Essays on Social Influence in Political Economy: How Expectations and Identity Affect Pro-Social Leading and Following

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:17463132">http://nrs.harvard.edu/urn-3:HUL.InstRepos:17463132</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Essays on Social Influence in Political Economy: How Expectations and Identity Affect Pro-Social Leading and Following

A dissertation presented
by
Mauricio Fernández Duque

to
The Committee on Higher Degrees in Political Economy and Government

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Political Economy and Government

Harvard University
Cambridge, Massachusetts
May 2015
Abstract

By social influence I understand the change in an individual’s thoughts, feelings, attitudes or behaviors that results from interactions with another individual or group. Political, commercial and public health campaigns rely at least partly on influence. Without influence, we have a hard time explaining voter turnout, fads or contagious health behaviors. In my research I focus on pro-social behavior and the de-decentralized provision of public goods, and I ask when and why people are influenced by others as well as when and why people attempt to influence others to “do the right thing”. These questions help us understand human motivation in social contexts, and thus may also help us design policies that can nudge behavior towards more socially desirable, welfare enhancing outcomes.

Despite the importance of influence, its study is scattered across disciplines. In my research, I seek to bridge the disciplinary gap through a three-pronged approach. First, I incorporate concepts found in psychology into a decision-theoretic framework. Second, I experimentally test for hypotheses that are derived from this formalization. Third, I use game theory to derive novel conclusions about how aggregate behavior changes when these concepts are incorporated and propose policy recommendations. My dissertation follows parts of this procedure and points to next steps for two psychology concepts: social identity adoption and social expectations.

In chapter 1, I write down a unifying model of social identity adoption that integrates different strands in the economics and psychology literature. I provide evidence for the main predictions of this model with a large scale field experiment on charitable giving in Mexico.
In chapter 2, joint with Michael Hiscox, we write down a model from which we derive conditions for distinguishing between a social expectations and an altruism explanation to pro-social influence. Results from a laboratory experiment show that most pro-social influence is due to social expectations. In chapter 3, I integrate this social expectations model into a sequential decision setting. I use this to derive a novel model of pluralistic ignorance, and argue that this model explains why uninformed individuals can be leaders in a way past models could not.
Contents

Abstract ................................................................. iii
Acknowledgments ................................................... ix

Introduction 1

1 Priming identity and Public Good Provision 5
  1.1 Introduction .................................................. 5
  1.2 Theoretical Framework ....................................... 10
    1.2.1 Alternative Rational Explanations ......................... 18
  1.3 Experimental Design ........................................ 21
    1.3.1 Our Partner ........................................... 21
    1.3.2 The Design ........................................... 21
    1.3.3 Randomization ......................................... 23
  1.4 Empirical Specification and Balance ...................... 26
  1.5 Results ..................................................... 31
    1.5.1 Nationalists increase contribution more with prime .... 31
    1.5.2 Only regionalists contribute relatively more to close states 33
    1.5.3 Regionalist’s bias diminishes with prime ............... 33
    1.5.4 All contribute more towards the poor states, especially nationalists 34
    1.5.5 Pro-poor bias diminishes with prime, especially for Nationalists ... 36
  1.6 Conclusions .................................................. 37

2 Social expectations in pro-social influence 39
  2.1 Introduction ................................................ 39
  2.2 Related Literature ......................................... 44
  2.3 Conceptual Framework ...................................... 47
  2.4 The Experimental Design .................................. 51
  2.5 Results ..................................................... 58
    2.5.1 Incorporating Beliefs .................................. 60
    2.5.2 Robustness Checks .................................... 63
  2.6 Conclusion .................................................. 64
List of Tables

1.1 Distribution of recipient cards by donor ............................................. 26
1.2 Summary statistics of covariates .......................................................... 26
1.3 Summary statistics of dependent and independent variables ..................... 30
1.4 Balance Table ....................................................................................... 32
1.5 Nationalist Prime and Pride ................................................................. 33
1.6 Regional bias for Regionalists and Nationalists ....................................... 34
1.7 Triple interaction with nationalist pride .................................................. 35
1.8 Poverty Bias ......................................................................................... 36
1.9 Poverty Bias and Prime ......................................................................... 37

2.1 Summary statistics .................................................................................. 58
2.2 Can and Cannot Consider Decisions ....................................................... 60
2.3 Decision To Replicate The Can Consider Scenario .................................. 61
2.4 Impact and Marginality Conditions ......................................................... 62
2.5 Types of Leaders .................................................................................... 63

A.1 Balance Table with Municipal GDP pc as a measure of regionalism .......... 109
A.2 Nationalist Prime and Pride .................................................................. 110
A.3 Regional bias according to pride ............................................................ 110
A.4 Triple interaction with nationalist pride ................................................ 111
A.5 Poverty Bias ......................................................................................... 112
A.6 Poverty Bias and Prime ....................................................................... 112
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Store placement</td>
<td>22</td>
</tr>
<tr>
<td>1.2</td>
<td>Card treatments</td>
<td>23</td>
</tr>
<tr>
<td>1.3</td>
<td>Donor stores and recipient states</td>
<td>24</td>
</tr>
<tr>
<td>2.1</td>
<td>Introductory Presentation Of Scenarios</td>
<td>54</td>
</tr>
<tr>
<td>2.2</td>
<td>Questionnaire of Introductory Presentation Of Scenarios</td>
<td>54</td>
</tr>
<tr>
<td>2.3</td>
<td>Questionnaire Answers of Introductory Presentation Of Scenarios</td>
<td>55</td>
</tr>
<tr>
<td>2.4</td>
<td>Explanation Of First Contribution For Leaders in Treatment 1</td>
<td>56</td>
</tr>
</tbody>
</table>
Acknowledgments

This work couldn’t have been completed without the help and support of many people who were there for me professionally and personally. Samuel Asher is the first to come to mind, as he has helped me out on both accounts. He’s been about as good friend as anyone can ask for, and turned the PhD around for me when it was at its darkest. Along with Michael Hiscox, they made sure I didn’t give up when I wanted to and helped me think through and mature my ideas. If the ideas contained herein are not particularly mature, then I probably should have paid more attention. My secondary advisors Iris Bohnet and Nava Ashraf took a thoughtful look at my work and I am very thankful of their insights and encouragement. There were a lot of professors, classmates and students whose conversations and knowledge I benefited from, too numerous to mention. My partnership with Un Kilo de Ayuda gave me the opportunity to think about how what I learn in the classroom applies in the field, and what the field can teach the classroom. The friendship, backing and advice from Jose Ignacio Ávalos and the team he oversees has been invaluable.

Personally, my parents and brother have been a constant source of support. As time goes by, I realize more and more how lucky I am to have them in my life. I must not forget my friends from Mexico - Adriana, Aline, Daniel, Helga, Gabriel, Gustavo, Jorge, Juan Pablo, Liv, Mariana, Maki, Tatiana, Victor and Yokoyani - as they would get upset. So would Nicholas, Philip, Radhika, Sarah and Serena, most likely. Your names are alphabetized, guys. I love you all and you have helped me laugh and dance through the hard times.
Introduction

By social influence I understand the change in an individual’s thoughts, feelings, attitudes or behaviors that results from interactions with another individual or group. Political, commercial and public health campaigns rely at least partly on influence. Without influence, we have a hard time explaining voter turnout, fads or contagious health behaviors. In my research I focus on pro-social behavior and the de-centralized provision of public goods, and I ask when and why people are influenced by others as well as when and why people attempt to influence others to “do the right thing”. These questions help us understand human motivation in social contexts, and thus may also help us design policies that can nudge behavior towards more socially desirable, welfare enhancing outcomes.

Despite the importance of influence, its study is scattered across disciplines and arguably does not occupy the central role it should. Although influence is a unifying concept in political science that could allow us to better understand how redistributive deals are reached, policies are supported and mobilization is achieved, there is not a systematic treatment of the concept from which these diverse phenomena are considered. Despite advances in behavioral economics, economists’ analysis of behavior change often crowd out influence-based explanations by their focus on incentives-based explanations. This is detrimental to both explanations, as we can improve our understanding of each by delimiting at their domains and exploring their interactions. Psychology is perhaps the discipline that has given the most central role to influence. However, psychologists typically are not focused on thinking about the aggregate result of behavioral tendencies such as markets and political decision making so central to the other disciplines. They are not
sufficiently studied through the lens of influence. Further, the analytical tools used by psychologists are usually hard to interpret to rational choice-minded social scientists.

In my research, I seek to bridge the gap between these disciplines through a three-pronged approach. First, I incorporate concepts found in psychology into a decision-theoretic framework. Second, I experimentally test for hypotheses that are derived from this formalization. Third, I use game theory to derive novel conclusions about how aggregate behavior changes when these concepts are incorporated and propose policy recommendations. My dissertation follows parts of this procedure and points to next steps for two psychology concepts: social identity (paper 1) and social expectations (papers 2 and 3).

*Priming identity and Public Good Provision.* Social identity theory states that individuals assume one of several possible identities, and that the identity they assume affects their behavior. There are several explanations of how an identity is chosen. The explanation that is most commonly considered by economists and formal political science is a ‘motivated cognition’ explanation, in which individuals choose the identity that maximizes their status or self-esteem. An alternative explanation is a ‘categorical thinking’ explanation, where an uninformative prime makes an individual unconsciously think about their identity differently. This latter explanation is a better fit for why Asian women do better on math exams when they are asked about their nationality instead of their gender. In our first paper, I write a model that combines both explanations of identity adoption. We then test hypotheses from this model in a large scale field experiment. In this experiment, we varied the content of donation cards that are bought in supermarkets in Mexico and have no added value. The cards asked for a donation to a specific state in Mexico. We varied the recipient state to test for regional bias, which is a politically relevant cleavage that is understudied in this context. We then varied whether the card primed nationalism by varying colors in the card, either with nationalist colors or without nationalist colors. As predicted from the model, we find heterogeneous effects depending on the regionalism of the state. We conclude that a persuader must take into account the identities individuals prefer due to a motivated cognition channel in order to understand how priming different identities
will affect their choice. We include a discussion of how to more carefully distinguish the theoretical channel, as well as discuss policy implications to be developed in more fully in subsequent work.

Psychologists were the first to point out that individuals behave differently depending on their perception of social expectations (e.g. the Asch experiments on conformity). Economists have explored different ways in which social expectations may affect behavior, focusing on reputational or preference complementarities. There is more recent work suggesting that social expectations might be a motivator unto itself (e.g. the Dana et al experiments). In the second and third paper of my dissertation, I focus on the impact of social expectations as a direct motivation on influence.

Social expectations in pro-social influence. In this paper we provide evidence for social expectations as affecting the motivation to influence. An individual might make a higher anonymous contribution to a charity when in a position to influence for different reasons. He may do it independently of what others think (to be pro-social, to self-signal or to meet a code of conduct) or because he cares about what others will think (meeting social expectations). Note that being in a position to influence means that the contribution is public, so these motivations are typically confounded. In order to disentangle these motivations, we write down a model with both types of motivations and design a lab experiment that distinguishes between the motivations according to the model. We first give individuals the opportunity to give more if their decision is public. We then experimentally disentangle whether those who gave more were motivated by caring what others think. We find that 20 out of 84 individuals try to influence, and 17 out of those 20 did so to meet social expectations. In the next paper I plug the model developed here into a sequential decision setting.

