for the hypothesized cutoff values to be in the given relation to each other, off-equilibrium path beliefs $m(x_1)$ and $m(y_1)$ must satisfy: 1) $\tilde{t}_{1,x} > \tilde{t}_{1,y} \Leftrightarrow H(m(x_1)) < H(m(y_1)) + \frac{v(x_2)-v(y_1)+v(y_1)-v(y_2)}{s}$, and 2) $\tilde{t}_{1,y} > 0 \Leftrightarrow H(m(y_1)) < H(\bar{t}) + \frac{v(y_2)-v(y_1)}{s}$.

3. $\tilde{t}_{1,y} > \tilde{t}_{2,x} > 0$: In this case, low types always choose x_1 and y_2 and are indifferent between ideals. Middle and high types always choose ρ_2, x_2 and y_2. With these choices, conditional expectations determine that $m(x_2) > m(y_2) = \bar{t} > m(x_1)$. But then we must have that $\tilde{t}_{1,x} > 0$, contradicting our hypothesis. Therefore no equilibrium is possible in this case.

4. $\tilde{t}_{2,x} > \tilde{t}_{1,y} > 0$: In this case, low types also always choose x_1 and y_2 and are indifferent between ideals, while middle and high types always choose ρ_1, x_1, and y_1. We therefore have rational expectation image values such that $m(y_1) > m(x_1) = \bar{t} > m(y_2)$. But then $\tilde{t}_{2,y} > 0$, contradicting our hypothesis. There is therefore no equilibrium possible in this case.

5. $\tilde{t}_{2,y} > \tilde{t}_{1,x} > 0$: In this case low types always choose x_2 and y_1 and are indifferent between ideals. Middle and high types always choose ρ_1, x_1, and y_1. We therefore have rational expectation image values such that $m(x_1) > m(y_1) = \bar{t} > m(x_2)$ and $m(y_2)$ is off-equilibrium. This and the cutoff values must satisfy the following two conditions for this equilibrium to exist (in which $m(x_1)$ and $m(x_2)$ are determined from Bayesian reasoning): 1) $\tilde{t}_{2,y} > \tilde{t}_{1,x} \Leftrightarrow H(m(y_2)) < H(\bar{t}) + H\left(\frac{\int_{\tilde{t}_{1,x}}^{\infty} t\phi(t)dt}{1-\Phi(\tilde{t}_{1,x})}\right) - H\left(\frac{\int_{0}^{\tilde{t}_{1,x}} t\phi(t)dt}{\Phi(\tilde{t}_{1,x})}\right) + \frac{v(y_1)-v(y_2)+v(x_1)-v(x_2)}{s}$, and 2) $\tilde{t}_{1,x} > 0 \Leftrightarrow H\left(\frac{\int_{\tilde{t}_{1,x}}^{\infty} t\phi(t)dt}{1-\Phi(\tilde{t}_{1,x})}\right) < H\left(\frac{\int_{0}^{\tilde{t}_{1,x}} t\phi(t)dt}{\Phi(\tilde{t}_{1,x})}\right) + \frac{v(x_2)-v(x_1)}{s}$.

6. $\tilde{t}_{1,x} > \tilde{t}_{2,y} > 0$: In this case, low types always choose x_2 and y_1 and are indifferent between ideals, while middle and high types choose ρ_2, x_2, and y_2. We therefore have rational expectation image values such that $m(y_2) > m(x_2) = \bar{t} > m(y_1)$ and $m(x_1)$ is off-equilibrium. This and the cutoff values must satisfy the following two conditions for this equilibrium to exist: 1) $\tilde{t}_{1,x} > \tilde{t}_{2,y} \Leftrightarrow H(m(x_1)) < H(\bar{t}) + H\left(\frac{\int_{\tilde{t}_{2,y}}^{\infty} t\phi(t)dt}{1-\Phi(\tilde{t}_{2,y})}\right) - H\left(\frac{\int_{0}^{\tilde{t}_{2,y}} t\phi(t)dt}{\Phi(\tilde{t}_{2,y})}\right) + \frac{v(y_2)-v(y_1)+v(x_2)-v(x_1)}{s}$, and 2) $\tilde{t}_{2,y} > 0 \Leftrightarrow H\left(\frac{\int_{\tilde{t}_{2,y}}^{\infty} t\phi(t)dt}{1-\Phi(\tilde{t}_{2,y})}\right) < H\left(\frac{\int_{0}^{\tilde{t}_{2,y}} t\phi(t)dt}{\Phi(\tilde{t}_{2,y})}\right) + \frac{v(y_1)-v(y_2)}{s}$.

7. $\tilde{t}_{2,y} > \tilde{t}_{2,x} > 0$: In this case, low types choose x_1 and y_1 and strictly prefer ρ_1. Middle and high types also choose ρ_1, x_1, and y_1. Therefore $m(x_1) = m(y_1) = \bar{t}$. Off-equilibrium beliefs must satisfy the two conditions: 1) $\tilde{t}_{2,y} > \tilde{t}_{2,x} \Leftrightarrow H(m(x_2)) > H(m(y_2)) + \frac{v(x_1)-v(x_2)+v(y_2)-v(y_1)}{s}$, and 2) $\tilde{t}_{2,x} > 0 \Leftrightarrow H(m(x_2)) < H(\bar{t}) + \frac{v(x_1)-v(x_2)}{s}$.

8. $\tilde{t}_{2,x} > \tilde{t}_{2,y} > 0$: In this case, low types choose x_1 and y_1 and strictly prefer ρ_1. Middle and high types choose ρ_1, x_1, and y_1. Therefore $m(x_1) = m(y_1) = \bar{t}$, and off-equilibrium

beliefs must satisfy the two conditions: 1) $\tilde{t}_{2,x} > \tilde{t}_{2,y} \Leftrightarrow H(m(x_2)) < H(m(y_2)) + \frac{v(x_1)-v(x_2)+v(y_2)-v(y_1)}{s}$, and 2) $\tilde{t}_{2,y} > 0 \Leftrightarrow H(m(y_2)) < H(\bar{t}) + \frac{v(y_1)-v(y_2)}{s}$.

These eight cases all fall into the four categories described in the proposition statement, completing the proof.

Proposition 8 part 2: *In the endogenous norm game, the following holds:*

2. For approval seekers, there exists an equilibrium of one of these same four types described in Proposition 8 part 1.

Proof: The demonstration of possible forms of equilibria in the proof of Proposition 8 part 1 isn't specific to respect seekers. Uniqueness follows as before: each type has a unique best response to their exogenous optimization problem parameters, which constitutes the equilibrium.

Proposition 9 part 1: *In the endogenous norm game, the following holds:*

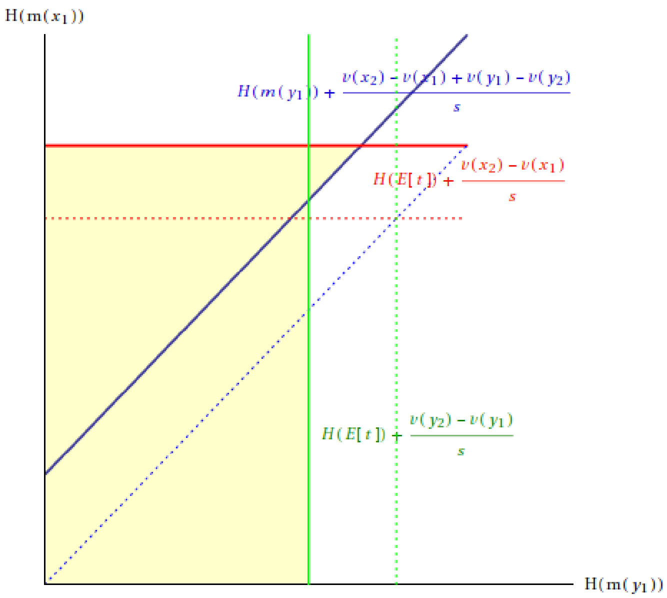
1. For respect seekers, at sufficiently small $s < \frac{v(y_1)-v(y_2)}{\Delta H}$, there is an equilibrium in which all types choose ρ_2 and some individuals defect to y_1. For sufficiently large $s > \frac{v(y_1)-v(y_2)}{H(\bar{t})}$, there is an equilibrium in which all types choose and comply perfectly with ρ_2. One of these equilibria always exists. For higher $x > \frac{v(x_2)-v(x_1)}{H(\bar{t})}$, there is additionally an equilibrium in which all types choose and comply perfectly with ρ_1. For mid-range $s \in \left[\frac{v(x_2)-v(x_1)+v(y_2)-v(y_1)}{H(\bar{t})+\Delta H}, \frac{v(x_2)-v(x_1)}{\Delta H}\right]$, there is an equilibrium in which all types choose ρ_1 but some individuals defect to x_2.

Proof: The conditions in case 1 and 2 in the proof of Proposition 8 part 1 determine the conditions necessary for a pooling equilibrium on ρ_2, x_2 and y_2. We can graph these conditions and determine when the appropriate region is nonempty, as in figure B.2. This figure shows three solid lines corresponding to bounds on off-equilibrium path beliefs, and the dotted lines of the same color represent the limit of these boundaries as $s \to \infty$. Additionally, it's not possible for $H(m)$ to be smaller than 0 or larger than $\overline{H}$, so we must take these bounds into account as well.

The upper shaded triangle between the blue and red lines represent the region of beliefs that support the equilibrium described in case 1 in the proof of Proposition 8 part 1. The lower shaded trapezoid between the green and blue lines are beliefs that support the equilibrium of case 2. The upper region is nonempty so long as the red bound intersects the vertical axis above the blue bound, and the blue bound intersects below $\overline{H}$. (Note that the red bound must intersect above 0 because the intercept approaches $H(\bar{t}) > 0$ from above.) These conditions are satisfied if $s > \frac{v(y_1)-v(y_2)}{H(\bar{t})}$ and if $s > \frac{v(x_2)-v(x_1)+v(y_1)-v(y_2)}{\overline{H}}$.

On the other hand, the case 2 equilibrium region is nonempty so long as the green line intersects the horizontal axis above 0. This occurs when $s > \frac{v(y_1)-v(y_2)}{H(\bar{t})}$, so this is the only condition needed for a pooling equilibrium on ρ_2, x_2, and y_2.

Figure B.2: Supporting Off-equilibrium Beliefs for ρ_2 Pooling Equilibrium



For the pooling equilibrium on ρ_1, x_1, and y_1, we can draw a similar figure based on the conditions in cases 7 and 8, which shows that the equilibrium can be supported *iff* $s > \frac{v(x_2)-v(x_1)}{H(\bar{t})}$.

For the partial pooling equilibrium on ρ_2 in which some types defect to y_1, case 6 above describe the necessary conditions on the cutoff values and the off-equilibrium beliefs $m(x_1)$. The first condition regards the off-equilibrium belief, requiring that $H(m(x_1)) < H(\bar{t}) + H(m(y_2)) - H(m(y_1)) + \frac{v(y_2)-v(y_1)+v(x_2)-v(x_1)}{s}$. The right hand side is strictly positive, so this is always satisfiable within the allowed range. The second condition requires that $H(m(y_2)) - H(m(y_1)) < \frac{v(y_1)-v(y_2)}{s}$. The LHS is always at least $\underline{\Delta H}$, so for equilibrium to exist we must have $s < \frac{v(y_1)-v(y_2)}{\underline{\Delta H}}$.

Case 5 describe the conditions for a partial pooling equilibrium on ρ_1 in which some types defect to x_2. These conditions can be combined as $H(m(y_2)) - H(\bar{t}) + \frac{v(x_2)-v(x_1)+v(y_2)-v(y_1)}{s} < H(m(x_1)) - H(m(x_2)) < \frac{v(x_2)-v(x_1)}{s}$. The off-equilibrium belief $m(y_2)$ can be simply set to 0, since anytime these inequalities hold they will also hold with $m(y_2) = 0$. Then we need the difference in image utilities of choosing x_1 and x_2 to fall within the appropriate range. In general, the lower and upper bounds of this difference depend on the distribution, but for mid-range values of s the equilibrium will exist.