Uninformed leaders. The Arab Spring has brought revolutions back into the spotlight of media and academics. Rational explanations of social revolutions tend to assume that getting the revolution started requires an extremist, a prominent individual or someone with privileged information. We argue that neither view fits the facts of what started the
Tunisian uprising, and propose a novel model of pluralistic ignorance that shows how an obscure, uninformed non-extremist can start a revolution. In the model, individuals have heterogeneous preferences over a public good and some have social expectations like in Paper 2. Unlike standard models of revolutions, individuals prefer to do what others want, not what others do. Sequential public good decisions may lead to the spread of misinformation about what others want, and thus to pluralistic ignorance. In this context, an individual who has no privileged private information may break the pluralistic ignorance if he is known to be representative of the distribution of preferences over the public good and if he takes an action that is known to reflect his personal preferences. We will argue that the model is a more natural explanation for several phenomena usually explained by standard herding approaches.
Chapter 1

Priming identity and Public Good Provision

1.1 Introduction

There is a broad family of theories in psychology with the central tenet that social identity is somehow chosen and affects behavior in important ways (e.g. the review in Swann and Bosson, 2010). The choice of identity has gained some traction in rational choice political science and economics, used to explain the size of the welfare state (Shayo, 2009), sectarianism (Penn, 2008), patience (Benjamin, Choi and Strickland, 2010), gender discrimination, social exclusion and the household division of labor (Akerlof and Kranton, 2000, 2010). Indeed, some scholars claim that identity choice is one of the most fundamental choices individuals make (Akerlof and Kranton, 2000), which determines what type of people and activities they seek out. However, there is still a fair amount of lingering doubt among rational choice scholars over the usefulness and definition of the concept of identity. Part of this may be due to the fact that some studies find identity to be highly manipulable, responding to environmental cues (e.g. Benjamin, Choi and Strickland, 2010). How can such a fundamental choice be at the same time so malleable?

In psychology, there are several answers to the question of how individuals adopt their
identity. However, the standard theories of identity choice in economics follows a specific answer, which is that individuals choose their identity to maximize their self esteem or social standing (Akerlof and Kranton, 2000; Shayo, 2009; Benabou and Tirole, 2011). I will refer to this as the motivated cognition approach. An important alternative view of identity choice in psychology is that individuals are hard-wired to think in terms of categories, and individuals have different identities that correspond to different categories through which they make decisions. I will refer to this alternative as the categorization approach. In a classic example, Shih, Pittinsky and Ambady, 1999 show that if Asian American women are primed to think about their race, they do better in a math exam than the control group, but do worse if primed to think about their gender. These behaviors are stereotype-consistent with an identity that is exogenously made salient. Further, one of these identities makes them perform worse. These facts are hard to reconcile with the idea that these women are choosing an identity to maximize their self esteem. In economics, Benjamin et al (2010) find different discount factors when a specific identity is made salient.

Despite different views on how identity is chosen, social identity theory consistently predicts that individuals are biased in favor individuals they identify with. If individuals’ contributions towards local public goods depend on whether they identify with the beneficiaries, how much can individuals’ contributions change based on manipulating their identity? Work in social psychology and political science has found that contributions towards others can be increased if individuals are led to see them as part of their in-group (e.g. Gaertner, 1999). However, if identity is partly a fundamental choice based on motivated cognition, shouldn’t we expect limits to this manipulation? In this paper we write down a model and explore these questions in the context of a field experiment on charitable giving in Mexico. Individuals will be making a choice to donate to an NGO focused on eradicating child malnutrition in Mexico. They choose whether to support the NGO’s efforts in a specific state. The choice either primes nationalism or not.

The starting point of our model is that from the available identities individuals can adopt, individuals prefer some over others. For example, those individuals for whom identifying
with their region gives a higher self esteem or utility than identifying with their nation will prefer their regional identity. Nationalistic individuals, in contrast are those who prefer their nationalist identity. This preference over identities is given by a motivated cognition channel. For example, individuals whose region is relatively richer than the national average may prefer to identify with their region than with the nation. We take these identity preferences as given and introduce the categorization component by building on the framework of categorical thinkers developed by Mullainathan, Schwarzstein and Shleifer, 2008.

Categorical thinking is a type of departure from Bayesian thinking that has received a lot of support in the psychology literature. Individuals group situations into categories, and cannot distinguish between co-categorized situations. Suppose individuals are asked to help the efforts of the NGO in a state \( Z \). The donation petition may be accompanied by a prime that highlights national pride. The prime affects how decision makers co-categorize their situations. Critically for our theory, this co-categorization in turn affects the identity they adopt to make their decision. If the prime does not highlight nationalist pride, they co-categorize the situation as one of ‘an individual giving to state \( Z \)’, for which they use their preferred identity. In this case, regionally biased individuals act more favorably towards \( Z \) if the state is in their region, since these states are part of \( Z \)’s ‘in-group’. Nationalists, in contrast, don’t discriminate among states based on their proximity, because all states in Mexico are part of their in-group. Nationalists are therefore more sensitive to the poverty of the state. If the prime does highlight nationalist pride, they co-categorize the situation as ‘a Mexican giving to a Mexican state’, for which they use their nationalist identity. Because the contribution is now seen as benefiting a Mexican state instead of the specific state \( Z \), both the regionalist’s and the nationalist’s biases (towards close and poor recipient states, respectively) diminish. Further, since the regionalists are now making a decision through their nationalist identity, they weigh in-group members less in their utility function. From this model, we formally derive and test six implications for nationalists and regionalists regarding their poverty bias, regional bias and the interaction of a bias with a prime.

In order to test our theory, we ran a large scale field experiment in 446 supermarkets in
9 states in Mexico. We partnered with Un Kilo de Ayuda, a Mexican NGO who provides food aid and nutritional services to poor families in rural Mexico. One of their fundraising strategies is to offer otherwise valueless cards in supermarket cashiers with the NGO’s logo and a plea to donate. These dollar cards are given by the customer to the cashier who scans it and the money is collected by the NGO. We introduced a novel design of this card, with a picture of children and a text that read ‘Thank you for donating to state Z.’ We varied the design of these cards along two dimensions. The first dimension was the state Z mentioned. We limited Z to take on the value of one of six states where the NGO operated. The second dimension which we varied was whether we primed nationalism. In order to do this, one version of the cards introduced the colors of the Mexican flag into the margins and text of the design. These colors were the only thing that changed in the design. This is important, as we will interpret this treatment as an uninformative message.

Our results support our predictions. First, we find that the prime affects nationalists more positively than regionalists. Second, we find an overall bias towards benefiting one’s region, which is only significant for regionalists. Third, we find weaker but consistent evidence that this bias displayed by regionalists diminishes with the prime, and this interaction is significantly different from the nationalist’s. Fourth, there is an overall bias towards benefiting the poorest recipients states, and fifth, this bias is only significant for nationalists. Sixth, we have some mixed evidence that this pro-poor bias diminishes more strongly for nationalists than for regionalists with the prime, but we argue that these mixed results are favorable to our interpretation.

Taken together, our results indicate that priming individuals to take on a specific identity will have different effects depending on what their preferred identities are. Regionalists will act with less of a bias against out-group Mexicans, but they will also be less other regarding overall. Nationalists will be less sensitive to poverty, which may be compensated by an increase in contributions. Having a better understanding of individuals’ identities can help increase local public good provision through judicious use of uninformative primes. Further, priming identities may lead to efficiency-equity tradeoffs in terms of the social welfare effect
of total voluntary contributions. On the theoretical side, by linking two strands of social identity theory we hope to contribute to pinning down and spurring further work on this elusive but important concept.

In addition to the literature on identity, this paper contributes to diverse strands of literatures. Substantively, we contribute to the understanding of nationalist sentiment and regionalism. Although what exactly nation building means is somewhat fuzzy (Meadwell, 1991; Eriksen, 1993; Herbst, 2000; Keating, 2003), a large part of it involves developing a common identity that will increase cooperation (Andersen, 1983, Miguel, 2004). Within this debate, a natural cleavage to look at is regionalism (Tewdwr-Jones et al, 2006; Keating, 2003), but it has been overlooked by the focus on ethnic (Miguel, 2004; Transue, 2007) and religious cleavages (Charnysh, Lucas and Singh, 2012; Sachs, 2009). Mexico is no exception to the nation-building enterprise or to regionalism (Brading, 1985; Van Young, 1992). During the XXth century, the PRI embarked on an intense nation-building project. However, regional strongholds had to be integrated into this project. Indeed, there is still a debate over whether the nationalist project erased regionalism (e.g. Brading, 1985; Pereyra, 2007; Boyer, 2011). Some argue that migration (Wayne, 1973) and democracy (Baker, 2009; Dominguez and McCann 1996; Klesner 1993, 1995; Magaloni 1999; Poire 1999) have increased regionalism. Our work provides novel evidence for regional bias in Mexico, and our theoretical work suggests that priming a national identity may have unintended consequences if the preferred identity of the population is not well understood or motivated to change.

To the extent that charitable decisions are indicative of broader social preferences, our work further contributes to the literature on preferences for redistribution (Alesina and La Ferrara, 2005; Alesina and Giuliano, 2009) and how these preferences depend on nationalism (Ahlerup and Olsson, 2007).

Although other papers have looked at the impact of a national prime on cooperation across cleavages (e.g. Charnysh, Lucas and Singh, 2012), this is the first large scale field experiment exploring this question. The fact that we are study several recipients for several states (as opposed to the in-group preferences of two groups) allows us to get variation on
regional bias as well as income disparities. The focus on only two groups has inevitably confounded differences across groups that we show impact contribution decisions. Further, this aids our understanding of social preferences, where there is a debate regarding the rationalizability of these preferences (e.g. Andreoni and Miller, 2002, Charness and Rabin, 2002). Our work provides a novel rationale for why social preferences are not rationalizable. Finally, our paper contributes to the large literature on fundraising (e.g. Shang and Croson, 2009; DellaVigna, List and Malmendier, 2014), by determining the conditions under which regional appeals and nationalist primes can raise contributions.

The rest of this paper is organized as follows. Section II presents the theoretical framework. Section III describes the NGO and the experimental setting in more detail. Section IV describes the empirical specification. Section V presents the results. Section VI concludes. The results for an alternative measure of nationalism is presented in the appendix.

1.2 Theoretical Framework

Suppose there is a decision-maker playing a dictator game in which he must decide how to split \( d \in [0, 1] \) between himself and a recipient. The recipient is a poor person of state \( Z \) in the decision-maker’s country. There are four types of recipient states: they are either rich or poor and they are either close or far from the decision-maker.

The decision-maker has social preferences that depend on the recipient’s status as an ‘in-group’ or an ‘out-group’ member. These will be partly determined by a motivated cognition channel and partly through a categorization channel. ‘In-group’ recipients have initial wealth \( y_{in} \) and utility \( u_{in}(y) \), while out group recipients have \( y_{out} \) and \( u_{out}(y) \). The decision-maker has utility \( u_i(1 - d, u_{in}, u_{out}) \). The decision-maker is faced with either an in-group or an out-group member as the recipient. If he is faced with an in-group recipient, he solves

\[
\max_{d \in [0, 1]} u(1 - d, u_{in}(y_{in} + d), u_{out}(y_{out}))
\]
If he is faced with an out-group recipient, he solves

$$\max_{d \in [0,1]} u(1 - d, u_{in}(y_{in}), u_{out}(y_{out} + d))$$

There are two types of decision makers $\theta \in \{R, N\}$, the ‘Regionalist’ type and the ‘Nationalist’ type. These types are determined exogenously in this model, and represent the motivated cognition aspect of identity choice. The motivated cognition approach to social identity theory claims that individuals choose the identity that maximizes their self esteem (e.g. Shayo, 2009). Individuals who live in regions that are relatively better off identify with their region, while the rest identify with their nation. We assume that regionalist decision makers consider recipients in close-by states as part of their in-group, and the rest as part of their out-group. In contrast, the high pride decision makers consider all recipients in states within the country as part of their in-group.

For simplicity, we will assume the following functional forms:

$$u = (x - d)^{\alpha} + u_{in}^{\beta} + u_{out}^{\gamma}, \quad 1 > \beta > \gamma > 0, \quad 1 > \alpha > 0$$

$$u_{in}(y - d) = u_{out}(y - d) = y - d$$

It is straightforward that the optimal $d^*$ increases in the recipient state’s wealth, and that decision makers are biased towards their in-group.

We are interested in studying how an uninformative message may change the identities and behavior of different types. Therefore, we turn to a model of uninformative persuasion. In order to do so, we follow the model of Mullainathan, Schwarzstein and Shleifer (2008) closely.

There are three similar, but observably distinct situations $s \in \{0, 1, 2\}$ under which the decision-maker is making a decision. If $s = 0$, then the situation is that of a decision-maker is deciding whether to make a contribution to a recipient of state $Z$. If $s = 1$, then the situation is that of a Mexican deciding whether to make a contribution to a recipient in Mexico. If $s = 2$, then the situation is that of the decision-maker is deciding whether to give to a state that is far/close and rich/poor. The prior probabilities of these situations are
\[ p(s = 0) > 0, \ p(s = 1) > 0 \ \text{and} \ p(s = 2) > 0. \]

We will further specify our assumptions over the decision problems faced by \( N \) and \( R \) types. Critically, the situations not only specify who the recipient is but also which identity they adopt when making the decision. Individuals' utility is affected by the identity they've adopted to make a decision as well as their preferences over identities given by the motivated cognition channel. When the situation does not prime individuals to adopt a certain identity, they choose their preferred identity. In all situations, \( N \) types adopt the nationalist identity, which is their preferred identity. However, \( R \) types adopt the identity of regionalism in situations \( s = 0 \) and \( s = 2 \), and through the identity of nationalism in situation \( s = 1 \). That is, they are not choosing through their preferred identity in \( s = 1 \).

In \( s = 0 \) and \( s = 2 \), only regionalist decision-makers care about the far/close dimension when making their contribution decision. That is, since \( N \) types see all of the states as part of their in-group, the utility of other states is similarly weighted for states of different proximity. For \( R \) types, in these situations far states are part of their out-group. The utility functions that an \( R \) and \( N \) type maximize for \( s \in \{0, 2\} \) are, respectively:

\[
u(d; s, \theta = R) = (x - d)^\alpha + (y_Z + d)^\beta + 1 \gamma - \beta
\]

\[
u(d; s, \theta = N) = (x - d)^\alpha + (y_Z + d)^\beta
\]

where 1 is an indicator function equal to 1 if \( Z \) is far. Notice that the regionalist decision-maker treats \( Z \) as an in-group member only if the state is close. The nationalist decision-maker, on the other hand, treats all states as in-group members.

If individuals are in \( s = 1 \), they are choosing how much to contribute to a recipient in Mexico. This recipient has initial wealth \( y_M = E(y_Z) \), which is lower than the wealth of a rich recipient state and higher than the wealth of a rich recipient state. Because this identity is not the \( R \) types' preferred, he values in-group members less than the in-group members of his preferred identity. We capture this by assuming that the marginal utility of benefiting an in-group members in \( s = 1 \) is \( \delta \) what it would be in \( s = 0 \) or \( s = 2 \).
The utility functions that an $R$ and $N$ type maximize for $s = 1$ are, respectively:

$$u(d; s = 1, \theta = R) = (x - d)^{\alpha} \delta (y_M + d)^{\beta}$$

$$u(d; s = 1, \theta = N) = (x - d)^{\alpha} (y_M + d)^{\beta}$$

There is another player who sends a message to the decision-maker before the latter’s decision is made. Think of the sender as the NGO who is choosing the advertisement, and the message as the nationalist prime. That is, when thinking about this message, fix the state $Z$ the card is indicating is the recipient state and concentrate on the variation in colors. We will therefore represent the message space as $m \in \{FC, S\}$, to denote whether the message has the flag’s colors or is a standard version. The sender has private information about the recipients. Since we are interested in arguing that message works because it is a prime that affects categorical thinking, we need to be precise as to why the message does not affect a Bayesian consumer. Our main assumption is that the message is uninformative. Given public information $r$, the Bayesian can observe that the true situation is $s = 0$. The interpretation is that the consumer believes the card is going to state $Z$, and does not place any positive probability on the card going to another nearby/far and rich/poor state or to another state in Mexico. The public information $r$ is the state $Z$ the card says will be the beneficiary state. Since $p(s = 0|r) = 1$ for the Bayesian, any message about the situation is irrelevant. We highlight this in the following:

**Fact 1.** The decision $d^*$ by a Bayesian consumer with $p(s = 0|r) = 1$ is unaffected by a message about $s$.

Now consider a categorical thinker. As opposed to a Bayesian thinker, categorical thinkers group situations into categories, and cannot distinguish between co-categorized situations. In particular, the categorical thinker cannot distinguish $C_1 = \{s_0, s_1\}$ or $C_2 = \{s_0, s_2\}$. That is, he either co-categorizes giving to state $Z$ as giving to (a) a Mexican state or (b) a far/close state (for regionalist decision-makers) or a rich/poor state (for nationalist decision-makers).
makers). The ‘choice’ of category is an automatic process. If the chosen category is $C_1$, then
\[ p(s = 0|C_1) = \frac{p(s = 0)}{p(s = 0) + p(s = 1)} \]
and the other probabilities are defined similarly.

In order to select the category through which he makes a decision, he ‘chooses’ the category $C(m)$ that is most likely given his information:
\[ C(r, m) \equiv \arg \max_{C \in \{C_1, C_2\}} \hat{p}(s \in C|m, r) \]

In order to unpack this decision, we need to understand the sender’s private information a bit better. We model the sender’s private information as being informative only in situations $s = 1$ and $s = 2$. That is, the sender observes private signal $i \in \{\text{far/close and rich/poor state, Mexican state}\}$. We further assume for simplicity that this private signal is perfectly informative to the sender: $p(s = 1|i = \text{Mexican state}, r) = p(s = 2|i = \text{far/close and rich/poor state}, r) = 1$. The interpretation is that the NGO is privately perfectly informed of whether the card benefits a state that is far/close and rich/poor from the perspective of the decision maker, or if it benefits a Mexican state. The sender observes his private signal, then sends a message to the decision-maker. The decision-maker holds beliefs over the probability of seeing the message given the situation and the sender’s private information, denoted by $\hat{\sigma}(m|i, s, r)$ and updated using Bayes’ rule whenever possible. The decision-maker updates his priors over the situation given the message he observes, and chooses $d$.

We can now more concretely specify that category $C_1$ is chosen if and only if:
\[
\hat{p}(s \in C_1|m, r) \geq \hat{p}(s \in C_2|m, r) \iff \hat{p}(s = 1|m, r) \geq \hat{p}(s = 2|m, r) \iff \\
\left(\sum_{i' \in i} \hat{\sigma}(m|i', s = 1, r)p(i'|s = 1, r))p(s = 1) \right) / \left(\sum_{s' \in \{0,1,2\}} \left(\sum_{i' \in i} \hat{\sigma}(m|i', s', r)p(i'|s', r))p(s') \right) \geq \right.
\left. \left(\sum_{i' \in i} \hat{\sigma}(m|i', s = 2, r)p(i'|s = 2, r))p(s = 2) \right) / \left(\sum_{s' \in \{0,1,2\}} \left(\sum_{i' \in i} \hat{\sigma}(m|i', s', r)p(i'|s', r))p(s') \right) \right) \iff \\
\left(\sum_{i' \in i} \hat{\sigma}(m|i', s = 1, r)p(i'|s = 1, r))p(s = 1) \right) / \left(\sum_{s' \in \{0,1,2\}} \left(\sum_{i' \in i} \hat{\sigma}(m|i', s', r)p(i'|s', r))p(s') \right) \geq \\
\left(\sum_{i' \in i} \hat{\sigma}(m|i', s = 2, r)p(i'|s = 2, r))p(s = 2) \right) / \left(\sum_{s' \in \{0,1,2\}} \left(\sum_{i' \in i} \hat{\sigma}(m|i', s', r)p(i'|s', r))p(s') \right) \right) \iff \\
\left(\sum_{i' \in i} \hat{\sigma}(m|i', s = 1, r)p(i'|s = 1, r) \right) p(s = 1) \geq \\
\left(\sum_{i' \in i} \hat{\sigma}(m|i', s = 2, r)p(i'|s = 2, r) \right) p(s = 2) \right)
\]
\[
\left( \sum_{i' \in i} \hat{\sigma}(m|s = 2, r)p(i'|s = 2, r) \right) p(s = 2) \Leftrightarrow \\
\hat{\sigma}(m|s = 1)p(s = 1) \geq \hat{\sigma}(m|s = 2)p(s = 2)
\]

In this setting then, a message that is uninformative to the Bayesian can affect the categorical decision-maker by changing how he co-categorizes a situation. If the categorical thinker believes that the sender is sufficiently more likely to send the message with the flag colors under \(s = 1\) than \(s = 2\), the uninformative message changes the way the decision-maker categorizes the decision. The interpretation is that the decision-maker may be led by the sender to think of the decision as one of giving to 'state Z, a Mexican state' as opposed to 'state Z, a far/close and rich/poor state'. Mullainathan, Schwarzstein and Shleifer (2008) refer to this uninformative persuasion as 'framing' the decision.

In order to see how the decision-maker’s decision is affected by framing, we must write out the expected utility of the categorical thinker explicitly. The expected utility of a categorical thinker who is using category \(C_i\) for \(i \in \{1, 2\}\) is:

\[
E^{C_i}U = p(s = 0|C_i)u(s = 0, \theta) + p(s = i|C_i)u(s = i, \theta)
\]

We can now explicitly write out the argument a categorical thinker maximizes depending on his category and type. An R type using \(C_2\) and giving to state Z chooses \(d_{Z_1,R}^2 \in [0, x]\) to maximize:

\[
(x - d_{Z_1,R}^2)^\alpha + (y_Z + d_{Z_1,R}^2)^{\beta + 1(\gamma - \beta)}
\]

An R type using \(C_1\) and giving to state Z chooses \(d_{Z_1,R}^1 \in [0, x]\) to maximize:

\[
(x - d_{Z_1,R}^1)^\alpha + p(s = 0|C_1)(y_Z + d_{Z_1,R}^1)^{\beta + 1(\gamma - \beta)} + \delta p(s = 1|C_1)(y_M + d_{Z_1,L}^1)^\beta
\]

An N type using \(C_2\) and giving to state Z chooses \(d_{Z_2,N}^2 \in [0, x]\) to maximize:

\[
(x - d_{Z_2,N}^2)^\alpha + (y_Z + d_{Z_2,N}^2)^\beta
\]
An N type using C₁ and giving to state Z chooses $d_{1,N}^Z \in [0, x]$ to maximize:

$$(x - d_{1,N}^Z) + p(s = 0|C_1)(y_Z + d_{1,N}^Z) + p(s = 1|C_1)(y_M + d_{1,N}^Z)$$

Deriving the optimal amount in each of these cases is straightforward. For the results below, we will assume that for all decision makers the wealth and proximity of a recipient states are independent. Further assume that a group of decision makers are given different messages with equal probability and faced with the same distribution of states. In order to state our main result, we will use the notation $E_Z(d_{C,θ})$, $E_C(d_{θ}^Z)$ and $E_p(d_{C,θ}^Z)$ to respectively denote the expected contribution over recipient states, categories and state proximities. By comparing the optimal amounts derived from the expressions above, we can derive the following:

**Result 1.** Suppose R and N type decision makers are categorical thinkers. Then if Z is close, Z' is far and both have the same wealth:

1. Nationalists increase contributions more than regionalists with prime for δ low enough.

$$E_Z(d_{1,N}) - E_Z(d_{2,N}) > E_Z(d_{1,R}) - E_Z(d_{2,R})$$

2. Only regionalists contribute relatively more to close states.

$$E_C(d_{R}^Z) - E_C(d_{R}^{Z'}) > 0 = E_C(d_{N}^Z) - E_C(d_{N}^{Z'})$$

3. Regionalist’s bias diminishes with prime

$$(d_{2,R}^Z - d_{2,R}^{Z'}) - (d_{1,R}^Z - d_{1,R}^{Z'}) > 0 = (d_{2,N}^Z - d_{2,N}^{Z'}) - (d_{1,N}^Z - d_{1,N}^{Z'})$$

If Z is poor and Z’ is rich:

4. All contribute more towards the poor states

$$E_C(d_{θ}^Z) > E_C(d_{θ}^{Z'})$$

5. Nationalists contribute relatively more than regionalists to poor states when states are far.
Otherwise, they contribute the same relative amount. For $C \in \{1, 2\}$,

$$d_{C,N}^Z - d_{C,N}^{Z'} \geq d_{C,R}^Z - d_{C,R}^{Z'}$$

with the inequality strict when $Z$ and $Z'$ are far.