Proposition 9 part 2: *In the endogenous norm game, the following holds:*

2. For approval seekers, at sufficiently small s, there is a unique equilibrium in which

> *everyone chooses ρ_2 and complies with it, perhaps imperfectly with low t individuals defecting to y_1. For higher s, there is an additional possible equilibrium in which everyone chooses ρ_1 and complies with it, perhaps imperfectly with low t individuals defecting to x_2.*

This is a simple outcome from a maximization problem. At small s, the benefit of believing in and following ρ_2 outweighs any image benefit of following the crowd, so the pooling equilibrium on ρ_1 isn't sustainable. At large enough s, people prefer to follow the crowd whichever ideal the crowd chooses, so both equilibria are possible.

Proposition 10: *If a group needs to set membership dues d to attract a fraction $q \leq p$ of the population, where p is the fraction of the population that is sympathetic to the group, then:*

1. *for approval seekers, d is strictly decreasing in s, and*
2. *for respect seekers, d is increasing in s over the range of s but* possibly *decreasing for small increases in s.*

Proof: This is a simple corollary of Propositions 1 and 2, with $p_1 = p$, $v_2 = 0$, and $v_1 = -d$.

Appendix C

Chapter 2 Experiment Details

C.1 Design

To test for extrinsic determinants of reciprocity-induced sharing, we conducted a reciprocity variant of the dictator game, a "double dictator game" (DDG). The DDG consists of two dictator games played consecutively, with the role of dictator and recipient switched in the second stage. The first dictator game is a "mini" dictator game over only \$2 in order to distinguish reciprocal motivations from distributional preferences. We compare the DDG to a standard dictator game (DG), and cross the design with sorting options. The six treatments are shown in table C.1. We employ a between-subjects design based on the design in Lazear, Malmendier, and Weber (2012) to compare games with and without sorting, holding constant the endowment. The design is similar to Dana, Cain, and Dawes (2006) and to Broberg, Ellingsen, and Johannesson (2007).

Table C.1: 3×2 Experiment Design

	Standard Dictator Game	Double Dictator Game	
		Kind choice in	Unkind choice in
		initial mini-game	initial mini-game
	(No Reciprocity)	(Positive Reciprocity)	(Negative Reciprocity)
No Sorting	**DG/NS**	**PR/NS**	**NR/NS**
Sorting	**DG/S**	**PR/S**	**NR/S**

All experiments were conducted at UC Berkeley. Subjects received a participation fee of \$5 and were informed that they might earn additional money. Half of the subjects were randomly assigned the role of recipient (and hence, in the DDG, the role of mini-dictator). The other half were assigned the role of dictators (and hence, in the DDG, of mini-recipients) and were moved to a separate room from the recipients, where they received their instructions.

Subjects were randomly assigned to their roles, and thus were randomly assigned to either positive or negative reciprocity treatments. DG treatments were conducted in separate sessions, and sorting and non-sorting treatments were also conducted in separate sessions at different times. Assignment to these treatments is thus not random; however, the same subject pool of UC Berkeley students and staff were solicited in each case, in the same manner, for experiments in the same lab, so it is likely that the samples are comparable.

DDG with and without sorting. The first-stage mini-dictators were told to divide $2 with a randomly-paired participant in the other room by circling one of two choices: keep $2 and give the paired participant $0; or keep $1 and give the paired participant $1. We offered only a binary choice (share $1 or $0) to generate clean reciprocity treatment assignments.

After these participants made their choices, participants in the other room were told about the first stage. After finding out how much of the $2 their paired participant shared, the matched partner played the second stage. In the variant without sorting, this was a $10 dictator game with the same person. Note that subjects did not know, initially, that the mini-game would be followed by a dictator game over $10 with the roles switched.[1] At the end of the experiment, the experimenter described the game to the participants in the other room and showed each of them how much money they received. Thus, recipients were guaranteed to know about the second-stage game and about the amount dictators had decided to share with them.

In the variant with sorting, second-stage dictators first decided whether or not to "participate." They received two envelopes, labeled "participate" and "don't participate." If they chose to participate, they opened the former envelope, which contained a sheet with the participant number of the paired recipient, and filled it out exactly as in the no sorting condition. If they chose not to participate, they opened the envelope marked "don't participate" (which did not contain a matched participant number) and wrote their participant number on the sheet inside. In that case, they received $10 without the option to divide the money.

In this variant, participants in the other room only found out about the dictator game if their matched dictator had chosen to participate. After collecting the envelopes, the experimenter separated receivers matched with participating and non-participating dictators. Those matched with non-participating dictators received the $5 participation fee and the earnings from the mini-dictator game, and the experiment ended. Those paired with participating dictators completed the experiment as in the no sorting treatment, meaning they were informed about the dictator game and shown how much the dictator had shared.

A total of 192 pairs of subjects (54 without and 138 with sorting) participated in the DDG. In 89 cases (46 percent), the first mover shared the two dollars (26 cases [48 percent] without and 63 cases [46 percent] with sorting). This led to 89 dictators in the PR condition (63 with sorting and 26 without) and 103 in the NR condition (75 with sorting and 28 without).

[1] Initial voluntary sharing can thus be interpreted as an act of kindness, rather than an attempt to induce reciprocal behavior, which avoids concerns about interpreting the dictators' reactions as reciprocity.

DG with and without sorting. The benchmark for comparison of the reciprocity environment are baseline dictator game (DG) conditions with and without avoidance options. The procedures for the DG variants closely mirror those of the DDG variants, just without the first stage.[2] Details are described thoroughly in Lazear, Malmendier, and Weber (2012). Overall, 182 student subjects participated in the DG treatments, including 45 dictators in the no-sorting condition and 46 in the sorting condition.

C.2 Results

Before we turn to the main results on the effect of sorting in reciprocity environments, we briefly discuss the evidence of reciprocity in the conditions without sorting.

Figure C.1 presents the average amounts shared, without sorting (left, dark columns), with sorting treatments. In the middle, striped columns, those who opt out and thus share nothing are treated as having shared \$0, and the right, grey columns show averages conditional on sorting in. The three sets of bars show average levels of sharing in the three reciprocity conditions: the baseline DG (no reciprocity), positive (PR), and negative (NR) reciprocity. Focusing on the conditions without sorting (left columns in each set of bars), we see that the average amount shared by dictators in the standard (no-sorting) DG treatment is \$2.00 in our experiment. In the (no-sorting) DDG, the average amount shared by dictators increases to \$2.39 in the PR treatment, and it decreases to \$0.70 in the NR treatment. Table 2.1, Column 1, shows that only the negative-reciprocity effect is significant. In other words, there is a significant negative-reciprocity effect ($-.130$) and an insignificant positive-reciprocity effect (.039),[3] which is consistent with previous experimental evidence, e.g., weak positive reciprocity but strong "concern withdrawal" in Charness and Rabin (2002).

Note, though, that the above averages are based on the second-stage endowment of \$10 and, hence, assume "narrow bracketing" in the second-stage dictator game. An alternative measure of reciprocity adds the amount of \$2 from the mini-DG back to the analysis (see Cox (2004)), in which case the effect of positive reciprocity is marginally significant and the effect of negative reciprocity is insignificant.[4] The subsequent analysis on the effects of

[2] One difference to the DDG is that the instructions emphasized the \$10 dictator game (including the decision not to participate in case of the sorting treatment) was the last decision that anybody in either room would make in this experiment. This did not seem necessary in the standard DG, consistent with how the DG is conducted in prior research.

[3] Standard errors are robust to heteroskedasticity and adjusted for small-sample bias, using the residual-variance estimator HC3, which approximates a jackknife estimator (MacKinnon and White 1985). If we cluster by session, standard errors in this and in all other estimations are very similar and typically slightly smaller, though unlikely to be reliable given the few clusters.

[4] Under this alternative approach, recipients end up with an average amount of \$1+\$2.39 = \$3.39 out of \$12 (28.3%) after sharing \$1, and with an average payoff of \$2 + \$0.70 = \$2.70 (22.5%) after sharing zero, compared to \$2 out of \$10 (20%) in the single dictator game. In this case, positive reciprocity induces a marginally significant increase in giving (t-statistic = 1.88, p-value = 0.06), and negative reciprocity does not have a significant effect. The lack of a significant negative-reciprocity effect reflects censoring at \$2: Dictators cannot reduce the amount obtained by recipients below \$2 if those kept the initial \$2.

sorting is unchanged when \$12 is used as the relevant endowment.

The Effect of Sorting on Sharing. We now turn to the question of what impact sorting has on sharing in the baseline DG and each of the reciprocity environments. As mentioned above, the average amount shared in the DG treatment without sorting is \$2.00, which is comparable to findings in previous experiments. Most subjects share a positive amount (64 percent). However, the introduction of sorting strongly decreases the average amount shared, to \$1.21, as shown in the left set of bars in Figure C.1. This decrease is statistically significant in a non-parametric rank-sum test ($z = 2.34$, $p = 0.02$). The sorting opportunity also decreases the frequency of sharing dramatically, to only 39% sharing a positive amount.

How does sorting affect sharing after positive or negative reciprocity has been induced? In Figure C.1, we see that positive reciprocity increases giving to an average level of \$2.39 without sorting (left bar in middle set). But, once again, sorting causes a large drop in average amounts shared, to \$1.71 (middle bar in middle set). In a negative reciprocity environment without sorting, sharing plummets to an average of \$.70, and the option to sort out further reduces sharing to \$.31 (left and middle bars in right set).

Figure 2.1 provides more details and shows the distribution of amounts shared in each condition. (We display the frequencies of subjects who opt out separately at the left.) We observe a sharp shift of the distribution to the left when sorting becomes possible, regardless of the reciprocity conditions. Hence, both the simple comparison of means and the distributional evidence suggest that sorting has a large impact on sharing even after reciprocity has been induced, inconsistent with a solely intrinsic motivation for reciprocal behavior.

Finally, Table 2.1 confirms the statistical significance of these findings, both in a linear regression (column 2) and in a tobit estimation (column 4). As before, standard errors are robust to heteroskedasticity and, in the linear regression, adjusted for small-sample bias.[5] Under both estimation procedures, sorting significantly reduces sharing.

In other words, givers who respond to a previous kind or unkind act are affected by the option to avoid the opportunity to give. This evidence suggest that the dominant approach to sharing under reciprocity, which relies on intrinsic factors, is incomplete. Extrinsic factors affect individuals' giving in reciprocity environments as well, inconsistent with the prevalent modeling approach.

The simple comparison of means suggests that the effect is smaller under positive reciprocity and larger under negative reciprocity than in the neutral DG setting. In Figure C.1, we see that the average amount shared decreases by 40 percent, from \$2.00 to \$1.21, in the DG condition; by 29 percent, from \$2.39 to \$1.71, in the PR condition; and by 56 percent, from \$0.70 to \$0.31, in the NR condition.

[5] Following MacKinnon and White (1985), we use the residual-variance estimator HC3 in the linear regressions, which approximates a jackknife estimator. In the tobit model, we perform a jackknife estimation, which produces slightly more conservative standard errors than the robust variance estimator.

The visual impression holds up to statistical testing. As Columns 3 and 5 of Table 2.1 reveal, the significant decrease due to sorting does not differ significantly, from the DG condition, in either the NR or PR conditions. The same picture emerges if we consider the frequency of sharing, i.e., the fraction of subjects who share any positive amount. The probit regression in the final column of Table 2.1 shows that 25 percent of sharers opt out, but the interactions of Sorting with either reciprocity condition, PR or NR, are statistically insignificant. In other words, the impact of sorting on average amounts shared is large and significant, and approximately invariant to the reciprocity setting. The structural results in Appendix C.3 will provide quantification of the role of intrinsic and extrinsic factors.

The Effect of Reciprocity on Sorting. The leftmost bars in each of the three graphs in Figure 2.1 reveal that the sorting option is used by a significant fraction of subjects in all treatments. In the DG condition, 50 percent of all subjects choose to opt out. In the PR condition, many still sort out, but fewer than in the DG baseline (32 percent). The NR condition incites more sorting behavior than the PR condition, but only slightly more than the baseline DG (59 percent). The differences between sorting under reciprocity and sorting in the neutral DG setting are either insignificant or only marginally significant. If we regress a dummy for "opting out" on a dummy for being in the PR treatment and a dummy for the NR treatment, as shown in column 1 of Table C.2, the intercept of .500 is highly significant, while the PR coefficient of -.183 is only marginally significant, and the NR coefficient of .087 is insignificant. The results are similar if we use a probit model, though the PR coefficient becomes more significant (marginal effect of .-186 with s.e. .094), as shown in column 2.