6. The pro-poor bias diminishes with the prime, more so for Nationalists when states are far.

$$(d_{2,N}^Z - d_{2,N}^{Z'}) - (d_{1,N}^Z - d_{1,N}^{Z'}) \geq (d_{2,R}^Z - d_{2,R}^{Z'}) - (d_{1,R}^Z - d_{1,R}^{Z'}) > 0$$

with the inequality strict when $Z$ and $Z'$ are far.

Result 1 shows a series of comparative statics that follow from comparing the four maximization decisions of $R$ and $N$ type categorical thinkers. Result 1.1 states that nationalist types increase contributions more with the prime, as making a choice to benefit others under a less preferred identity decreases contributions. Result 1.2 states that $R$ types have a bias towards giving to close states, which $N$ types do not. This is a direct consequence of how the regionalists’ preferred identity puts more weight on close states. Result 1.3 states that the bias displayed by the regionalists towards close states diminishes with the prime, as putting more weight on being in a situation in which they are ‘giving to a Mexican state’ makes decision makers less sensitive to the proximity of the recipient state. Result 1.4 says that all states contribute more to poor recipient states, Result 1.5 further specifies that nationalists contribute more on average to poor states, and result 1.6 specifies that the bias diminishes for both types with the prime but more so for nationalists. Thus, the model predicts different types of biases depending on the main identity players adopt through a motivated cognition channel, and predicts that this main bias diminishes with the prime.

There are further results that can be derived from this model, regarding triple or quadruple interactions. We refrain from those because they become statistical results of higher order interactions can be explained by different mechanisms and require more power.

In the empirical section, we will design an experiment that tests these predictions.
1.2.1 Alternative Rational Explanations

This paper has made an assumption about the Bayesian decision-maker’s beliefs. If we relax this assumption, we can offer alternative explanations for the theoretical results that are consistent with a rational actor. In this subsection, we will sketch out these explanations. These explanations potentially fit all the facts of the experiment. Other explanations, like that some colors are more attention-grabbing, might explain a difference in overall sales but not a the results predicted by the model. Arguments regarding a higher social reputation of buying one type of card over the other are also insufficient in and of themselves to explain the behavior patterns that are predicted. We argue that the alternative explanations given below are the mechanisms through which a social reputation story would need to go through in order to explain these patterns. We will sketch out how to test these alternative explanations with follow up experiments or surveys.

Suppose that a Bayesian does not believe that a card that is going to state \( Z \) will go exclusively to state \( Z \). If the message \( m \in \{ FC, S \} \) increases the Bayesian’s beliefs that they are in situation 2, then this rational actor would act the same way as the categorical thinker. The intuition would be that the Bayesian believes that the card that says it will benefit state \( Z \) will perhaps partly benefit state \( Z \), but it will also benefit other states. The flag colors makes the Bayesian infer that the states that will be benefited will be distributed more evenly across Mexico than he would have thought without the priming. There are two reasons why the nationalist message would make the Bayesian change his beliefs. The first is that the Bayesian thinks that the message is informative of some private information the sender is trying to convey. For example, the NGO is conveying to the Bayesian through the colors that what the card means when it says it will benefit state \( Z \) is that it will benefit state \( Z \) among other Mexican states. The second reason is that the Bayesian distrusts the message of the sender less, who was trying to get the Bayesian to believe that the proceeds were going to state \( Z \). So the nationalist message may make the Bayesian take the NGO less seriously in its claim to target the proceeds. In either case, the nationalist prime leads the Bayesian to believe that the proceeds are going to a wider set of states than the card without
the prime. In contrast, the nationalist prime leads the coarse thinker to think of state Z as a Mexican state more than it does without the nationalist prime. Note that these are testable implications. Individuals may be asked to predict the recipient states of the donation with both types of cards.

One can build an alternative rational explanation of why the prime works that relies on the decision-maker’s trust of the NGO’s differential efficiency. This explanation is outside of the current model, and will only be sketched. Suppose the pattern of contributions without a prime is explained by the decision-maker’s belief over how efficiently the NGO works in different states. Then the prime may work through changing the decision-maker’s trust in the NGO’s efficiency in some states. For example, take the regionalist states. They may believe that the NGO is less efficient in further away states, perhaps because they are harder to verify. Then the prime may change this belief by making the decision-maker believe that they are equally efficient across states. Perhaps they learn something about the NGO’s efficiency through observing what the NGO thinks it knows about fundraising through its nationalist message. This is a just-so story that gets even more convoluted in order to simultaneously explain as the differential contribution of high pride states, but it is one we cannot rule out with the data. In order to test for this, we could see whether the perception of efficiency is affected by exposure to different cards in the way predicted by this theory.

A third concern is that the prime may be sending an unexpected message that is having an effect of decisions that are unintended. The main candidate in this regard is that the message is sending a political message. In particular, the colors of the flag are the same colors of the party who was hegemonic for 70 years during the twentieth century. On the other hand, the standard card has light turquoise colors which replace the colors of the flag. Even though the colors of the main contending party, the Partido Acción Nacional who was in power at the federal level during the study, are dark blue, it is plausible that the colors were related by the consumers. Notice, first, that consumers observed only one card. The political connotation of the cards is harder to observe in isolation than when they are contrasted. Second, even if the colors of the flag made consumers think of the
formerly hegemonic party, this is not necessarily contrary to the nationalist prime. One of the main tasks of the PRI was the nation building project (e.g. Chávez, 2009), and its 70 year hegemony makes it somewhat synonymous with a nation-wide project. In any case, whether or not the different cards make individuals think of the party or not is testable. However, if this is the message decision-makers are inferring, what’s important for the validity of our hypothesis is whether our results are an artifact of some other effect we’re ignoring through the different message. The natural candidate is that the NGO has private information about the recipient state’s party-in-power, or perhaps party affiliation. To test for different messages, we could ask respondents to tell us what comes to mind when looking at either card, and look for statistically significant patterns. We expect that the significant difference would be on mentions of nationalism.

A different type of alternative explanation would be that what is driving the difference in the prime’s effect is not a motivated cognition preference for a regionalist identity, but some other variable. This alternative variable would have to explain both the initial difference in bias and why both groups differ in their response to the prime. For example, poor states may give more to their region because there is spatial correlation in wealth. They respond more to primes because they are less educated so can be swayed more easily by non-informative messages. When they think they are giving to a recipient in Mexico as opposed to in their region, they give less because Mexico has more wealth on average. Note that this explanation is not consistent with all the predictions of the model. On the one hand, it would predict that individuals in rich states would be biased towards farther states. If they are less susceptible to non-informative messages, their pro-poor bias would not diminish with the prime. More generally, the types of hypotheses that come out of combining the motivated cognition and categorical social identity theories do not seem to be easily captured by an omitted variable.
1.3 Experimental Design

In this section we talk in more detail about our NGO partner, the treatments and research design, as well as the randomization.

1.3.1 Our Partner

The partner we are working with is Un Kilo de Ayuda, a Mexican NGO with the mission statement of eradicating child poverty in Mexico. In the period of study, they were serving 900 rural communities in eight states in Mexico, reaching 60 families on average per community. Their intervention includes periodic height, weight, cognitive development and hemoglobin measurements, beneficiary training on best feeding and hygiene practices as well as a package of 10 food items chosen to supplement the nutritional deficiencies in rural Mexico.

Un Kilo de Ayuda is the first NGO in Mexico to fundraise through cards sold at supermarket cashiers. They have followed this strategy for 25 years. These cards are placed in over 4,000 points of sale around the country. They are roughly the size of a postcard and have images or text inviting shoppers to donate to the NGO. The interested shopper hands the card to the cashier to scan. They cost 3 pesos (about a quarter) or 10 pesos (about 75 cents), depending on the point of sale. They normally don’t include anything of direct value to the consumer, although sometimes the card comes with a small gift (e.g. a bracelet, a packet of hand sanitizer). Since these cards typically have little value added, they are often left with the cashier. Through this method the NGO raises approximately 15% of their revenue.

1.3.2 The Design

Our experiment varied the design of the cards in six supermarket chains across 9 states. We implemented the design in all the supermarkets that sold 10 peso cards in the 9 states we chose. The 9 states included six of the states where the NGO was operating at the time (Guerrero, Sinaloa, Yucatán, Puebla, Estado de México and Veracruz) as well as the three
states with the largest cities in the country (Mexico City, Jalisco and Nuevo León). The total number of supermarkets in those states was 560, which raised 1.8 million pesos (approx. USD 120,000) in 2011. Figure 1.1 shows the location of the supermarkets across Mexico.

We placed cards across supermarkets that varied along two dimensions. The first is the state the card whose proceeds it will benefit: State of Mexico, Puebla, Sinaloa, Guerrero, Yucatán or Veracruz. Note that we limited the recipient states to states where the NGO was operating in order to avoid deception. The second dimension is whether the state card had the colors of the flag or not. An important part of our theoretical argument is that this prime is an uninformative message that highlights nationalism. We present an example of four cards in Figure 1.2, varying the recipient state and the nationalist prime.

In each of those supermarkets, UKA had two bar codes activated which were meant to be used exclusively for our cards for the duration of our experiment. Therefore, our unit of analysis is a half-supermarket. We tried to get as close as possible to putting at least 20 cards of each type per state, considering that we needed at least four type of cards: own state and one other state in a nationalist and non-nationalist version. In some states that was not possible. We also made sure that if state A had cards of state B, and both were recipient states, state B would have cards of state A. This meant that all states with only one other state card had State of Mexico as its ‘other state’. Figure 1.3 shows the allocation of cards to
stores. The small colored dots represent stores, the larger black dots represent lines, and the colored lines link the store to the recipient state. They are color coded by recipient, where yellow indicates Yucatán, orange State of Mexico, light blue Sinaloa, green Guerrero, red Veracruz and dark blue Puebla.

The frequency of the sales data we obtained depended on the chain where the cards were sold. We obtained daily data from Soriana and weekly data from Casa Ley. Obtaining sales data from Comercial Mexicana required one of the NGO staffers checking in daily on sales from the computers at the NGO, which was not possible on weekends and there were some days in which the staffer was unavailable. We ended up with a frequency that was somewhat irregular, but higher frequency than weekly. The other three supermarket chains only reported monthly data. Aside from sales, we know the chain, number of cashiers and address of the store.

1.3.3 Randomization

We took into account several considerations in order to randomize. First, there was a different number of types of cards per each states. Some states had 4 types of cards,
while others had 12. Second, each store was assigned two types of cards, so we needed to randomize within the supermarket and worry about possible spillovers. Third, we knew different supermarkets would give us different frequency data, and that there might be unobserved heterogeneity across supermarkets. Our approach was to first stratify by state and supermarket chain. Within these state-chain strata, we formed tuples or clusters of stores by matching the stores and then randomizing the treatments within the tuple (Bruhn et al, 2008). We matched by minimizing the sum of the Mahalanobis distance of all pairings within the cluster. We were able to set all tuples with the same supermarket chain except for one tuple. Within each store, we needed to decide which aisle received which of the two cards that were assigned. For each pair of aisles we flipped the virtual equivalent of a fair coin to decide which one would receive which card. We then made lists assigning cards to supermarket aisles which we sent to the staffers in charge of placing the cards. The staffers received these lists along with packet of cards that were marked per store. We conducted a partial audit of the cards, and found 2 supermarkets where the cards were misplaced. They were excluded from the sample.