Intrinsic factors as the sole motivation for reciprocity are also threatened by our observation of "spiteful non-sharing," i.e., in the fraction of subjects who sort in but share nothing. As can be seen in Figure 2.1 and C.2, only a small number of subjects sort in and share zero in the DG (11 percent). Positive reciprocity reduces this rate to 3 percent (a 72 percent reduction). Meanwhile, in the NR condition, 20 percent of dictators sort in to share nothing (a 46 percent increase over DG).

As columns (3) and (4) of Table C.2 reveal, however, the differences are not significant in an OLS regression, and only the reduction after kind treatment (PR) is marginally significant in a probit estimation. However, the change in the rate of spiteful non-sharers between NR and PR is highly significant under both econometric models. Given the magnitude of the effect and the strength of the PR/NR comparison, and the fact that the effect is likely underestimated due to censoring at zero, we can infer that reciprocity has an impact on spiteful non-sharing.

The Distribution of Gifts. As a final avenue through which we can examine the impact of sorting we zoom into different areas of the distribution of shared amounts. As we saw already in Figure 2.1, the introduction of a sorting option significantly changes the distributions of positive gifts in all three conditions. The graphical evidence suggested a strong shift to the left in all three graphs.

But how does the sorting option affect different regions of the distribution? To investigate this question, we categorize dictators into three groups: zero-sharers (including those who

Figure C.1: Average Amounts Shared with and without Reciprocity

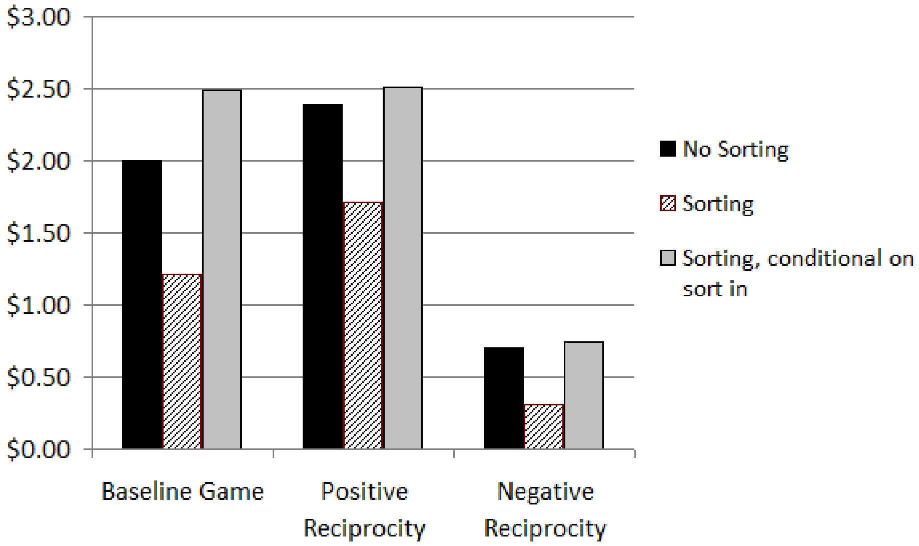
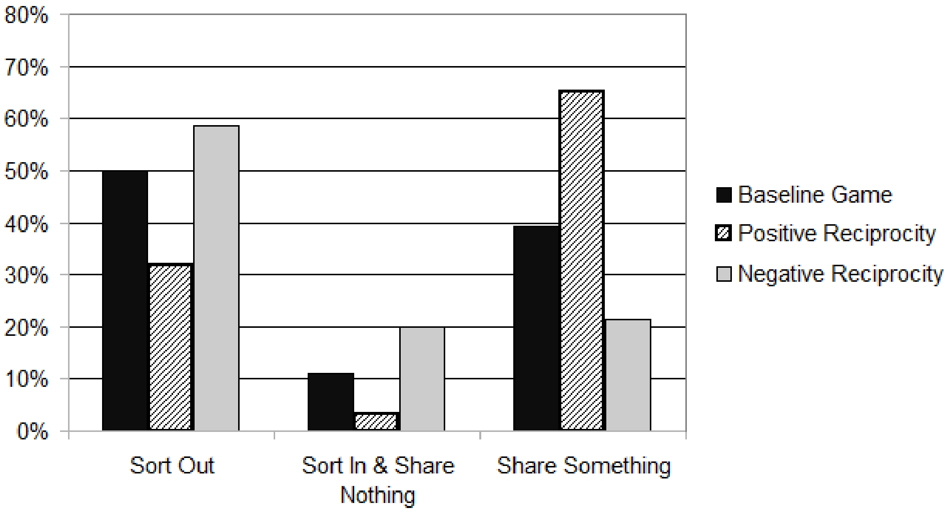


Figure C.2: Sorting In to Share Something or Nothing



opt out), low sharers (who share but less than 25 percent of the endowment), and high sharers (who share at least 25 percent). Across all conditions, 52 percent are zero-sharers, 26 percent are low sharers, and 22 percent are high sharers.[6] We then regress indicator variables for each of those groups on an indicator for the sorting option, separately for all three groups. As shown in Table C.3, the sorting option significantly increases the fraction of subjects who share nothing, by 20.9 percent (column 1). This is consistent with the strong sorting effect across treatments. We also find that the sorting option has no significant effect on the fraction of low sharers (column 3). This reflects the switch of low sharers to zero sharing, on the one hand, and the switch of high sharers to low sharing, on the other hand. Finally, the sorting option significantly reduces the fraction of high sharers, by 16.0 percent (column 5).

We also find that the effect of the sorting option in the three areas of the distribution does not vary significantly across the neutral, positive-reciprocity, and negative-reciprocity settings. Columns (2), (4), and (6) show that the interaction of the Sorting indicator and dummies for the positive-reciprocity and the negative-reciprocity settings are statically insignificant or at most (in one case) marginally significant. In other words, the effect of the sorting option in different areas of the distribution appears to be surprisingly similar across different treatments, which is hard to reconcile with the leading theories of reciprocal behavior.

C.3 Structural Estimation

We estimate the parameters μ_{α_r}, σ_{α_r}, μ_{β_r} and σ_{β_r} for each $r \in \{\text{DG,NR,PR}\}$, using a minimum distance estimator given by $(m(\theta) - M)'W(m(\theta) - M)$, where M is the vector of true moments given by the data, $m(\theta)$ is the vector of moments predicted by the theory at vector of parameters θ, and W is the weighting matrix. For W, we use the diagonal of the inverse of the variance-covariance matrix.[7] For the vector of moments, we break down the choices of giving into bins: exactly 0, from 25 cents to \$2.50, from \$2.75 to \$4.75, exactly \$5, and more than \$5. In the sorting conditions, an additional moment specifies the fraction who sort out. Altogether, we have 11 moments in each reciprocity environment, or 33 total. In the baseline estimations, we also assume that α_r is normally distributed according to $N(\mu_{\alpha_r}, \sigma_{\alpha_r})$ and β_r is similarly distributed according to $N(\mu_{\beta_r}, \sigma_{\beta_r})$. We will vary both the bin sizes and the distributional assumptions below.

The theoretical moments are simulated in Matlab using adaptive Simpson quadrature for numerical integration, implemented as the *quad* routine. An individual i with type parameters α_i and β_i in a particular reciprocity environment will share $x^s = 5+1/(2(\alpha_i+\beta_i))$ (or the closest element of the discrete choice set) if he cannot opt out; he will sort in and share x^s even if he can opt out if $U(x^s) > 10 - 25\alpha$ (and otherwise will sort out). This

[6] Results do not substantially change if slightly different category definitions are used.

[7] The same pattern of results emerges if using the identity matrix as the weighting matrix instead, but we omit these robustness checks for brevity.

threshold allows us to simply integrate over the distribution of types within the respective intervals to calculate the total fraction that fall within each choice category.

We determine the vector of parameters $\hat{\theta}$ that minimizes the distance estimator using *fmincon*, which is Matlab's implementation of Powell's (1983) sequential quadratic programming algorithm. We impose the constraints that σ_{α_r} and σ_{β_r} are positive and at most 20, and that μ_{α_r} and μ_{β_r} are between -20 and 20.[8] To make sure to find the global minimum, rather than a local minimum, we choose starting points randomly from a uniform distribution on the allowable ranges and run *fmincon* on 5 to 120 of these starting point vectors, depending on the model specification, until the best estimates were clearly in concurrence. The best estimate is typically found in at least half of the runs.

The minimum distance estimate is typically asymptotically normal, with a variance of $(\hat{G}'W\hat{G})^{-1}(\hat{G}'W\hat{\Lambda}W\hat{G})(\hat{G}'W\hat{G})^{-1}/N$, where $\hat{\Lambda} \equiv Var(m(\hat{\theta}))$ and $\hat{G} \equiv \frac{1}{N}\sum_{i=1}^{N} \nabla_{\theta} m_i(\hat{\theta})$ (Wooldridge 2002). We calculate $\nabla_{\theta} m_i(\hat{\theta})$ numerically using an adaptive finite difference algorithm.

Column 1 of Table C.4 shows the results, which are described in the main text. Columns 2 through 7 show that these findings are robust to a number of important robustness checks. In Column 2 we show the results of an estimation using a larger set of 39 moments. We break the choice set into a finer categorization, namely, exactly \$0, 25 cents to \$1.50, \$1.75 to \$3, \$3.25 to \$4.75, \$5 exactly, or more than \$5. The pattern of results is very similar, with μ_{α} generally being slightly more negative and μ_{β} being more positive.

We also estimate specifications were we alter the assumption that α and β are normally distributed, and assume a uniform distribution, with $\alpha_i \in [\mu_{\alpha} - \sigma_{\alpha}\sqrt{3}, \mu_{\alpha} + \sigma_{\alpha}\sqrt{3}]$ and $\beta_i \in [\mu_{\beta} - \sigma_{\beta}\sqrt{3}, \mu_{\beta} + \sigma_{\beta}\sqrt{3}]$, within each reciprocity environment. These are shown in columns 3 and 4 for the model with 33 and with 39 moments respectively.

As another robustness check, we also alter the model specification

$$U_r(x) = x + (\alpha_r + \beta_r \mathbb{1}(\text{sort in}))(f - (x-5)^2). \tag{C.1}$$

In the baseline specification, $f = 0$. The alternate specification allows for $f > 0$, which means that extrinsic factors can increase the giver's utility, e.g., in the form of pride, when giving at least a threshold amount. We estimate this model with $f = 5$ (column 5) and $f = 15$ (column 6), which implies that giving at least \$2.76 or \$1.13, respectively, is sufficient to feel pride or to be viewed favorably by your peers.[9] The results are again very similar for the specification with $f = 5$. Mean intrinsic motivation is lower than in the baseline specification, very similar to the estimation with 39 moments in column 2, though the standard deviation becomes smaller. Extrinsic motivation is also similar to prior estimates, and at the higher end. The NR estimates closely resemble those in the baseline estimation, and the PR estimates resemble most closely those in the model with 39 moments. In estimates in the specification

[8] These bounds are imposed in order to speed convergence. We verify that the global minimum lies within that range with runs using much larger bounds.

[9] Note that f, which in principle could be identified separately, is not well identified in practice within our dataset, thus motivating our tests using a set of particular values.

with $f = 15$ is somewhat different. As even low amounts of giving are assumed to suffice to generate positive image, the model tends to dismiss internal altruism as the main determinant of giving – it becomes significantly negative in all settings. The estimates of external motives, instead, become quite a bit larger, sometimes doubling in size. The main insight, however, is that the magnitudes remain comparable even under this somewhat stark assumption about external factors.

Finally, in column 7, we alter the G and H functions yet again to allow for sharing of more than half the endowment. The utility function now becomes

$$U_r(x) = x - (\alpha_r + \beta_r \mathbb{1}(\text{sort in}))x^2. \tag{C.2}$$

This effectively redefines α and β relative to a new maximum level of altruism, so we cannot compare the magnitudes of the estimates to the other specifications. Their relative levels (between treatments), however, are interpreted similarly. This change also requires a different breakdown of the data into moments, since a set of measure 0 is predicted to share exactly 5. We combine sharing exactly 5 and sharing more than 5, for a total of 27 moments instead of 33.