We faced a couple of challenges in the implementation of the experiment. The placement
of cards was done by NGO staffers who lived in the states that were part of the experiment or nearby states. From Mexico City, we sent them boxes of cards to place. According reports from the NGO, the staffer from Sinaloa did an unreliable job in placing the cards. This can be seen from the omissions in the data, but internal audits by the NGO suggest that we cannot trust the appropriate placement in the stores that were reached. Indeed, the staffer was fired shortly thereafter for her performance in the months prior to and during the implementation. We therefore took out the observations from Sinaloa. Data from the Chedraui supermarket were also unreliable, as the reported amounts did not match the card prices.

We believe spillovers are not a big concern in this context for two reasons. The first is that different types of cards were placed at different cashier aisles. It is typically difficult to see the cards in another cashier, particularly the subtle differences between the designs. One may be worried about customers who visit supermarkets several times during the experiment, who would be exposed to different cards. Relatedly, the cards might get shuffled across aisles. In order to minimize these concerns, we focus on the sale of cards for a period of approximately two weeks. This was another reason to ignore card sales in Chedraui, as well as Almacenes Zaragoza and Tiendas Garcés, who reported monthly sales. Since the cards were placed at different points throughout October 2012, the sales measure would be conflated with sales of cards not in the experiment. Furthermore, Almacenes Zaragoza and Tiendas Garcés almost reported no sales in the period of study. For these reasons, we ignore Almacenes Zaragoza, Tiendas Garcés and Chedraui in the analysis. Given our clustered randomization, this does not affect the balance of the rest of the sample.

Outside of Sinaloa, attrition was generally not a large concern. 2 stores in Guerrero, 2 in the State of Mexico and 1 in Jalisco did not receive the cards. The exception is Nuevo León, who only received one of the two boxes of cards. They placed cards in 19 of the stores. The placement of cards was designed to maximize the number of tuples that were originally assigned. In all, the elimination of chains, observations from Sinaloa and outliers drops the tuples we have observations for from 140 to 82. Tuples for which we have the full set of
observations are 55. Table 1.1 shows the distribution of cards from donor state (rows) to recipient states (columns).

**Table 1.1: Distribution of recipient cards by donor**

<table>
<thead>
<tr>
<th>S of</th>
<th>Mex</th>
<th>Gue</th>
<th>Pue</th>
<th>Sin</th>
<th>Ver</th>
<th>Yuc</th>
<th>Tot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mexico City</td>
<td>25</td>
<td>21</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>22</td>
<td>134</td>
</tr>
<tr>
<td>State of Mexico</td>
<td>29</td>
<td>28</td>
<td>28</td>
<td>30</td>
<td>28</td>
<td>29</td>
<td>172</td>
</tr>
<tr>
<td>Guerrero</td>
<td>13</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>Jalisco</td>
<td>17</td>
<td>18</td>
<td>16</td>
<td>15</td>
<td>18</td>
<td>18</td>
<td>104</td>
</tr>
<tr>
<td>Nuevo León</td>
<td>13</td>
<td>15</td>
<td>13</td>
<td>17</td>
<td>14</td>
<td>12</td>
<td>84</td>
</tr>
<tr>
<td>Puebla</td>
<td>15</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>Veracruz</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>12</td>
<td>0</td>
<td>42</td>
</tr>
<tr>
<td>Yucatán</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>141</td>
<td>97</td>
<td>93</td>
<td>101</td>
<td>95</td>
<td>95</td>
<td>622</td>
</tr>
</tbody>
</table>

The table shows number of half-stores in states in the rows that had cards with recipients in the columns. S of Mex = State of Mexico, Gue = Guerrero, Pue = Puebla, Sin = Sinaloa, Ver = Veracruz, Yuc = Yucatán.

Table 1.2 below shows summary statistics. The experimental data we analyze includes 311 supermarkets. Store size varies between 3 and 35 cashiers, with a mean of 15. Notice that the sales in August, two months before the experiment, range from 0 to 3,504.7. Stores with no sales usually indicate that the cards had not been introduced yet. Notice further that very large supermarkets and very large sales are very infrequent.

**Table 1.2: Summary statistics of covariates**

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>S. D.</th>
<th>Min</th>
<th>Max</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latitude</td>
<td>20.46853</td>
<td>2.273281</td>
<td>16.77844</td>
<td>25.8257</td>
<td>311</td>
</tr>
<tr>
<td>Longitude</td>
<td>-99.39202</td>
<td>2.878353</td>
<td>-105.2403</td>
<td>-88.14884</td>
<td>311</td>
</tr>
<tr>
<td>Cashiers per store</td>
<td>14.83923</td>
<td>6.49587</td>
<td>3</td>
<td>35</td>
<td>311</td>
</tr>
<tr>
<td>Sales August 2012</td>
<td>481.1377</td>
<td>507.3783</td>
<td>0</td>
<td>3504.7</td>
<td>311</td>
</tr>
</tbody>
</table>

### 1.4 Empirical Specification and Balance

The empirical specification we will use throughout is the following:
\[ Sales_j = \alpha + \beta \text{IndepVar}_j + \gamma_1 \text{Donor}_j + \gamma_2 \text{Recipient}_j + \gamma_3 \text{tuple}_j + \delta X_j + \epsilon_j \]

where \( j \) is a half store. Our dependent variable is the sales for a period of about two weeks of sales, divided by cash registers and days. That is, how many cards were sold daily per cash register for a given type of card. Throughout our regressions, we control for sales in the month of August of 2012, the latitude and longitude of the store and the number of cashiers. We include donor fixed effects, recipient fixed effects and tuple fixed effects. As a conservative measure of our standard errors, we will cluster by state-chain. We will show the results with all our observations. One way to ensure balance is to restrict the sample to tuples for which we have complete tuples. The results yield qualitatively similar results and are not included.

For some specifications, it will be more natural to make comparisons across the states that have cards for all recipient states. We refer to these as 12-tuple states, since their tuples or clusters contain 12 types of cards. We will specify when the regressions only include them.

In order to test our theoretical predictions, summarized in Result 1, we need to define several variables.

**Regionalist versus nationalist:** We will be interested in making comparisons between regionalist and nationalist decision makers. In order to measure regionalism and nationalism, we use the results of a 2010 survey raised by the Mexican bank Banamex and the NGO Este País (Moreno, 2011). This is the only survey that asks about regional values in Mexico that is representative at the state level. One of the benefits of using a state-level variable is that, given that we clustered at the state level, our estimates of card purchases per treatment are balanced in expectation. The survey, called The National Survey of Values that Unite and Divide Mexicans (in Spanish, Encuesta Nacional de Valores sobre lo que nos Une y Divide a los Mexicanos or ENVUD), asks a series of questions regarding economic, political and social issues.

In order to classify states according to their nationalist pride, we used the state-level answer to the following question: On a scale of 1 to 10, where 1 means ‘not at all proud’
and 10 means ‘very proud’, how proud are you to be Mexican? (in Spanish: ¿En una escala del 1 al 10, donde 1 significa ‘nada orgulloso’ y 10 significa ‘muy orgulloso’, usted qué tan orgulloso está de ser mexicano?). We will use the continuous version of this measure as well as divide it into the half highest and half lowest values, which we will refer to as the ‘nationalist’ and ‘regionalist’, respectively.

One concern with this measure is that it is too aggregated. Unfortunately, there do not exist surveys of nationalism that are representative at a smaller political unit. A plausible alternative measure is municipal level GDP. According to the motivated cognition approach to social identity theory, individuals adopt their identity in order to maximize their self esteem. Therefore, when choosing whether to identify at a national or at a regional level, scholars have argued that they compare their relative well-being to that of the region and to that of the nation. The better off their region is, the more likely they are to identify with their region. Municipal level GDP gives us a measure of how well off an individuals’ region is with respect to the national level. Consistent with this, our survey measure has a -.48 correlation with state level GDP per capita. In the appendix, we show that using this alternative measure yields similar results.

**Region:** In order to study regional bias, we want to compare whether there is a preference to give to states that are closer to those who are further. Given that we’re dealing with states that are exposed to different numbers of recipient states, as well as states who are not exposed to their own state as a recipient, we focus on the half closest states as our main dependent variable. One possible objection to this approach is that a more natural measure of region would be desired. However, Mexico is known for having a particularly large amount of regional classifications (Van Young, 1992), which make it hard to know which one to use. Proximity is a natural candidate, and common to many of these different classifications.

There are different ways to measure distance. Ideally, we want a measure that takes into account not just linear distance between states, but how much it takes to get there. We also need to take into account the irregular shape of the states in order to figure out where to
calculate the start point and end point of the road connecting two states. We follow the intuitive approach of calculating the closest distance between state borders using Google Maps. Calculating the distance between two specific cities within the state - say the capital - is problematic since Mexico’s mountainous terrain may make it hard for certain specific cities to reach each other. This alternative measure yields qualitatively similar (although weaker) results.

**Poverty:** We will be interested in looking at whether individuals are biased towards benefiting recipients of poorer states. There are many ways to measure poverty of a state. One may consider GDP per capita. However, this measure is unsatisfactory, as it does not measure how many people in the state are in a state of poverty. The two other measures are amount of poor and poverty per capita. One may wish to use the total amount of poor: after all, when trying to help the most a decision maker may prefer to target the place where the most poor are. However, this ignores that the number of poor in a large state may be very small in comparison to the total population. When a decision-maker thinks about the ‘poverty problem’ of a state, we assume the focus is on poverty per capita. The intuition is that poverty is a local public good problem, and donating to a charity contributes to that public good. Along the same lines, in states that the proportion of poor is very small, the poor may be more disperse among the population and harder to reach. We therefore focus on poverty per capita, which we take from the official statistics produced by CONEVAL in 2007.

Notice that most states are exposed to cards benefiting their own state and one other state, so the confound between wealth and regional bias is strongest. For this reason, when looking at poverty we will restrict our analysis to 12-tuple states. We will look at both the continuous measure of the recipient state’s poverty per capita and at the poorest recipient state among 12-tuple states, Guerrero. The latter is our preferred measure, as Guerrero is generally considered to be one of the poorest states in Mexico.

Table 1.3 shows the summary statistics of the dependent and independent variables. First notice that our dependent variable has a mean of .21 cards sold daily per aisle. However,
the maximum value is 6.72 cards daily per aisle. Because we had very few observations with such high values and we had multiple types of cards, we are concerned that these high values may be driving our results by unbalancing some tuples. We therefore ran a specification of the dependent variable as-is and a version of the dependent variable where we winsorized the 9 highest values in the sample, which ranged from 1.73 to 6.72. We present both results when they are qualitatively different.

Notice that about half of the recipients (43.7%) were nationalist types according to the ENVUD survey. About half of the recipient states were half closest to decision makers. From among the 227 supermarkets that were in 12-tuple states, 16% of them had a card with the poorest recipient state, Guerrero.

It is worth pointing out that the values of the survey answers of nationalist pride for our donor states are in general high. Note that this is expected from a question that suffers from social desirability bias. Second, and consistently, 60% of individual responses in the nation-wide survey are of value 10, 14% are 9 and 13% are 8. So this state-level variation captures a fair amount of the variation in responses. The summary of the poverty per capita variable is presented for the six recipient states. Notice that there is a large variation, from Guerrero with 67.4 poor per capita to Sinaloa with 36.5.

**Table 1.3: Summary statistics of dependent and independent variables**

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>S. D.</th>
<th>Min</th>
<th>Max</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily cards per aisle</td>
<td>.2159879</td>
<td>.456454</td>
<td>0</td>
<td>6.724902</td>
<td>311</td>
</tr>
<tr>
<td>Nationalist</td>
<td>.437299</td>
<td>.4964522</td>
<td>0</td>
<td>1</td>
<td>311</td>
</tr>
<tr>
<td>Municipal GDP pc (thousand pesos)</td>
<td>20.74326</td>
<td>8.085779</td>
<td>4.88753</td>
<td>47.24912</td>
<td>311</td>
</tr>
<tr>
<td>Half closest</td>
<td>.5064309</td>
<td>.500361</td>
<td>0</td>
<td>1</td>
<td>311</td>
</tr>
<tr>
<td>Poorest recipient</td>
<td>.1659919</td>
<td>.3724505</td>
<td>0</td>
<td>1</td>
<td>227</td>
</tr>
<tr>
<td>Mexican pride by state</td>
<td>9.145</td>
<td>.1855494</td>
<td>8.93</td>
<td>9.46</td>
<td>8</td>
</tr>
<tr>
<td>Pov pc by recipient state</td>
<td>52.33333</td>
<td>11.80723</td>
<td>36.5</td>
<td>67.4</td>
<td>6</td>
</tr>
</tbody>
</table>

Our independent variables of interest will be the main effect and interaction of our treatments. We are interested in looking at whether nationalism has an effect, whether the recipient state is part of the region and the interaction of the two. Table 1.4 presents the
balance table of these four bins for covariates of location, store size and past sales. The first four rows of results present the results for all states in the sample. We are interested in looking at these bins for regionalist and nationalist decision makers, so the next sets of four rows present the balance for each of these sub-groups. The covariates are in general balanced. Further, because we cluster randomized by state-chain, we know can exclude supermarkets chains and states while remaining an expected balance. The standard errors in the table are clustered by state-chain. It is worth noting that regionalist states in general sell less than nationalist states. This is expected, as regionalist states are on average poorer.

As we discuss in Section 2.1, the results we will present are hard to reconcile with a theory in which all results are driven by an income effect. What is important is that, within each type of state we find balance on the different treatments.

### 1.5 Results

In this section we present our main results. These sections will provide evidence for each of the predictions of Result 1 in the Theoretical Framework. We will proceed through each of the predictions in turn.

#### 1.5.1 Nationalists increase contribution more with prime

In this subsection we provide evidence for Result 1.1, that nationalist types increase contributions more with the flag colors.

Table 1.5 shows our first result. The first thing to notice is that the nationalist prime induced by the flag colors on the card has an insignificant effect on overall card sales, as shown in column 1. Columns 2 and 3 show that this effect is insignificantly positive for nationalists, and insignificantly negative for regionalists. However, this difference of differences is statistically significant in column 5, where all observations are pooled and the dependent variable is winsorized. Although significance is found only on the winsorized version, the results are close to significance without this manipulation.
Table 1.4: Balance Table

<table>
<thead>
<tr>
<th></th>
<th>Latitude</th>
<th>Longitude</th>
<th>Cashiers</th>
<th>Past Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closest, with prime</td>
<td>20.38</td>
<td>-99.40</td>
<td>15.50</td>
<td>509.8</td>
</tr>
<tr>
<td></td>
<td>(0.787)</td>
<td>(0.809)</td>
<td>(0.638)</td>
<td>(104.2)</td>
</tr>
<tr>
<td>Closest, no prime</td>
<td>20.28</td>
<td>-99.23</td>
<td>14.95</td>
<td>483.6</td>
</tr>
<tr>
<td></td>
<td>(0.737)</td>
<td>(0.850)</td>
<td>(0.946)</td>
<td>(90.43)</td>
</tr>
<tr>
<td>Furthest, with prime</td>
<td>20.52</td>
<td>-99.59</td>
<td>14.25</td>
<td>483.6</td>
</tr>
<tr>
<td></td>
<td>(0.862)</td>
<td>(0.758)</td>
<td>(0.844)</td>
<td>(90.43)</td>
</tr>
<tr>
<td>Furthest, no prime</td>
<td>20.45</td>
<td>-99.35</td>
<td>14.64</td>
<td>465.2</td>
</tr>
<tr>
<td></td>
<td>(0.797)</td>
<td>(0.744)</td>
<td>(0.837)</td>
<td>(88.88)</td>
</tr>
<tr>
<td><strong>Nationalist states</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closest, with prime</td>
<td>21.09</td>
<td>-100.8</td>
<td>15.90</td>
<td>605.7</td>
</tr>
<tr>
<td></td>
<td>(1.316)</td>
<td>(0.947)</td>
<td>(0.575)</td>
<td>(180.3)</td>
</tr>
<tr>
<td>Closest, no prime</td>
<td>20.91</td>
<td>-100.7</td>
<td>14.58</td>
<td>569.9</td>
</tr>
<tr>
<td></td>
<td>(1.287)</td>
<td>(0.929)</td>
<td>(1.155)</td>
<td>(157.9)</td>
</tr>
<tr>
<td>Furthest, with prime</td>
<td>21.21</td>
<td>-100.9</td>
<td>14.43</td>
<td>569.9</td>
</tr>
<tr>
<td></td>
<td>(1.374)</td>
<td>(0.918)</td>
<td>(1.229)</td>
<td>(157.9)</td>
</tr>
<tr>
<td>Furthest, no prime</td>
<td>21.12</td>
<td>-100.7</td>
<td>14.76</td>
<td>547.4</td>
</tr>
<tr>
<td></td>
<td>(1.350)</td>
<td>(0.868)</td>
<td>(1.225)</td>
<td>(156.3)</td>
</tr>
<tr>
<td><strong>Regionalist states</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closest, with prime</td>
<td>19.47</td>
<td>-97.58</td>
<td>14.99</td>
<td>387.8</td>
</tr>
<tr>
<td></td>
<td>(0.165)</td>
<td>(1.079)</td>
<td>(1.235)</td>
<td>(62.24)</td>
</tr>
<tr>
<td>Closest, no prime</td>
<td>19.52</td>
<td>-97.42</td>
<td>15.39</td>
<td>380.3</td>
</tr>
<tr>
<td></td>
<td>(0.187)</td>
<td>(1.212)</td>
<td>(1.641)</td>
<td>(73.87)</td>
</tr>
<tr>
<td>Furthest, with prime</td>
<td>19.53</td>
<td>-97.79</td>
<td>13.98</td>
<td>380.3</td>
</tr>
<tr>
<td></td>
<td>(0.139)</td>
<td>(1.036)</td>
<td>(1.148)</td>
<td>(73.87)</td>
</tr>
<tr>
<td>Furthest, no prime</td>
<td>19.59</td>
<td>-97.67</td>
<td>14.49</td>
<td>361.3</td>
</tr>
<tr>
<td></td>
<td>(0.154)</td>
<td>(1.136)</td>
<td>(1.204)</td>
<td>(36.73)</td>
</tr>
</tbody>
</table>

Standard errors in parentheses
Table 1.5: Nationalist Prime and Pride

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Nat</th>
<th>Reg</th>
<th>All</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flag colors</td>
<td>0.0466</td>
<td>0.104</td>
<td>-0.0129</td>
<td>-0.0381</td>
<td>-0.0764</td>
</tr>
<tr>
<td></td>
<td>(0.0571)</td>
<td>(0.103)</td>
<td>(0.0358)</td>
<td>(0.0699)</td>
<td>(0.0489)</td>
</tr>
<tr>
<td>Nationalism</td>
<td>-0.202</td>
<td>-0.337</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.691)</td>
<td>(0.534)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interaction</td>
<td>0.310</td>
<td>0.311*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.211)</td>
<td>(0.156)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>2.080</td>
<td>3.340</td>
<td>1.899</td>
<td>2.083</td>
<td>1.806</td>
</tr>
<tr>
<td></td>
<td>(1.598)</td>
<td>(1.974)</td>
<td>(2.770)</td>
<td>(1.921)</td>
<td>(1.771)</td>
</tr>
<tr>
<td>Observations</td>
<td>622</td>
<td>272</td>
<td>350</td>
<td>622</td>
<td>622</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.287</td>
<td>0.292</td>
<td>0.293</td>
<td>0.287</td>
<td>0.339</td>
</tr>
<tr>
<td>Winsorized DV</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Dependent variable is daily card sales per aisle
Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

1.5.2 Only regionalists contribute relatively more to close states

In this subsection we provide evidence for Result 1.2, that regionalists are biased in their contribution decisions towards states that are in their region, and that nationalists do not exhibit this bias.

Table 1.6 presents our results. Columns 1 and 2 show the aggregate results. There is a bias in favor of contributing to closer states, of .06 extra daily cards per aisle. When split up by nationalism and regionalism, it can be seen that the bias is driven by the regionalist states, who have at least twice the bias that the nationalist states do. The bias is only significant in the non-winsorized results for the aggregate of states.

1.5.3 Regionalist’s bias diminishes with prime

In this subsection we provide evidence for Result 1.3, that the bias we found regionalists to have diminishes with the nationalist prime, and does not affect nationalist’s decision over whether to contribute to a state depending on its proximity.

Table 1.7 presents our results. Notice that in column 4, the regionalist states display
Table 1.6: Regional bias for Regionalists and Nationalists

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>All</th>
<th>Nat</th>
<th>Nat</th>
<th>Reg</th>
<th>Reg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closest</td>
<td>0.0524</td>
<td>0.0660**</td>
<td>0.00878</td>
<td>0.0435</td>
<td>0.0893*</td>
<td>0.0957**</td>
</tr>
<tr>
<td></td>
<td>(0.0377)</td>
<td>(0.0271)</td>
<td>(0.0287)</td>
<td>(0.0243)</td>
<td>(0.0458)</td>
<td>(0.0385)</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.517</td>
<td>0.586</td>
<td>3.564*</td>
<td>1.511</td>
<td>-8.594</td>
<td>-4.848</td>
</tr>
<tr>
<td></td>
<td>(3.402)</td>
<td>(1.892)</td>
<td>(1.760)</td>
<td>(0.949)</td>
<td>(5.238)</td>
<td>(4.587)</td>
</tr>
<tr>
<td>Obs</td>
<td>622</td>
<td>622</td>
<td>272</td>
<td>272</td>
<td>350</td>
<td>350</td>
</tr>
<tr>
<td>Adj $R^2$</td>
<td>0.335</td>
<td>0.363</td>
<td>0.281</td>
<td>0.360</td>
<td>0.337</td>
<td>0.335</td>
</tr>
<tr>
<td>Winsor</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A bias towards close states: they buy .12 extra cards daily per aisle for cards with closer recipient states. Although it is insignificant, the coefficient on the interaction suggests that this bias diminishes by half with the prime. The story is very different for the nationalist states, shown in column 3. They do not display a bias based on proximity, and if anything they give more to close-by states with the prime. That the pattern of contributions for both type of states is different is shown in the positive and significant triple interaction of column 1.

1.5.4 All contribute more towards the poor states, especially nationalists

In this subsection we provide evidence for Result 1.4 and 1.5, by showing that there is a bias towards contributing to the poorest recipient states, and that this bias is highest for the nationalists.

As discussed previously, for this and the next part of the analysis we will focus on 12-tuple states because they have six recipient states. Other states have at most three recipient states, and we would like to be able to disentangle poverty and regional bias. By narrowing the sample, this reduces the state-chains per group to cluster the standard errors. We therefore will report when robust standard errors changes significance in this and the next.