When comparing the estimates of this baseline model to models requiring β to be reciprocity invariant, as described in the main text, we also vary the specification and assumptions of the model, as we did in Table C.4. In all estimations, we replicate both patterns: requiring beta to be constant only slightly hurts the predictive power of the model, but requiring beta to be zero drastically reduces it. (We omit these results for brevity.)

For completeness, Table C.5 shows the actual moments along with the predicted moments under the three models of Table 2.3. These detailed predictions mostly just break down the overall pattern of results described above, but we note two details that could lead to future refinements of reciprocity models. First, our distributional assumptions make it difficult for the model to match the fraction of small gifts, simply because the weight on the functions inducing sharing ($\alpha + \beta$) must lie between .1026 and .2105 for the dictator to give between 25 cents and \$2.50, but only has to fall between .2105 and 4 to induce sharing between \$2.75 and \$4.75. Hence, it might be of value to explore alternative (more exotic) distribution or non-parametric estimations that allow for bunching to capture actual giving choices. In our contexts, standard distribution functions sufficed to generate robust results. Second, the data from the positive reciprocity conditions appears to fit the model less well than the other two reciprocity conditions, mainly due to the additional jump in small gifts up to \$1. This could reflect a desire to return the \$1 previously shared by the recipient, in line with Sugden's (1984) view of reciprocity as an obligation to return favors at a minimal level. We omit such an *ad hoc* adjustment to the model for now, but a more detailed study of the function form of extrinsic motives for reciprocal sharing might require a more detailed exploration of this pattern.

Table C.2: Rates of Sorting and Spiteful Non-sharing, Sorting conditions

Model:	**OLS**	**Probit**	**OLS**	**Probit**
Dependent Variable:	Sorted Out		Spiteful Non-sharer	
	(1)	(2)	(3)	(4)
Constant	0.500^{***}		0.109^{**}	
	(0.075)		(0.047)	
Positive Reciprocity (PR)	-0.183^{*}	-0.186^{**}	-0.077	-0.097^{*}
	(0.096)	(0.094)	(0.052)	(0.052)
Negative Reciprocity (NR)	0.087	0.087	0.091	0.072
	(0.095)	(0.094)	(0.066)	(0.057)
PR-NR			-0.168^{***}	-0.168^{***}
			(0.001)	(0.004)
Observations	184	184	184	184
(pseudo) R^2	0.055	0.041	0.050	0.076

Notes: "Sorted out" is a dummy variable for choosing the sorting out option in the sorting conditions; "spiteful non-sharer" is a dummy variable for a participant who chooses to sort in but shares nothing. Independent variables are condition group dummies. The probit model shows marginal effects. Robust standard errors are in parentheses (with bias-correction (HC3) in the linear case, see MacKinnon and White (1985).

* - $p < 0.1$; ** - $p < 0.05$; *** - $p < 0.01$

Table C.3: Zero, Low, and High Sharers: Effects of Reciprocity and Sorting

Dependent Variable:	Zero sharers		Low sharers		High sharers	
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.384^{***}	0.356^{***}	0.293^{***}	0.222^{***}	0.323^{***}	0.422^{***}
	(0.049)	(0.073)	(0.046)	(0.063)	(0.048)	(0.075)
Positive Reciprocity		-0.240^{**}		0.239^{**}		0.001
		(0.098)		(0.120)		(0.126)
Negative Reciprocity		0.323^{***}		0.028		-0.351^{***}
		(0.117)		(0.106)		(0.091)
Sorting	0.209^{***}	0.253^{**}	−0.048	−0.027	-0.160^{***}	-0.227^{***}
	(0.061)	(0.104)	(0.056)	(0.087)	(0.055)	(0.096)
Sorting × Positive Reciprocity		−0.019		−0.054		0.073
		(0.137)		(0.148)		(0.150)
Sorting × Negative Reciprocity		−0.145		−0.063		0.208^{*}
		(0.146)		(0.129)		(0.112)
Observations	283	283	283	283	283	283
R^2	0.04	0.194	0.003	0.055	0.034	0.120

Notes: "Zero-sharers" is a dummy variable for either sorting out or sorting in but sharing nothing. "Low-sharers" is a dummy variable for sharing more than 0 but less than 25 percent of the endowment. "High-sharers" is a dummy variable for sharing at least 25 percent of the endowment. Independent variables are condition group dummies. Robust standard errors are in parentheses (with bias-correction (HC3), see MacKinnon and White (1985).

* - $p < 0.1$; ** - $p < 0.05$; *** - $p < 0.01$

Table C.4: Structural Estimation and Robustness Checks

Specification		(1)	(2)	(3)	(4)	(5)	(6)	(7)
	μ_α	-1.732 (0.486)	-2.214 (0.596)	-2.259 (0.562)	-2.819 (0.681)	-2.214 (0.556)	-3.937 (0.608)	0.003 (0.003)
Dictator	σ_α	3.569 (0.757)	4.171 (0.879)	3.389 (0.593)	3.907 (0.716)	3.147 (0.463)	1.422 (2.165)	0.005 (0.005)
Game	μ_β	2.560 (0.489)	2.888 (0.560)	3.043 (0.573)	3.496 (0.661)	3.037 (0.555)	4.441 (0.603)	0.003 (0.003)
	σ_β	3.159 (0.725)	3.646 (0.827)	3.057 (0.649)	3.589 (0.741)	3.492 (0.687)	4.244 (0.604)	0.006 (0.006)
	μ_α	-5.723 (2.073)	-6.618 (2.499)	-8.866 (4.053)	-9.733 (4.670)	-5.787 (2.021)	-6.916 (2.233)	0.003 (0.003)
Negative	σ_α	5.789 (2.124)	6.149 (2.234)	7.358 (3.352)	7.774 (3.693)	5.693 (1.921)	5.396 (2.705)	0.004 (0.004)
Reciprocity	μ_β	1.010 (0.657)	1.627 (0.846)	2.006 (0.965)	2.603 (1.106)	1.115 (0.747)	2.002 (1.141)	0.000 (0.000)
	σ_β	1.627 (1.071)	2.670 (1.431)	2.507 (1.202)	3.267 (1.389)	1.748 (1.161)	2.944 (1.600)	0.001 (0.001)
	μ_α	0.153 (0.140)	-0.063 (0.248)	0.110 (0.216)	-0.136 (0.358)	-0.773 (0.304)	-2.346 (0.773)	0.001 (0.001)
Positive	σ_α	1.247 (1.005)	2.060 (0.992)	1.199 (1.048)	1.711 (1.120)	2.047 (0.191)	2.039 (2.102)	0.001 (0.001)
Reciprocity	μ_β	2.893 (0.450)	3.136 (0.503)	2.922 (0.404)	3.329 (0.461)	3.196 (0.492)	4.116 (0.997)	0.001 (0.001)
	σ_β	3.485 (0.489)	4.003 (0.664)	2.891 (0.304)	3.469 (0.434)	3.455 (0.569)	3.436 (0.835)	0.001 (0.001)
Weighted SSE		288.671	279.286	284.820	276.454	287.512	296.274	53.142

Notes: GMM estimation results for baseline specification and six robustness checks. α refers to the weight on intrinsic motives, and β the weight on extrinsic motives. Specification details (see section 5.1 in the main text for more details): **(1)** Baseline specification, 33 moments and normally distributed parameters. **(2)** 39 moments, normal distributions. **(3)** 33 moments, uniform distributions. **(4)** 39 moments, uniform distributions. **(5)** Alternative specification with $f = 5$ (see equation C.1). **(6)** Alternative specification with $f = 15$ (see equation C.1). **(7)** Alternative functions G and H, with 27 moments (see equation C.2).

Table C.5: Moments and Model Predictions

		Actual	Baseline Model	Constant Social Pressure	No Social Pressure
Dictator Game	Sort in, give 0	0.1087	0.1725	0.1973	0.0000
	Sort in, give $0.25 - 2.50$	0.2174	0.0015	0.0017	0.0071
	Sort in, give $2.75 - 4.75$	0.0435	0.0911	0.0924	0.2284
	Sort in, give 5	0.1087	0.1769	0.1686	0.2382
	Sort in, give > 5	0.0217	0.0000	0.0000	0.0000
	Sort Out	0.5000	0.5580	0.5399	0.5262
	Give 0	0.3556	0.4395	0.4428	0.5262
	Give $0.25 - 2.50$	0.2667	0.0089	0.0089	0.0071
	Give $2.50 - 4.75$	0.2000	0.2986	0.2971	0.2284
	Give 5	0.1778	0.2529	0.2512	0.2382
	Give > 5	0.0000	0.0000	0.0000	0.0000
Negative Reciprocity	Sort in, give 0	0.2000	0.2347	0.2146	0.0000
	Sort in, give $0.25 - 2.50$	0.1600	0.0010	0.0006	0.0052
	Sort in, give $2.75 - 4.75$	0.0400	0.0716	0.0329	0.1488
	Sort in, give 5	0.0133	0.0724	0.0875	0.1257
	Sort in, give > 5	0.0000	0.0000	0.0000	0.0000
	Sort Out	0.5867	0.6203	0.6645	0.7203
	Give 0	0.6786	0.7883	0.7515	0.7203
	Give $0.25 - 2.50$	0.2500	0.0052	0.0044	0.0052
	Give $2.50 - 4.75$	0.0000	0.1328	0.1262	0.1488
	Give 5	0.0357	0.0737	0.1179	0.1257
	Give > 5	0.0357	0.0000	0.0000	0.0000
Positive Reciprocity	Sort in, give 0	0.0317	0.1689	0.1824	0.0000
	Sort in, give $0.25 - 2.50$	0.3968	0.0034	0.0033	0.0080
	Sort in, give $2.75 - 4.75$	0.1111	0.1723	0.1702	0.2808
	Sort in, give 5	0.1270	0.2463	0.2505	0.3085
	Sort in, give > 5	0.0159	0.0000	0.0000	0.0000
	Sort Out	0.3175	0.4091	0.3936	0.4027
	Give 0	0.1154	0.2133	0.2645	0.4027
	Give $0.25 - 2.50$	0.4615	0.0086	0.0091	0.0080
	Give $2.50 - 4.75$	0.1538	0.3799	0.3705	0.2808
	Give 5	0.2692	0.3983	0.3558	0.3085
	Give > 5	0.0000	0.0000	0.0000	0.0000
Weighted SSE			288.7	291.8	380.2

Notes: Actual and predicted moments from GMM estimations of baseline model, model requiring that social pressure be invariant to reciprocity, and model requiring no social pressure.

Appendix D

Chapter 3 Experiment Instructions

Text in italics is not read aloud. Instructions are read by two research assistants: the narrator and the demonstration player.

NARRATOR: We will start by giving you an overview of this game. Each person in this room has been matched with a partner in the other room. Everyone in this game will earn money which will be divided between you and your partner. Each person will earn money by clicking a device like this *demonstrate* and each of you will decide how you want to divide the money that your partner earns between you and your partner. At the end of the game, one of the two rooms will be randomly selected, and only the decisions made in the room that we select will determine actual payouts.

First, we will explain one part of this game that might be a little confusing. Please listen carefully, and if there is anything you dont understand, you are free to ask questions at any time. In this game, your partner in the next room will earn money and you will decide how you want to divide that money between the two of you. However, you may or may not get to keep this money at the end of the game. While we are playing the game in this room, your partner in the next room will also be making decisions about how she would divide money money that you will earn in the game. After we finish playing the game, we are going to pick one of the two rooms this one or the other one. Only the decisions made in the room that we pick will count. So, if this room is picked, we will pay both you and partner based only on the decisions that you made in the game you will decide how to divide the money that your partner earns. However, if we pick the other room, we will pay both you and your partner based only on the decisions that your partner made your partner will decide how to divide the money that you earn.

So, if we choose this room, the total amount of money you take home is determined by your decisions and how much your partner earns. If we choose the other room, the amount of money you take home is determined by your partners decisions and how much you earn. *Repeat this paragraph once to confirm understanding.*

How will we choose which room and whose decisions determine how much you and your partner take home from the game? After we have finished playing the game, we will place these two round plastic disks into a cup like this *demonstrate using the yellow and red disks.*

Without looking, we will pull one of the two disks out of the cup *demonstrate using the yellow and red disks.* We are in Room **A**. So, if we pull out the **yellow** disk labeled **A** we will divide the money your partner earned and your decisions will determine how much everyone takes home today; if we pull out the **red** disk labeled **B** well divide the money that you earned and the decisions made in the other room will determine how much everyone takes home. So, if we choose **A** well pay you and your partner based on your decisions; if we choose **B** well pay you and your partner based on the decisions that your partner makes in the next room.