Table 1.8 presents our results. The independent variable in the first three columns...
Table 1.7: Triple interaction with nationalist pride

<table>
<thead>
<tr>
<th>Triple interaction</th>
<th>All</th>
<th>Nationalist</th>
<th>Regionalist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closest</td>
<td>0.0509</td>
<td>0.0609</td>
<td>-0.0177</td>
</tr>
<tr>
<td></td>
<td>(0.0568)</td>
<td>(0.0363)</td>
<td>(0.0278)</td>
</tr>
<tr>
<td>Flag colors</td>
<td>0.00208</td>
<td>0.0466</td>
<td>0.0334</td>
</tr>
<tr>
<td></td>
<td>(0.0359)</td>
<td>(0.0469)</td>
<td>(0.0432)</td>
</tr>
<tr>
<td>Nationalism</td>
<td>-0.959*</td>
<td>(0.463)</td>
<td></td>
</tr>
<tr>
<td>Closest *</td>
<td>-0.0594</td>
<td>-0.0165</td>
<td>0.0565</td>
</tr>
<tr>
<td></td>
<td>(0.0572)</td>
<td>(0.0493)</td>
<td>(0.0360)</td>
</tr>
<tr>
<td>Flag colors</td>
<td>0.00301</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nationalism</td>
<td>0.0752</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nationalism *</td>
<td>0.109</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flag colors</td>
<td>(0.0911)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Triple</td>
<td>0.132**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interaction</td>
<td>(0.0539)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>622</td>
<td>622</td>
<td>272</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.337</td>
<td>0.334</td>
<td>0.355</td>
</tr>
<tr>
<td>Winsorized</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
are whether the card benefited Guerrero, the poorest state. In the last three columns, the independent variable is the poverty per capita of the recipient states in the cards. We see a similar pattern in both cases. There is an overall bias towards benefiting the poorest states, and nationalist states have a stronger bias. This pattern can also be found in the full sample, that is including non 12-tuple states. It should be noted that the coefficients from the fourth and fifth column become shy of significance with robust standard errors (p-value of .102 and .156, respectively).

<table>
<thead>
<tr>
<th>Table 1.8: Poverty Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td><strong>All Nat Reg All Nat Reg</strong></td>
</tr>
<tr>
<td>Poverty pc 0.033 0.054 -0.013 0.002* 0.003** 0.002</td>
</tr>
<tr>
<td>(0.051) (0.013) (0.063) (0.001) (0.00007) (0.001)</td>
</tr>
<tr>
<td>Observations 494 172 322 492 170 322</td>
</tr>
<tr>
<td>Adjusted R² 0.473 0.333 0.359 0.472 0.333</td>
</tr>
<tr>
<td>DV Poorest Poorest Cont Cont Cont</td>
</tr>
<tr>
<td>Winsorized No No No No No No</td>
</tr>
</tbody>
</table>

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

1.5.5 Pro-poor bias diminishes with prime, especially for Nationalists

In this subsection we provide evidence for Result 1.6, by showing that the bias towards the poor diminishes with a nationalist prime, and that diminution is higher for Nationalists.

Table 1.9 shows our results. As in the last subsection, the independent variable in the first three columns are whether the card benefited Guerrero, the poorest state. In the last three columns, the independent variable is the poverty per capita of the recipient states in the cards. The pattern of results now depends on the variable we use. With the poorest recipients as the independent variable, the results are in the expected direction. On aggregate, individuals contribute more when the recipient state is Guerrero and this bias diminishes with the prime. The nationalist’s pro-Guerrero bias is 4 times larger than the regionalist’s, and this bias becomes negative with the prime. Although the effect is not significant with the clustered standard errors, all the coefficients in the second column are
significant at at least the 0.05 significance level with robust standard errors. In contrast, the continuous measure of poverty per capita shows that it is the regionalists who have a pro-poor bias, and whose bias disappears with the nationalist prime. In either case, all groups display a bias towards the poor which diminishes or is reversed with the nationalist prime.

As we’ve argued, our preferred measure of regionalism is that of the poorest state, as Guerrero along with Oaxaca and Chiapas (not in the sample) are commonly perceived to be the poorest region in Mexico. Indeed, these three states rank among the four poorest in terms of both poverty per capita and GDP per capita.

Table 1.9: Poverty Bias and Prime

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Nat</th>
<th>Reg</th>
<th>All</th>
<th>Nat</th>
<th>Reg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poverty pc</td>
<td>0.13*</td>
<td>0.29</td>
<td>0.052</td>
<td>0.0039</td>
<td>0.0018</td>
<td>0.0050*</td>
</tr>
<tr>
<td></td>
<td>(0.062)</td>
<td>(0.11)</td>
<td>(0.051)</td>
<td>(0.0019)</td>
<td>(0.0056)</td>
<td>(0.0022)</td>
</tr>
<tr>
<td>Flag colors</td>
<td>0.075</td>
<td>0.17</td>
<td>0.0045</td>
<td>0.241**</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>(0.074)</td>
<td>(0.15)</td>
<td>(0.057)</td>
<td>(0.094)</td>
<td>(0.087)</td>
<td>(0.15)</td>
</tr>
<tr>
<td>Poverty pc *</td>
<td>-0.19*</td>
<td>-0.45</td>
<td>-0.12</td>
<td>-0.0036</td>
<td>-0.0019</td>
<td>-0.0053*</td>
</tr>
<tr>
<td>Flag colors</td>
<td>(0.089)</td>
<td>(0.22)</td>
<td>(0.11)</td>
<td>(0.0021)</td>
<td>(0.0056)</td>
<td>(0.0023)</td>
</tr>
<tr>
<td>Observations</td>
<td>494</td>
<td>172</td>
<td>322</td>
<td>492</td>
<td>170</td>
<td>322</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.365</td>
<td>0.489</td>
<td>0.333</td>
<td>0.300</td>
<td>0.338</td>
<td>0.302</td>
</tr>
<tr>
<td>DV</td>
<td>Poorest</td>
<td>Poorest</td>
<td>Poorest</td>
<td>Cont</td>
<td>Cont</td>
<td>Cont</td>
</tr>
<tr>
<td>Winsorized</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

1.6 Conclusions

In this paper we presented a model of identity behavior which combines insights from the motivated cognition and categorization explanations of identity adoption. We showed empirically that if an individual is led to adopt an identity that maximizes his self-esteem, he will behave more charitably to others who share his identity than if he is led to adopt an identity that gives him lower self esteem.

This paints a nuanced picture of how a persuader can affect others’ behavior. Although
the persuader may be able to shift the identity through which an individual makes a decision, without an informative message the persuader cannot manipulate how the individual ranks that identity. Therefore, how behavior changes when the identity of a target individual is shifted also depends on the identity ranking. In our case, we provided evidence that increasing contributions with a nationalist prime depended on whether individuals were regionalist or nationalist.

From a policymaking perspective, the implication of our results are that contributions towards a public good can be increased by better understanding the preference over identities individuals have and a targeted priming of identities. For example, we could have increased revenue in our sample by placing cards with flag colors in nationalist states and without flag colors in regionalist states.

Notice further that manipulating an individual’s bias may lead to an efficiency-equity tradeoff. Even though nationalists weakly increased their overall contribution with the prime, they also diminished their contributions towards the poorest state. More generally, if the prime makes individuals less sensitive to the recipients’ characteristics, they might not target their contributions in a way that is socially optimal.

This linkage of the categorical and motivated cognition approach to social identity theory leads to other lines of inquiry. For example, the motivated cognition channel of identity is presumably modified by informative messages regarding the social status of a given identity. How does a shift in identity due to motivated cognition affect the manipulability of the categories individuals use? Can manipulating categories lead individuals to increase their preference of identity, perhaps by increasing exposure to other groups? We hope to address these questions in future work.
Chapter 2

Social expectations in pro-social influence

2.1 Introduction

Why are public goods sometimes provided in the absence of centralized enforcement agents? There is strong evidence in the field (e.g. Ostrom, 1990; Fisman and Miguel, 2007) and in the lab (e.g. Fehr and Gächter, 2000) that individuals may sustain positive contributions towards a public good, escaping the logic of collective action (Olson, 1965). Culturalist explanations of economic and political development often rely on the variation of non-enforced public good provision to explain the variation of macro-level outcomes across countries (e.g. Weber, 1905; Banfield, 1958, Putnam et al, 1993; Fukuyama, 1995). One avenue through which these contributions may be sustained could be through self enforcing social information. Individuals respond positively to social information (Cialdini and Trost, 1998; Bruno and Meier, 2004, Richard and Randal, 2008, Alpízar, Carlsson and Johansson-Stenman, 2008, Shang and Croson, 2009; Croson and Shang, 2009; Kessler, 2013; Allcot, 2011), so public contributions towards a public good may lead to more contributions. In order to understand the sustainability of this avenue, we need to understand the behavior and motivations of

---

1Co-authored with Michael Hiscox
individuals whose contributions will be publicized to others who will then make their own contributions. We will refer to the former individuals as being in the position of ‘first movers’, and the latter as being in the position of ‘second movers’. We’ll say an individual leads when he contributes differently when in a first mover position. In this paper, we are interested in whether and why leaders lead in this narrow sense.

Our narrow definition of leadership applies to many types of political and economic processes. In the most dramatic, martyrs who die for their cause or leaders who head a protest or a military strike are making a hefty public contribution to a public good. In the most mundane, people who make a point of recycling, voting or donating publicly are also leading. Because of the nature of a public good, often individuals must be influenced into taking actions that are not in their self interest. When first movers lead, they potentially take on the task of influencing. A standard economic explanation of why they do so is that they are motivated by self image considerations (Bénabou and Tirole, 2006; Ariely, Bracha and Meier, 2009). These self image explanations allow for non-material selfish motivations to outweigh the cost of engaging in collective action. An individual who is sporting an ‘I Voted’ sticker on the day of the elections is presumably doing so in order to get recognition from others. According to this view, influencing others is a byproduct of the social image motivator. A different explanation is that individuals influence others in order to gain the self interested benefits that come from more people providing a public good (e.g. Hermalin, 1998). The role of the ‘I Voted’ sticker in this explanation is to get more people to vote for the candidate whose policies will benefit the bearer of the sticker. In this paper, we’ll explore whether influencing others to contribute to a public good without direct material benefits is a motivator unto itself. That is, do people wear the ‘I Voted’ sticker to get others to vote because they think it is a good thing independently of their own self-interest? This would imply that individuals lead in order to increase the followers’ contributions from what they would had given had they seen a different contribution. We call this motivation ‘altruistic influence’ or ‘altruistic leadership’.

The experimental literature on decision making has become skeptical of explanations of
behavior that rely on altruism. One reason for this is that individuals display reluctance in their altruism. The seminal experiment was done by Dana, Cain and Dawes, 2006. They have subjects make a decision of how to divide 10 dollars with another player in an anonymous dictator game. They then offer the dictators to switch from the dictator game decision to instead receive 9 dollars and exit the game without the recipient learning about the game. They find that dictators choose the ‘silent exit’ option. This displays reluctance in that they are taking back the altruistic action they engaged in. Notice that this is strange behavior from an efficiency perspective: in principle they could have made everyone at least as well off by giving one dollar and not exiting. To shed light on people’s motivations, they show that people don’t exit silently when the recipient never would have known where the money came from. Reluctance is therefore explained as individuals engaging in altruistic actions in order to avoid the cost of disappointing others. Other work has also found evidence for reluctant altruism (Dana, Weber and Xiang, 2007, DellaVigna, List and Malmendier, 2012, but see Grossman, 2010). Is leadership reluctant? That is, if individuals do indeed lead for non-material or social image reasons, would they take their leadership back when they can avoid social expectations? We call this ‘reluctant leadership’, and explore it in this paper.

We provide experimental evidence for leadership, and we further show that it is reluctant. Our design distinguishes between altruistic leadership and reluctant leadership by having first movers make a series of decisions and elicit their expectations about the impact of their decisions on second movers. First movers divide ten dollars between themselves and a charity in three different scenarios. They know that after they make their decisions, a lottery will select one of the decisions. Note that this strategy method allows us to keep the first mover’s decisions on the same margin. In the first two scenarios, the first mover divides the ten dollars prior to a second mover who will know what the first mover did and will make his own division with his own ten dollars. The only difference between the scenarios is whether the second mover knows what the first mover did before making his decision or after. We call these the ‘Can-Consider’ and ‘Cannot-Consider’ scenarios, since only in the
former can the second mover consider what the first mover did when making a decision. We randomized whether subjects received the Can-Consider scenario as their first scenario.