So, either your decisions or your partners decisions but not both will determine how much both you and your partner take home at the end of the game. In this game, your partner will earn money and you will decide how to divide it between yourself and your partner. Later on in the game, you will also earn money. Your partner will decide how to divide that money between the two of you. However, youll only get to keep the share of your partners earnings that you allocate yourself and your partner will only get the rest of the money that they earn if we pull the **yellow** disk out of the cup at the end of the game. Otherwise, how much you take home today will be determined by how much you earn and the decisions your partner makes.

Are there any questions so far?

Everything weve said so far doesnt effect you until after the game it doesnt effect the choices you will make, or what you need to understand to play the game. Now, we will explain in more detail how your partner earn money in this game. This will also be the way that you will earn money later in the game.

How much money your partner earns in this game is determined by how much your partner works. In this game, your partner will earn money by clicking this machine. It is called a clicker. Each time you squeeze the clicker, the number on the front increases by one. *Walk around and show players the front face of the clicker as you click.* Every time you squeeze the clicker, the number goes up by one it never decreases. So, the number on the front of the clicker shows the number of times youve clicked. In this game, your partner will be paid money based on the number of times she clicks. Well ask your partner to hold the clicker in one hand, and well give her ten minutes in which she can click as much as she wants. Your partner has to keep the clicker in one hand she cant switch hands, or click with two hands, or click on the table. *Demonstrate each of these.* The more times she clicks the clicker during the ten minutes, the more money she will be paid.

How much your partner is paid depends on how much she clicks. For each number of times she might click, the poster on the wall shows you how much she will be paid. So, shell be paid 30 shillings if she clicks 200 times or less. If she clicks between 201 and 400 times, she will be paid 60 shillings. *Point out these examples on the wall poster.* The more times she click, the more money she earns.

The poster on the wall show you all the possible amounts of money that your partner might earn in the game. The left side indicates all the possible numbers of times your partner might click, and the right side indicates all the possible amounts she might be paid. For example, if she clicks 1,900 times, the number of times that she clicks is between 1,801 and

2,000, so she will be paid 300 shillings. *Point out this row on the poster.* On the other hand, if she clicks 2,100 times, she will be paid 330 shillings because 2,100 is between 2,001 and 2,400. Point out this row on the poster. So, the more times she clicks, the more money shell be paid. You can look at the poster to see how much your partner will be paid for any number of times that she might click.

Here are two examples of earning money in this game. This is *name of demonstration player. Indicate the demonstration player. The demonstration player should click the demonstration clicker very slowly while the next sentence is being read, then stop and hand it to the Narrator as he or she says this. Name of demonstration player* did not click very much only 650 times in ten minutes. How much will she be paid? *Point to the correct row on the poster.* The number of times she clicked is between 601 and 800, so *name of demonstration player* will be paid 90 shillings. *The Narrator hands the clicker back to the demonstration player.*

What if *name of demonstration player* had clicked many times? *Demonstration player clicks very rapidly for ten seconds, stops, and hands the clicker back to the Narrator.* If *name of demonstration player* clicked 3100 times in ten minutes, how much would she be paid? *Point to the correct row on the poster.* The number of times she clicked is between 3001 and 3250, so *name of demonstration player* will be paid 450 shillings.

So, your partner will be paid money in this game for squeezing a tally counter. *Hold up the clicker.* The more times she clicks, the more shell be paid.

Before we explain the rest of the game, we are going to let you try clicking the counters. We are going to give each of you a counter. Dont start clicking yet! Well give you one minute to try clicking. You will not be paid any money for your work during this trial period it is just to give you a sense of what clicking the counters will be like. Your partner will be given ten minutes to click the counter the amount that you would be able to click in that time might be about ten times as much as you can in one minute. Remember, you have to hold the clicker in one hand you cant switch back and forth. Leave your other hand on the desk or at your side.

Research assistants should hand each person a clicker, making sure that they do not start clicking immediately. Once everyone has a clicker in their hand, set a cell phone timer to time one minute. Tell the players when to begin, and then count down the last five seconds before they have to stop. When time is called, make sure that everyone stops. They should not touch their clickers after time is called. As the Narrator continues with the instructions, the other research assistant should collect all the clickers, remove the tape, reset them to zero, and re-cover them with tape so that they cannot be manipulated.

Part III

Weve explained how your partner will earn money in this game by clicking a tally counter. Now we will explain about your partners. This is the last thing you need to understand about this game.

Weve divided you into two groups. We randomly assigned half of the seat numbers to each room. Everyone in this room is in Group **A**, and everyone in the other room is in Group **B**. Every player in Group **A** has been matched with a player in Group **B**. That person is your partner. You will never learn your partners identity, nor will your partner ever learn who you are.

In this game, your partner will earn money by clicking a tally counter and you will decide how you want to divide the money that you earn between yourself and your partner in the next room. You can divide it any way you like the choice is yours. You can give yourself all of your partners earnings, give your partner all of her earnings, or do anything in between. The only way your partner can receive any money in this game is for you to give it to her. The money that you give to your partner will be the only money that your partner receives if we draw the **yellow disk** out of the bowl at the end of the game. *Hold up the yellow disk to remind the audience about the disks.*

Before you earn money by clicking, well ask you to indicate how much of your partners earnings you would like to give yourself and how much you want to give to your partner. You are allowed to give as much or as little to your partner as you want it is your decision. How much you want to give to your partner might depend on how much you earn. You have tried clicking the counters you know what it is like. For each possible amount that your partner might end up earning, well ask you to indicate how you would divide that money between yourself and your partner. Later, when we find out how much your partner has actually earned, well divide the money that shes earned between you and your partner however you told us to.

Before we give your partner the opportunity click the counters, we will give each of you a set of forms like this. *Hold up a decision packet.* The packet lists all the possible amounts of money that your partner might earn: on each page of the form, there is one possible amount of money that your partner might earn by clicking. *Indicate sample decision sheet on the wall of the room.* So, the first page is 30 shillings, the second page is 60 shillings, all the way up to 600 shillings. You can see that one each page, you can actually see the coins that your partner will earn.

On each page, youll indicate how you want to divide the money that your partner earns between you and your partner by circling the coins that you want to give to yourself, and not circling the bills and coins that you want to give to your partner.

Lets look at a couple of examples. *Name of demonstration player* is about to make her decisions about how shed like to divide the money that her partner earns between herself and her partner. *The demonstration player stands in front of the first sample decision sheet.* Remember, shell circle any coins she wants to give to herself, and shell leave unmarked any coins that she wants to give to her partner. *Name of demonstration player*, are you ready to start?

DEMONSTRATION PLAYER: Yes, but I have a question. After the game, can I find out who my partner is, so that I know whether I gave her enough money?

NARRATOR: No. Youll never know who your partner is.

DEMONSTRATION PLAYER: OK. *Demonstration player turns to face the first decision sheet. She mimes thinking for a few seconds before speaking.* I will give myself twenty shillings and give my partner ten shillings. *Demonstration player circles twenty shillings on the sample decision sheet.*

NARRATOR: Well go through the sheets in your packet one at a time. Next, *name of demonstration player* will decide how she wants to divide the money if her partner earns shillings.

DEMONSTRATION PLAYER: *Demonstration player again thinks for a few seconds.* I will give myself all of the money. *Demonstration player circles each of the rows of coins.*

NARRATOR: Remember, you can divide the money that your partner earned any way you want the decision is yours. You can give your partner all of her earnings, or none of it. You can do whatever you want to do. Each possible amount that your partner might earn is represented on one of the pages of your packet. As the amounts get higher, you will have to divide more coins between yourself and your partner.

DEMONSTRATION PLAYER: *Demonstration player moves over to the sample page with three hundred shillings on it.* I will give myself 150 shillings and give my partner 150 shillings. *She circles 150 shillings worth of coins on the sheet.*

NARRATOR: We will ask you how you want to divide all of the possible amounts your partner might earn: 30 shillings, 60 shillings, all the way up to 600 shillings. Then, if the **yellow** disk is picked at the end of the game, well find out how much your partner actually earned and divide your partners earnings exactly how you told us to.

Are there any questions? Answer any questions. OK, we are finished explaining the game. Now, lets review.

First, we are going to ask you to indicate in a decision packet how you want to divide your partners earnings between yourself and your partner. Youll tell us how you would divide all of the possible amounts that your partner might end up earning by clicking.

After all of you have made your decisions about how you want to divide your partners earnings, each person will be given ten minutes to click the tally counter. The more times you click, the more money you earn; the poster on the wall tells you how much you will earn for each possible number of times you might click. After we determine how much your partner has earned, we will divide that money between you and your partner however you told us to.

At the end of the game, well place these two plastic disks into a cup. Well pull one of the disks out without looking. If we pull out the **yellow disk** with a letter **A** on it, youll take home the money your partner earned minus whatever amount you allocate to your partner. Your partner will take home only the money that you allocate them.

Are we ready to begin?

*Answer any questions, and then pass out the decision packets and pens. It is **very important** that there is **no talking at all** while the subjects make their decisions.*

After all the players have made their decisions, collect the packets and all of the pens. Then read the following. Youve all decided how you want to divide the money that you earn. Now we are going to let you click the counters to determine how much you are paid. Well place a counter on the desk in front of you. Please do not pick it up until we tell you to. *Pass out the re-taped clickers, making sure that each is set at zero. Then set a mobile phone timer for ten minutes and tell the players to start. Inform them when they have one minute left, and then count down the last five seconds.*

Bibliography

[1] Alan I. Abramowitz and Kyle L. Saunders. “Is Polarization a Myth?” In: *Journal of Politics* 70.02 (Mar. 2008), pp. 542–555.

[2] George A. Akerlof. “A theory of social custom, of which unemployment may be one consequence”. In: *Quarterly Journal of Economics* 94.4 (1980), pp. 749–775.

[3] George A. Akerlof. “Labor contracts as partial gift exchange”. In: *Quarterly Journal of Economics* 97.4 (1982), pp. 543–569.

[4] George A. Akerlof and Rachel E. Kranton. “Economics and Identity”. In: *Quarterly Journal of Economics* CXV.3 (2000), pp. 715–753.

[5] George A. Akerlof and Janet L. Yellen. “Fairness and unemployment”. In: *American Economic Review: Papers and Proceedings* 78.2 (Feb. 1988), pp. 44–49.

[6] George A. Akerlof and Janet L. Yellen. “The fair wage-effort hypothesis and unemployment”. In: *Quarterly Journal of Economics* 105.2 (1990), pp. 255–283.

[7] Gani Aldashev et al. “Using the law to change the custom”. In: *Journal of Development Economics* 97.2 (Mar. 2011), pp. 182–200.

[8] Alberto Alesina and Edward L. Glaeser. *Fighting Poverty in the US and Europe: A World of Difference.* New York: Oxford University Press, 2004.

[9] Ingvild Almå s et al. “Fairness and the development of inequality acceptance.” In: *Science* 328.5982 (May 2010), pp. 1176–8.

[10] Francisco Alpizar, Fredrik Carlsson, and Olof Johansson-Stenman. “Does context matter more for hypothetical thanforactual contributions? Evidence from a natural field experiment”. In: *Experimental Economics* 11.3 (Feb. 2008), pp. 299–314.

[11] T.W. Anderson and Herman Rubin. “Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations”. In: *Annals of Mathematical Statistics* 20.1 (1949), pp. 46–63.

[12] James Andreoni. “Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence”. In: *Journal of Political Economy* 97.6 (Jan. 1989), pp. 1447–58.

[13] James Andreoni. “Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving”. In: *The Economic Journal* 100.401 (June 1990), pp. 464–477.

[14] James Andreoni and B. Douglas Bernheim. “Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects”. In: *Econometrica* 77.5 (2009), pp. 1607–1636.

[15] James Andreoni and John H. Miller. “Giving According to GARP : An Experimental Test of the Consistency of Preferences for Altruism”. In: *Econometrica* 70.2 (2002), pp. 737–753.

[16] James Andreoni and Ragan Petrie. “Public goods experiments without confidentiality: a glimpse into fund-raising”. In: *Journal of Public Economics* 88.7-8 (July 2004), pp. 1605–1623.

[17] James Andreoni and John Karl Scholz. “An Econometric Analysis of Charitable Giving With Interdependent Preferences”. In: *Economic Inquiry* 36.3 (July 1998), pp. 410–428.

[18] James Andreoni and Lise Vesterlund. “Which is the fair sex? Gender differences in altruism”. In: *Quarterly Journal of Economics* 116.1 (2001), pp. 293–312.

[19] Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics*. Princeton: Princeton University Press, 2009.

[20] Dan Ariely, Anat Bracha, and Stephan Meier. “Doing good or doing well? Image motivation and monetary incentives in behaving prosocially”. In: *American Economic Review* 99.1 (2009), pp. 544–555.

[21] Michael Armer and Armer Youtz. “Formal Education and Individual Modernity in an African Society”. In: *American Journal of Sociology* 76.4 (1976), pp. 604–626.

[22] Solomon E. Asch. “Opinions and Social Pressure”. In: *Scientific American* 193.5 (1955), pp. 31–35.

[23] Nava Ashraf, Oriana Bandiera, and Kelsey Jack. “No margin, no mission? A Field Experiment on Incentives for Pro-Social Tasks”. Working Paper. 2012.

[24] Nava Ashraf, Oriana Bandiera, and Scott Lee. “Awards Unbundled: Evidence from a Natural Field Experiment”. Working Paper. 2013.

[25] International Bible Association. *The Holy Bible: New International Version*. 1985.

[26] Philip Babcock et al. “Letting Down the Team? Evidence of Social Effects of Team Incentives”. NBER Working Paper Series No. 16687. 2010.

[27] Oriana Bandiera, Iwan Barankay, and Imran Rasul. “Social Incentives in the Workplace”. In: *Review of Economic Studies* 77.2 (Apr. 2010), pp. 417–458.

[28] Oriana Bandiera, Iwan Barankay, and Imran Rasul. “Social Preferences and the Response to Incentives: Evidence from Personnel Data”. In: *Quarterly Journal of Economics* 120.3 (Aug. 2005), pp. 917–962.

[29] Nicholas Bardsley. “Dictator game giving: altruism or artefact?” In: *Experimental Economics* 11.2 (Sept. 2007), pp. 122–133.

[30] Abigail Barr and Mattea Stein. "Status and egalitarianism in traditional communities: An analysis of funeral attendance in six Zimbabwean villages". Working Paper. 2008.

[31] Abigail Barr et al. "Homo-æqualis: A cross-society experimental analysis of three bargaining games". University of Oxford Department of Economics Discussion Paper Series No. 422. 2009.

[32] Robert J. Barro. "Democracy and growth". In: *Journal of Economic Growth* 1.1 (Mar. 1996), pp. 1–27.

[33] Robert J. Barro. "Education and Economic Growth". In: *The Contribution of Human and Social Capital to Sustained Economic Growth and Well-Being*. Ed. by J.F. Helliwell. OECD, 2001.

[34] Pierpaolo Battigalli and Martin Dufwenberg. "Dynamic psychological games". In: *Journal of Economic Theory* 144.1 (Jan. 2009), pp. 1–35.

[35] Pierpaolo Battigalli and Martin Dufwenberg. "Guilt in Games". In: *American Economic Review* 97.2 (May 2007), pp. 170–176.

[36] Roy F. Baumeister et al. "Bad is stronger than good." In: *Review of General Psychology* 5.4 (2001), pp. 323–370.

[37] Gary S. Becker. "A theory of social interactions". In: *Journal of Political Economy* 82.6 (1974), pp. 1063–1093.

[38] Gary S. Becker. "Notes on an Economic Analysis of Philanthropy". NBER Working Paper Series. 1961.

[39] René H.F.P. Bekkers. "Measuring altruistic behavior in surveys: The all-or-nothing dictator game". In: *Survey Research Methods* 1.3 (2007), pp. 139–144.

[40] Roland Bénabou and Jean Tirole. "Identity, morals, and taboos: Beliefs as assets". In: *Quarterly Journal of Economics* 126.2 (2011), pp. 805–855.

[41] Roland Bénabou and Jean Tirole. "Incentives and prosocial behavior". In: *American Economic Review* 96.5 (2006), pp. 1652–1678.

[42] Joyce Berg, John Dickhaut, and Kevin A. McCabe. "Trust, Reciprocity, and Social History". In: *Games and Economic Behavior* 10.1 (1995), pp. 122–142.

[43] Theodore Bergstrom, Lawrence Blume, and Hal Varian. "On the private provision of public goods". In: *Journal of Public Economics* 29.1 (1986), pp. 25–49.

[44] Tanguy Bernard, Alain De Janvry, and Elisabeth Sadoulet. "When Does Community Conservatism Constrain Village Organizations?" In: *Economic Development and Cultural Change* January (2010), pp. 1–36.

[45] B. Douglas Bernheim. "A Theory of Conformity". In: *Journal of Political Economy* 102.5 (Jan. 1994), pp. 841–877.

[46] Truman F. Bewley. *Why Wages Dont Fall During a Recession*. Harvard University Press, 2009.

[47] Christina Bicchieri. *The Grammar of Society: The Nature and Dynamics of Social Norms*. New York: Cambridge University Press, 2006.

[48] Sally Blount. "When Social Outcomes Arent Fair: The Effect of Causal Attributions on Preferences". In: *Organizational Behavior and Human Decision Processes* 63.2 (Aug. 1995), pp. 131–144.

[49] Iris Bohnet and Bruno S. Frey. "Social distance and other-regarding behavior in dictator games: Comment". In: *American Economic Review* 89.1 (1999), pp. 335–339.

[50] Iris Bohnet and Bruno S. Frey. "The sound of silence in prisoner's dilemma and dictator games". In: *Journal of Economic Behavior & Organization* 38.1 (1999), pp. 43–57.

[51] Gary E. Bolton and Axel Ockenfels. "ERC: A theory of equity, reciprocity, and competition". In: *American Economic Review* 90.1 (2000), pp. 166–193.

[52] Samuel Bowles and Herbert Gintis. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton: Princeton University Press, 2011.

[53] Jordi Brandts and Carles Solà. "Reference points and negative reciprocity in simple sequential games". In: *Games and Economic Behavior* 36.2 (2001), pp. 138–157.

[54] Troyen A. Brennan et al. "Health industry practices that create conflicts of interest". In: *Journal of the American Medical Association* 295.4 (2006), pp. 429–433.

[55] Tomas Broberg, Tore Ellingsen, and Magnus Johannesson. "Is generosity involuntary?" In: *Economics Letters* 94.1 (Jan. 2007), pp. 32–37.

[56] J. Michelle Brock, Andreas Lange, and Erkut Y. Ozbay. "Dictating the Risk: Experimental Evidence on Giving in Risky Environments". In: *American Economic Review* 103.1 (Feb. 2013), pp. 415–437.

[57] Jeffrey V. Butler, Paola Giuliano, and Luigi Guiso. "Trust and Cheating". NBER Working Paper Series No. 18509. 2012.

[58] John C. Caldwell and Pat Caldwell. "The Cultural Context of High Fertility in Africa sub-Saharan". In: *Population and Development Review* 13.3 (1987), pp. 409–437.

[59] Colin F. Camerer. *Behavioral Game Theory*. Princeton University Press Princeton, NJ, 2003.

[60] Colin F. Camerer and Richard H. Thaler. "Anomalies: Ultimatums, dictators and manners". In: *Journal of Economic Perspectives* 9.2 (1995), pp. 209–219.

[61] Lisa Cameron, Paul Gertler, and Manisha Shah. "The dirty business of open defecation : Lessons from a sanitation intervention". In: *Pacific Conference for Development Economics*. 2013.

[62] Alexander W. Cappelen et al. "The Pluralism of Fairness Ideals: An Experimental Approach". In: *American Economic Review* 97.3 (June 2007), pp. 818–827.

[63] Katherine Grace Carman. “Social influences and the private provision of public goods: Evidence from charitable contributions in the workplace”. Harvard University Job Market Paper. 2004.

[64] Jeffrey Paul Carpenter. “Endogenous Social Preferences”. In: *Review of Radical Political Economics* 37.1 (Mar. 2005), pp. 63–84.

[65] Jeffrey Paul Carpenter and Peter Hans Matthews. “Social reciprocity”. IZA Discussion Paper Series No. 1347. 2004.

[66] Jeffrey Paul Carpenter, Peter Hans Matthews, and Okomboli Ongonga. “Why Punish? Social reciprocity and the enforcement of prosocial norms”. In: *Journal of Evolutionary Economics* 14.4 (Oct. 2004), pp. 407–429.

[67] Jeffrey Paul Carpenter and Caitlin Knowles Myers. “Why volunteer? Evidence on the role of altruism, image, and incentives”. In: *Journal of Public Economics* 94.11-12 (Dec. 2010), pp. 911–920.

[68] Anne Case et al. “Paying the piper: the high cost of funerals in South Africa”. NBER Working Paper Series No. 14456. 2008.

[69] Timothy N. Cason and Feisal U. Khan. “A laboratory study of voluntary public goods provision with imperfect monitoring and communication”. In: *Journal of Development Economics* 58.2 (Apr. 1999), pp. 533–552.

[70] Gary Charness and Matthew Rabin. “Understanding Social Preferences with Simple Tests”. In: *Quarterly Journal of Economics* 117.3 (2002), pp. 817–869.

[71] Todd L. Cherry. “Mental accounting and other-regarding behavior: Evidence from the lab”. In: *Journal of Economic Psychology* 22.5 (Oct. 2001), pp. 605–615.

[72] Todd L. Cherry, Peter Frykblom, and Jason F. Shogren. “Hardnose the Dictator”. In: *American Economic Review* 92.4 (Sept. 2002), pp. 1218–1221.

[73] In-Koo Cho and David M. Kreps. “Signaling Games and Stable Equilibria”. In: *Quarterly Journal of Economics* 102.2 (1987), pp. 179–221.

[74] Robert B. Cialdini. “Crafting Normative Messages to Protect the Environment”. In: *Current Directions in Psychological Science* 12.4 (2003), pp. 105–109.

[75] Robert B. Cialdini. *Influence: The Psychology of Persuasion.* HarperCollins, 1993.

[76] Robert B. Cialdini, Carl A. Kallgren, and Raymond R. Reno. “A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior”. In: *Advances in Experimental Social Psychology* 24 (1991), pp. 201–234.

[77] David J. Cooper and John H. Kagel. “Other-Regarding Preferences: A Selective Survey of Experimental Results”. In: *Handbook of Experimental Economics, Vol. 2.* Ed. by John H. Kagel and Alvin E. Roth. Vol. 2. Princeton, NJ: Princeton University Press.

[78] James C. Cox. “How to identify trust and reciprocity”. In: *Games and Economic Behavior* 46 (2004), pp. 260–281.

[79] James C. Cox, Daniel Friedman, and Steven Gjerstad. “A tractable model of reciprocity and fairness”. In: *Games and Economic Behavior* 59.1 (Apr. 2007), pp. 17–45.

[80] Rachel T.A. Croson. “Theories of Commitment, Altruism and Reciprocity: Evidence From Linear Public Goods Games”. In: *Economic Inquiry* 45.2 (Apr. 2007), pp. 199–216.

[81] Rachel T.A. Croson, Enrique Fatas, and Tibor Neugebauer. “Reciprocity, matching and conditional cooperation in two public goods games”. In: *Economics Letters* 87.1 (Apr. 2005), pp. 95–101.

[82] Jason Dana, Daylian M. Cain, and Robyn M. Dawes. “What you dont know wont hurt me: Costly (but quiet) exit in dictator games”. In: *Organizational Behavior and Human Decision Processes* 100 (July 2006), pp. 193–201.

[83] Jason Dana, Roberto A. Weber, and Jason Xi Kuang. “Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness”. In: *Economic Theory* 33 (Sept. 2007), pp. 67–80.

[84] Angus Deaton. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Baltimore, MD: Johns Hopkins University Press, 1997.

[85] Stefano DellaVigna, John A. List, and Ulrike Malmendier. “Testing for altruism and social pressure in charitable giving”. In: *Quarterly Journal of Economics* 127.1 (2012), pp. 1–56.

[86] Stefano DellaVigna et al. “The Importance of Being Marginal: Gender Differences in Generosity”. In: *American Economic Review: Papers and Proceedings* 103.3 (May 2013), pp. 586–590.

[87] Stefano DellaVigna et al. “Voting to Tell Others”. Working Paper. 2013.

[88] Geert Dhaene and Jan Bouckaert. “Sequential reciprocity in two-player, two-stage games: An experimental analysis”. In: *Games and Economic Behavior* 70.2 (Nov. 2010), pp. 289–303.

[89] Rafael Di Tella, Sebastian Galiani, and Ernesto Schargrodsky. “The formation of beliefs: evidence from the allocation of land titles to squatters”. In: *Quarterly Journal of Economics* 122.1 (2007), pp. 209–241.

[90] David L. Dickinson and Jill Tiefenthaler. “What Is Fair? Experimental Evidence”. In: *Southern Economic Journal* 69.2 (2002), pp. 414–428.

[91] Esther Duflo. “Schooling and labor market consequences of school construction in Indonesia: Evidence from an unusual policy experiment”. In: *American Economic Review* 19.4 (2001), pp. 795–813.

[92] Jean-Marie Dufour. “Some Impossibility Theorems in Econometrics With Applications to Structural and Dynamic Models”. In: *Econometrica* 65.6 (1997), pp. 1365–1387.

[93] Martin Dufwenberg and Georg Kirchsteiger. “A theory of sequential reciprocity”. In: *Games and Economic Behavior* 47.2 (May 2004), pp. 268–298.

[94] Richard A. Easterlin. “Why isn’t the whole world developed?” In: *Journal of Economic History* 41.1 (Mar. 1981), pp. 1–19.

[95] René Fahr and Bernd Irlenbusch. “Fairness as a constraint on trust in reciprocity: earned property rights in a reciprocal exchange experiment”. In: *Economics Letters* 66.3 (2000), pp. 275–282.

[96] Armin Falk. “Gift Exchange in the Field”. In: *Econometrica* 75.5 (Sept. 2007), pp. 1501–1511.

[97] Armin Falk, Ernst Fehr, and Urs Fischbacher. “Testing theories of fairness - Intentions matter”. In: *Games and Economic Behavior* 62.1 (Jan. 2008), pp. 287–303.

[98] Armin Falk and Urs Fischbacher. “A theory of reciprocity”. In: *Games and Economic Behavior* 54.2 (Feb. 2006), pp. 293–315.

[99] Armin Falk and Andrea Ichino. “Clean evidence on peer effects”. In: *Journal of Labor Economics* 24.1 (2006), pp. 39–57.

[100] James D. Fearon, Macartan Humphreys, and Jeremy M. Weinstein. “Can Development Aid Contribute to Social Cohesion after Civil War? Evidence from a Field Experiment in Post-Conflict Liberia”. In: *American Economic Review* 99.2 (Apr. 2009), pp. 287–291.

[101] Ernst Fehr, Helen Bernhard, and Bettina Rockenbach. “Egalitarianism in young children.” In: *Nature* 454.7208 (Aug. 2008), pp. 1079–83.

[102] Ernst Fehr and Urs Fischbacher. “Third-party punishment and social norms”. In: *Evolution and Human Behavior* 25.2 (Mar. 2004), pp. 63–87.

[103] Ernst Fehr and Simon Gächter. “Cooperation and punishment in public goods experiments”. In: *American Economic Review* 151.3712 (Feb. 2000), pp. 867–8.

[104] Ernst Fehr and Simon Gächter. “Fairness and Retaliation: The Economics of Reciprocity”. In: *Journal of Economic Perspectives* 14.3 (Aug. 2000), pp. 159–182.

[105] Ernst Fehr, Simon Gächter, and Georg Kirchsteiger. “Reciprocity as a contract enforcement device: Experimental evidence”. In: *Econometrica* (1997).

[106] Ernst Fehr, Lorenz Goette, and Christian Zehnder. “A Behavioral Account of the Labor Market: The Role of Fairness Concerns”. In: *Annual Review of Economics* 1.1 (Sept. 2009), pp. 355–384.

[107] Ernst Fehr, Georg Kirchsteiger, and Arno Riedl. “Does Fairness Prevent Market Clearing? An Experimental Investigation”. In: *Quarterly Journal of Economics* 108.2 (1993), pp. 437–459.

[108] Ernst Fehr and Klaus M. Schmidt. “A theory of fairness, competition, and cooperation”. In: *Quarterly Journal of Economics* 114.3 (1999), pp. 817–868.

[109] Lars P. Feld and Bruno S. Frey. “Tax compliance as the result of a psychological tax contract: The role of incentives and responsive regulation”. In: *Law & Policy* 29.1 (2007), pp. 102–120.

[110] Frederico Finan and Laura Schechter. “VoteBuying and Reciprocity”. In: *Econometrica* 80.2 (2012), pp. 863–881.

[111] Morris P. Fiorina and Samuel J. Abrams. “Political Polarization in the American Public”. In: *Annual Review of Political Science* 11.1 (June 2008), pp. 563–588.

[112] Urs Fischbacher, Simon Gächter, and Ernst Fehr. “Are people conditionally cooperative? Evidence from a public goods experiment”. In: *Economics Letters* 71.3 (June 2001), pp. 397–404.

[113] Raymond Fisman, Pamela Jakiela, and Shachar Kariv. “How Did Distributional Preferences Change During the Great Recession?” Working Paper. 2013.

[114] Raymond Fisman, Shachar Kariv, and Daniel Markovits. “Exposure to Ideology and Distributional Preferences”. Working Paper. 2009.

[115] Raymond Fisman, Shachar Kariv, and Daniel Markovits. “Individual Preferences for Giving”. In: *American Economic Review* 97.5 (2007), pp. 1858–1876.

[116] Christina M. Fong and Erzo F.P. Luttmer. “What determines giving to Hurricane Katrina victims? Experimental evidence on racial group loyalty”. In: *American Economic Journal: Applied Economics* 1.2 (2009), pp. 64–87.

[117] Robert Forsythe et al. “Fairness in Simple Bargaining Experiments”. In: *Games and Economic Behavior* 6 (1994), pp. 347–369.

[118] James H. Fowler. “Altruism and turnout”. In: *Journal of Politics* 68.3 (2006), pp. 674–683.

[119] Axel Franzen and Sonja Pointner. “Anonymity in the dictator game revisited”. In: *Journal of Economic Behavior & Organization* 81.1 (Jan. 2012), pp. 74–81.

[120] Bruno S. Frey and Stephan Meier. “Social Comparisons and Pro-social Behavior: Testing ”Conditional Cooperation” in a Field Experiment”. In: *American Economic Review* 94.5 (2004), pp. 1717–1722.

[121] Willa Friedman et al. “Economics as Liberation?” NBER Working Paper Series No. 16939. 2011.

[122] Bruce Fuller, Judith D. Singer, and Margaret Keiley. “Why do Daughters Leave School in Southern Africa? Family Economy and Mothers’ Commitments”. In: *Social Forces* 74.2 (Dec. 1995), p. 657.

[123] John Gale, Kenneth G. Binmore, and Larry Samuelson. “Learning to be imperfect: the ultimatum game”. In: *Games and Economic Behavior* 8 (1995), pp. 56–90.

[124] Luis Garicano, Ignacio Palacios-Huerta, and Canice Prendergast. "Favoritism Under Social Pressure". In: *Review of Economics and Statistics* 87.2 (May 2005), pp. 208–216.

[125] John Geanakoplos, David Pearce, and Ennio Stacchetti. "Psychological games and sequential rationality". In: *Games and Economic Behavior* 1.1 (Mar. 1989), pp. 60–79.

[126] Alan S. Gerber, Donald P. Green, and Christopher W. Larimer. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment". In: *American Political Science Review* 102.01 (Feb. 2008), pp. 33–48.

[127] Francesca Gino, Shahar Ayal, and Dan Ariely. "Contagion and differentiation in unethical behavior: the effect of one bad apple on the barrel." In: *Psychological science* 20.3 (Mar. 2009), pp. 393–8.

[128] Herbert Gintis. "Strong Reciprocity and Human Sociality". In: *Journal of Theoretical Biology* 206.2 (2000), pp. 169–179.

[129] Herbert Gintis. "The evolution of private property". In: *Journal of Economic Behavior & Organization* 64.1 (2007), pp. 1–16.

[130] Edward L. Glaeser et al. "Do Institutions Cause Growth?" In: *Journal of Economic Growth* 9.3 (Sept. 2004), pp. 271–303.

[131] Amihai Glazer and Kai A. Konrad. "A signaling explanation for charity". In: *American Economic Review* 86.4 (1996), pp. 1019–1028.

[132] Uri Gneezy. "Deception: The role of consequences". In: *American Economic Review* 95.1 (2005), pp. 384–394.

[133] Uri Gneezy and John A. List. "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments". In: *Econometrica* 74.5 (Sept. 2006), pp. 1365–1384.

[134] Jacob K. Goeree and Leeat Yariv. "Conformity in the Lab". California Institute of Technology Working Paper. 2007.

[135] Claudia Goldin and Lawrence F. Katz. *The Race between Education and Technology.* Cambridge, MA: Harvard University Press, 2008.

[136] Jerald Greenberg. "Employee theft as a reaction to underpayment inequity: The hidden cost of pay cuts." In: *Journal of Applied Psychology* 75.5 (1990), pp. 561–568.

[137] Fiona Greig. "Gender and the social costs of claiming value: An experimental approach". In: *Journal of Economic Behavior & Organization* 76.3 (Dec. 2010), pp. 549–562.

[138] Sanford J. Grossman and Motty Perry. "Perfect sequential equilibrium". In: *Journal of economic theory* 39.1 (1986), pp. 97–119.

[139] Zachary Grossman. "Self-signaling and social-signaling in giving". UC Santa Barbara Department of Economics Departmental Working Papers. 2012.

[140] Zachary Grossman. "Social-Signaling with Anonymity: Rule-Rationality or Beliefs-Based Altruism?" UC Santa Barbara Department of Economics Working Paper. 2013.

[141] Faruk Gul and Wolfgang Pesendorfer. "The canonical type space for interdependent preferences". Princeton University Working Paper. 2006.

[142] Eric A. Hanushek and Dennis D. Kimko. "Schooling, labor-force quality, and the growth of nations". In: *American Economic Review* 90.5 (2000), pp. 1184–1208.

[143] William T. Harbaugh. "The Prestige Motive for Making Charitable Transfers". In: *American Economic Review* 88.2 (Feb. 1998), pp. 277–282.

[144] William T. Harbaugh. "What do donations buy? A model of philanthropy based on prestige and warm glow". In: *Journal of Public Economics* 67.2 (Feb. 1998), pp. 269–284.

[145] Joseph Patrick Henrich, Steven J. Heine, and Ara Norenzayan. "The weirdest people in the world?" In: *Behavioral and Brain Sciences* 33.2-3 (2010), pp. 61–83.

[146] Joseph Patrick Henrich et al. *Foundations in Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-scale Societies*. New York, NY: Oxford University Press, 2004.

[147] Joseph Patrick Henrich et al. "In search of homo economicus: behavioral experiments in 15 small-scale societies". In: *American Economic Review* 91.2 (2001), pp. 73–78.

[148] Joseph Patrick Henrich et al. "Markets, religion, community size, and the evolution of fairness and punishment". In: *Science* 327.5972 (Mar. 2010), pp. 1480–4.

[149] Elizabeth Hoffman, Kevin A. McCabe, and Vernon L. Smith. "Behavioral Foundations of Reciprocity: Experimental Economics and Evolutionary Psychology". In: *Economic Inquiry* 36.3 (July 1998), pp. 335–352.

[150] Elizabeth Hoffman, Kevin A. McCabe, and Vernon L. Smith. "Social distance and other-regarding behavior in dictator games". In: *American Economic Review* 86.3 (1996), pp. 653–660.

[151] Elizabeth Hoffman et al. "Preferences, property rights, and anonymity in bargaining games". In: *Games and Economic Behavior* 7.3 (1994), pp. 346–380.

[152] Ronald Inglehart and Wayne E. Baker. "Modernization, cultural change, and the persistence of traditional values". In: *American Sociological Review* 65.1 (Feb. 2000), pp. 19–51.

[153] Alex Inkeles. "Making Men Modern: On the Causes and Consequences of Individual Change in Six Developing Countries". In: *American Journal of Sociology* 75.2 (1969), pp. 208–225.

[154] Alex Inkeles and David H. Smith. *Becoming modern: Individual change in six developing countries*. Cambridge, MA: Harvard University Press, 1974.

[155] Sasha Issenberg. "Gay-Marriage Strategists Plot PsyOps: The Inevitability Campaign". In: *New York Magazine* (2013).

[156] Pamela Jakiela. "How Fair Shares Compare: Experimental Evidence from Two Cultures". U.C. Berkeley Job Market Paper. 2009.

[157] Magnus Johannesson. "Non-reciprocal altruism in dictator games". In: *Economics Letters* 69 (2000), pp. 137–142.

[158] Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler. "Fairness as a constraint on profit seeking: Entitlements in the market". In: *American Economic Review* 76.4 (1986), pp. 728–741.

[159] Dean Karlan. "Hey look at me: The effect of giving circles on giving". NBER Working Paper Series No. 17737. 2012.

[160] Claudia Keser and Frans van Winden. "Conditional Cooperation and Voluntary Contributions to Public Goods". In: *Scandinavian Journal of Economics* 102.1 (Mar. 2000), pp. 23–39.

[161] Alexander K. Koch and Hans Theo Normann. "Giving in dictator games: Regard for others or regard by others?" In: *Southern Economic Journal* 75.1 (2008), pp. 223–231.

[162] James Konow. "Fair shares: Accountability and cognitive dissonance in allocation decisions". In: *American Economic Review* 90.4 (2000), pp. 1072–1091.

[163] Michael Kremer, Edward Miguel, and Rebecca Thornton. "Incentives to Learn". In: *Review of Economics and Statistics* 91.3 (Aug. 2009), pp. 437–456.

[164] Anirudh Krishna et al. "Escaping Poverty and Becoming Poor in 20 Kenyan Villages". In: *Journal of Human Development* 5.2 (July 2004), pp. 211–226.

[165] Alan B. Krueger and Alexandre Mas. "Strikes, Scabs, and Tread Separations: Labor Strife and the Production of Defective Bridgestone/Firestone Tires". In: *Journal of Political Economy* 112.2 (Apr. 2004), pp. 253–289.

[166] Erin L. Krupka, Stephen Leider, and Ming Jiang. "A Meeting of the Minds: Contracts and Social Norms". Working Paper. 2012.

[167] Erin L. Krupka and Roberto A. Weber. "Identifying social norms using coordination games: Why does dictator game sharing vary?" In: *Journal of the European Economic Association* 11.3 (June 2013), pp. 495–524.

[168] Sebastian Kube, Michel André Maréchal, and Clemens Puppe. "Do wage cuts damage work morale? Evidence from a natural field experiment". In: *Journal of the European Economic Association* 11.4 (2013). Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 471, pp. 853–870.

[169] Sebastian Kube, Michel André Maréchal, and Clemens Puppe. “The Currency of Reciprocity: Gift Exchange in the Workplace”. In: *American Economic Review* 102.4 (June 2012), pp. 1644–1662.

[170] Geoffrey C. Layman, Thomas M. Carsey, and Juliana Menasce Horowitz. “Party Polarization in American Politics: Characteristics, Causes, and Consequences”. In: *Annual Review of Political Science* 9.1 (June 2006), pp. 83–110.

[171] Edward P. Lazear, Ulrike Malmendier, and Roberto A. Weber. “Sorting in Experiments with Application to Social Preferences”. In: *American Economic Journal: Applied Economics* 4.1 (2012), pp. 136–164.

[172] John O. Ledyard. “Public goods: A survey of experimental research”. In: *The Handbook of Experimental Economics*. Ed. by John H. Kagel and Alvin E. Roth. Princeton University Press, 1997.

[173] David K. Levine. “Modeling Altruism and Spitefulness in Experiments”. In: *Review of Economic Dynamics* 1.3 (July 1998), pp. 593–622.

[174] John A. List. “On the Interpretation of Giving in Dictator Games”. In: *Journal of Political Economy* 115.3 (2007), pp. 482–493.

[175] John A. List. “The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions”. In: *Journal of Political Economy* 114.1 (Feb. 2006), pp. 1–37.

[176] John A. List and Todd L. Cherry. “Examining the role of fairness in high stakes allocation decisions”. In: *Journal of Economic Behavior & Organization* 65.1 (Jan. 2008), pp. 1–8.

[177] John Locke. *Second treatise of government*. 1690.

[178] Adrienne M. Lucas and Isaac M. Mbiti. “Access, Sorting, and Achievement: The Short-Run Effects of Free Primary Education in Kenya”. In: *American Economic Journal: Applied Economics2* 4.4 (2012), pp. 226–253.

[179] Kate Macintyre, Lisanne Brown, and Stephen Sosler. “It’s not what you know, but who you knew: examining the relationship between behavior change and AIDS mortality in Africa”. In: *AIDS Education and Prevention* 13.2 (Apr. 2001), pp. 160–74.

[180] James G. MacKinnon and Halbert White. “Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties”. In: *Journal of Econometrics* 29.3 (Sept. 1985), pp. 305–325.

[181] Ulrike Malmendier and Stefan Nagel. “Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?” In: *Quarterly Journal of Economics* 126.1 (2011), pp. 373–416.

[182] Ulrike Malmendier and Klaus M. Schmidt. “You Owe Me”. NBER Working Paper Series No. 18543. 2012.

[183] Ulrike Malmendier and Vera L. te Velde. “Social image, self-image and beliefs-based altruism”. Working Paper. 2013.

[184] Ulrike Malmendier, Vera L. te Velde, and Roberto A. Weber. “Rethinking Reciprocity”. Working Paper. 2013.

[185] Alexandre Mas and Enrico Moretti. “Peers at work”. In: *American Economic Review* 99.1 (Mar. 2009), pp. 112–145.

[186] Robert Mattes and Michael Bratton. “Learning about democracy in Africa: Awareness, performance, and experience”. In: *American Journal of Political Science* 51.1 (Jan. 2007), pp. 192–217.

[187] Dylan Matthews. *In 2011, only 15 senators backed same-sex marriage. Now 49 do.* 2013. URL: `http://www.washingtonpost.com/blogs/wonkblog/wp/2013/04/02/in-2011-only-15-senators-backed-same-sex-marriage-now-49-do/` (visited on 05/18/2013).

[188] Nolan M. McCarty et al. *Polarized America: The dance of ideology and unequal riches.* Cambridge, MA: MIT Press, 2006.

[189] Jonathan Meer. “Brother, Can You Spare a Dime? Peer Pressure in Charitable Solicitation”. In: *Journal of Public Economics* 95.7-8 (Dec. 2011), pp. 926–941.

[190] Stanley Milgram. “Behavioral study of obedience.” In: *Journal of Abnormal and Social Psychology* 67.4 (1963), pp. 371–378.

[191] Arthur M. Okun. *Prices and Quantities: A Macroeconomic Analysis.* Reference, Information and Interdisciplinary Subjects Series - G. Brookings Institution Press, 1981.

[192] Elinor Ostrom. “Collective action and the evolution of social norms”. In: *Journal of Economic Perspectives* 14.3 (2000), pp. 137–158.

[193] Elinor Ostrom, James Walker, and Roy Gardner. “Covenants with and without a sword: Self-governance is possible”. In: *American Political Science Review* 86.2 (1992), pp. 404–417.

[194] Owen Ozier. “The Impact of Secondary Schooling in Kenya : A Regression Discontinuity Analysis”. Working Paper. 2010.

[195] Elizabeth Levy Paluck and Donald P. Green. “Deference, Dissent, and Dispute Resolution: An Experimental Intervention Using Mass Media to Change Norms and Behavior in Rwanda”. In: *American Political Science Review* 103.4 (Oct. 2009), pp. 622–644.

[196] Madan M. Pillutla, Deepak Malhotra, and J. Keith Murnighan. “Attributions of trust and the calculus of reciprocity”. In: *Journal of Experimental Social Psychology* 39.5 (Sept. 2003), pp. 448–455.

[197] Jean-Philippe Platteau. *Institutions, Social Norms and Economic Development.* Amsterdam: Harwood Academic Publishers, 2000.

[198] Richard A. Posner and Eric B. Rasmusen. “Creating and enforcing norms, with special reference to sanctions”. In: *International Review of Law and Economics* 19.3 (Sept. 1999), pp. 369–382.

[199] Andrew Postlewaite. “Social Norms and Social Assets”. In: *Annual Review of Economics* 3.1 (Sept. 2011), pp. 239–259.

[200] M.J.D. Powell. “Variable metric methods for constrained optimization”. In: *Mathematical Programming The State of the Art*. Ed. by Achim Bachem, Bernhard Korte, and Martin Grötschel. New York: Springer Verlag, 1983, pp. 288–311.

[201] Matthew Rabin. “Incorporating fairness into game theory and economics”. In: *American Economic Review* 83.5 (1993), pp. 1281–1302.

[202] Matthew Rabin. “Moral preferences, moral constraints, and self-serving biases”. Mimeo. 1995.

[203] Ernesto Reuben and Arno Riedl. “Enforcement of Contribution Norms in Public Good Games with Heterogeneous Populations”. Working Paper. 2011.

[204] Bruce Rind and Daniel J. Benjamin. “Effects of Public Image Concerns and Self-Image on Compliance”. In: *Journal of Social Psychology* 134.1 (1994), pp. 19–25.

[205] JulioJ. Rotemberg. “Minimally acceptable altruism and the ultimatum game”. In: *Journal of Economic Behavior & Organization* 66.3-4 (June 2008), pp. 457–476.

[206] Alvin E. Roth and Ido Erev. “Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term”. In: *Games and Economic Behavior* 8.1 (1995), pp. 164–212.

[207] Alvin E. Roth, Vesna Prasnikar, and M. Okuno-Fujiwara. “Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study”. In: *American Economic Review* 81.5 (1991), pp. 1068–1095.

[208] Bradley J. Ruffle. “More Is Better, But Fair Is Fair: Tipping in Dictator and Ultimatum Games 1”. In: *Games and Economic Behavior* 23.2 (1998), pp. 247–265.

[209] Kay L. Satow. “Social approval and helping”. In: *Journal of Experimental Social Psychology* 11 (1975), pp. 501–509.

[210] Paul B. Seabright. “Continuous preferences and discontinuous choices: How altruists respond to incentives”. In: *The BE Journal of Theoretical Economics* 9.1 (2009), p. 14.

[211] Jane Sell and Rick K. Wilson. “Levels of information and contributions to public goods”. In: *Social Forces* 70.1 (1991), pp. 107–124.

[212] Nate Silver. *Gay Marriage Opponents Now in Minority*. 2011. URL: `http://fivethirtyeight.blogs.nytimes.com/2011/04/20/gay-marriage-opponents-now-in-minority/` (visited on 05/18/2013).

[213] Joel Sobel. “Interdependent preferences and reciprocity”. In: *Journal of Economic Literature* 151.3712 (Feb. 2005), pp. 867–8.

[214] Adriaan R. Soetevent. “Anonymity in giving in a natural context: a field experiment in 30 churches”. In: *Journal of Public Economics* 89.11-12 (Dec. 2005), pp. 2301–2323.

[215] Douglas Staiger and James H. Stock. “Instrumental Variables Regression with Weak Instruments”. In: *Econometrica* 65.3 (1997), pp. 557–586.

[216] James H. Stock and Motohiro Yogo. “Testing for weak instruments in linear IV regression”. NBER Technical Working Paper Series No. 284. 2002.

[217] Robert Sugden. “Reciprocity: The supply of public goods through voluntary contributions”. In: *The Economic Journal* 94.376 (1984), pp. 772–787.

[218] Jakob Svensson. “Investment, property rights and political instability: Theory and evidence”. In: *European Economic Review* 42.7 (July 1998), pp. 1317–1341.

[219] Vera L. te Velde. “Beliefs-based altruism and motivated reasoning”. Working Paper. 2013.

[220] Joël van der Weele et al. “Resisting moral wiggle room: How robust is reciprocity?” In: *American Economic Journal: Microeconomics* (2014). IZA Discussion Paper No. 5374.

[221] Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press, 2002.

[222] Brigham Young. *Journal of Discourses, Volume 5*. 1858.