Does giving more in the Can-Consider scenario than in the Cannot-Consider scenario provide evidence for altruistic leadership? Not necessarily. Note that there might be higher social expectations when put in a position where actions can influence others, and a higher contribution might be driven by the desire to avoid the cost of not meeting these expectations. This would be the driver of reluctant leadership. Note that this line of reasoning does not rely on social image concerns, since the contributions are anonymous. Rather, it assumes individuals care about what others expect them to do. This can be justified through psychological game theory (Geanakoplos, Pearce and Stacchetti, 1989; Broberg, Ellingsen and Johannesson, 2007; Andreoni, Rao and Trachtman, 2011; Ockenfels and Werner, 2011; Lazear, Malmendier, Weber, 2012; Taubinskky, 2013), in which higher order beliefs may be arguments in an individual’s utility function, or type-dependent preferences (Levine, 1995, chapter 3 of the dissertation). In order to distinguish reluctant leadership from altruistic leadership, we turn to the third of three decisions leaders are asked to make with their ten dollars. In this third decision, they cannot make a new unconstrained division of ten dollars. Rather, they are asked to choose between replicating the Can-Consider scenario or the Cannot-Consider scenario along with the decisions they made in those scenarios. For those who gave more under the Can-Consider scenario, this decision distinguishes altruistic and reluctant leadership.

Consider first the decision the altruistic leader faces. The reason he leads is to get others to contribute more to the public good. In the Can-Consider scenario, the first mover faces a multiplier on his contribution: what the charity receives is not only determined by what he gives, but also the impact of what he gives on what the follower gives. If second movers respond positively to the amount a leader contributes, an altruistic leader can ‘purchase’ a higher contribution to charity at a lower ‘price’. Under certain conditions on the first mover’s expectations of second movers’ response, by giving more in the Can-Consider scenario the first mover is revealing a preference for the allocation outcome in that scenario. These
conditions are a positive impact condition, in which the first mover thinks that the second mover will respond positively to the first mover giving more, and a marginal comparability condition, in which the first mover expects that the decision of the first and second movers are on a comparable margin. By replicating the Can-Consider (Cannot-Consider) scenario under these conditions, he is making it more (less) likely that a higher utility outcome obtains. The decision the altruistic leader will make is therefore unambiguous.

Now consider the decision the reluctant leader faces, and let a positive or zero multiplier condition as well as a marginal comparability condition hold. The leader will give more in the Can-Consider scenario only to avoid the higher social expectation associated with that scenario. By contributing more, the leader avoids the cost of being farther from the social expectation. When offered the option of replicating the Can-Consider and the Cannot-Consider scenario, he will choose to replicate the Cannot-Consider scenario. This way, the leader can avoid the social expectation cost, and does not have to contribute more to avoid it. Thus, if the conditions on expectations hold our three decisions are a sufficient statistic to distinguish between altruistic and reluctant leaders.

We find that 23% of our sample give more when they are first movers than when they are second, and on average being a first mover increases contributions by 20%. We further provide evidence that 80% of the leaders are reluctant in the sense that they choose to replicate the Cannot-Consider scenario. When we take reported beliefs into account to test for the positive multiplier condition, we are able to identify 40% of the leaders as reluctant, and only 10% as altruistic.

The rest of this paper is organized as follows. In the next section, we frame our contribution in the context of related literature. In section III we present the formalization of the conceptual framework we outlined in this introduction. In section IV, we outline the design of the experiment. Section IV presents our results. The fifth section concludes.
2.2 Related Literature

Studying altruistic and reluctant leadership in a non-experimental setting is hard to do because social image concerns, material self-interest and influence are bundled up in everyday scenarios. The experimental design offers a setting in which individuals can choose how much to contribute publicly but anonymously to a public good that does not benefit subjects directly (a charity). Settings that do not fulfill these conditions introduce the confounds we’ve discussed. Surprisingly, most of the experimental work in the literature that have studied altruistic leadership have failed to take these confounds into account. Many of them study a public goods game where the individual gets material benefits from what other players contribute (Meidinger and Villeval, 2002, Moxnes and van der Heijden, 2003; Potters, Sefton and Vesterlund, 2005; Güth, Levati, Sutter and van der Heijden, 2007; Komai and Grossman, 2007; Levati, Sutter and van der Heijden, 2007; Levy, Padgitt, Peart, Glöckner, Irlenbusch, Kube, Nicklisch, and Normann, 2009; Gachter, Nosenzo, Renner and Sefton, 2009 and 2010; Koji, Kobayashi and Suehiro, 2010; Rivas and Sutter 2009; Houser and Xiao, 2011; Komai, Grossman and Deters, 2011; Bracha, Menietti and Vesterlund, 2011; d’Adda, 2011; Arbak and Villeval, 2013). This setting is problematic since a first mover may be leading in order to increase his own benefit from the public good. Nevertheless, some authors use the results of this experiment to infer altruistic leadership motivations (e.g. Arbak and Villeval, 2013). Other experimental studies make the identity of the subject public or partially public (e.g. Jack and Recalde, 2013). An example where the identity is partially identifiable is d’Adda, 2011, who divides groups in villages in Colombia into subgroups of six to eight based on publicly recognizable status tiers. Second movers see first movers’ contributions as well as the status tier the leader was in. Concern for reputation (perhaps at the status tier level) confounds altruistic motivations.

A second concern with the literature on leadership is that several authors do not elicit expectations of the first mover’s impact on the second mover. Although whether a first mover’s action influences others is important, whether a first mover is intending to influence others is necessary for understanding leadership motivations. For example, consider a first
mover who thought second movers responded strongly negatively to what he gave. Then by giving less in a position to influence, the first mover is arguably acting as an altruistic leader. We should therefore be skeptical of conclusions made over intentions without information about the first mover’s expectations. An article that does take expectations into account is Gachter, Nosenzo, Renner and Sefton, 2010, who show that first movers who lead the most are those that have the highest expectation of their influence on followers. A related concern is the use of between subjects comparison in different scenarios. If those who give more are part of the minority who think contributions have a negative effect, we might still see a higher contribution of first movers on average and a positive expected response function on average, but no individual who gave more because he expected that would make others give more.

The Karlan and McConnell, 2012 article is closest to ours, so it is worth pointing out the differences a bit more carefully. In their design, they give individuals three sets of five dollars. With the first set they give privately. With the second they give publicly, since the experimenters write the names of all participants on a board with their corresponding amount. There are two conditions, in the first their contribution is announced before the third decision, and in the second their contribution is announced after. These conditions are closest to our Can Consider and Cannot Consider scenarios. Of note, they do not find a significant difference between contributions in these two scenarios. There are three main differences between their design and ours. The first is that the three decisions are made with different sets of 5 dollars. That means that each decision the subjects are making are on different margins, which makes it difficult to make any utility comparisons. We avoid this by using the strategy method with individuals’ three choices. The second is that social image motivations are present and salient. The social image motivation may be what’s most important or salient in both conditions, which may explain why they act similarly. The third difference is that everyone’s contribution is made public. This creates a strong free rider problem: the impact any first mover’s contribution may have on second movers is dampened by the information of what others give. A different complication that arises is
that individuals’ relative contribution becomes a relevant consideration, which is present in both the Can Consider and the Cannot Consider treatments. The anonymity in our design and the first mover’s lack of information regarding what others gave avoids these issues.

Past theoretical work on leadership as defined here has mostly focused on the leader’s private information on the productivity of the public good (Hermalin, 1998; Andreoni, 2005; Komai et al, 2007; Kobayashi and Suehiro, 2008; Majumdar and Mukand, 2010; Hatsumi, 2013). There are a few behavioral explanations, such as players disliking effort differentials (Huck and Rey-Biel, 2006), being impure altruists (Romano and Yildirim, 2001, or have a preference for fairness (Duffy and Muñoz García, 2010; Abe, Kobayashi and Suehiro, 2012). This paper explores the theoretical and empirical role social expectations have in a theory of leadership.

More generally, the narrowly defined leadership we are exploring is a fundamental component of political processes, as it impacts formal and informal positions of authority. While the former must balance their mechanisms of sanctions and rewards with obtaining voluntary compliance, the latter must rely most heavily on influencing others. When studying leadership broadly understood, game theory inspired political science has mostly looked at the formal positions of authority. Much of the focus has been on the impact different institutional settings a leader can be placed in have on the political outcomes (Shepsle, 1995) or the type of leader that is selected given different institutions (Coate and Morris, 1985; Rogoff, 1990; Besley, 2005; Acemoglu and Egorov, 2010) - although Majumdar and Mukand, 2010 and Acemoglu and Jackson, 2012 are exceptions. Recent empirical work, however, has shown the importance of individual leaders in determining outcomes (Bertrand and Schoar, 2003; Jones and Olken, 2005 and 2009; Mukunda, 2012). This paper is an attempt to contribute to a systematic exploration of a fundamental and underexplored aspect of political processes.
2.3 Conceptual Framework

Consider an environment with two players, a first mover $i$ and a second mover $j$. Both players must choose an amount out of a dollar to keep for themselves and the amount to give to a public good. The first mover makes his choice before the second mover. We will consider a scenario where the second mover sees what the leader gave before making a decision (the Can Consider scenario), and another one where the second mover does not (the Cannot Consider scenario). We will refer to the amount the first mover and the second movers give in the Can Consider scenario as $m_i$ and $m_j$, respectively. The first mover’s decision in the Can Consider scenario is given by:

$$
\max_{m_i \in [0,1]} EU(m_i, m_j, \text{Can}) =
$$

$$
u(m_i) + (1 - m_i) + \gamma(1 - E_i(m_j|m_i)) - ac(E_{soc}(m_i|\text{Can}) - m_i)
$$

There are three additively separable terms in the utility function. The first is the utility $u$ the first mover receives from the amount $m_i$ he keeps. This term is concave ($u' > 0, u'' < 0$). The second term is the amount of money that goes to the public good. This term is divided into the amount that the first mover contributes directly ($1 - m_i$) and the amount that the second mover contributes conditional on the first mover’s contribution ($\gamma(1 - E_i(m_j|m_i))$).

There are a couple of notes to point out about this component. The first is that it captures the intuition that interpreting whether a first mover is leading when a second mover will see what he gave depends on what the first mover’s expectations are over how the second mover will respond. The second is that the first mover is risk neutral with respect to the second mover’s contribution. This assumption can be modified at the cost of unnecessary complexity. The parameter $\gamma \geq 0$ captures how important the other player’s contribution is to the first mover’s choice. If $\gamma > 0$, the first mover does take into account the second mover’s reaction to his contribution. This motivates the following:

**Definition 1.** A first mover is an altruistic leader if and only if $\gamma > 0$. 
In our setup, our definition of altruistic leadership is consistent with that of a pure altruist who takes the second mover’s contribution into account, or of an impure altruist who cares about self-signaling (e.g. Benabou and Tirole, 2006). It is also consistent with someone who only cares about how much he influences a second mover to give, independently of its effect on the charity. Our design does not distinguish between these different motivations if \( \gamma > 0 \).

The third term captures the cost of social expectations. The social expectations of \( i \) in scenario \( s \) term, \( E_{soc}(m_i|s) \) for \( s \in \{Can, Can't\} \), is a somewhat complicated object. Although it is written in shorthand, what it really captures is the first mover’s expectation over what others expect the first mover will do in scenario \( s \). A more appropriate way of writing the term would be \( E_i(E_{-i}(m_i|s)) \), where \( E_{-i}(m_i|s) \) is the average expectation of \( m_i \) from all players who are not \( i \) in scenario \( s \). We take this term as exogenous in this setup. This term is endogenized in chapter 3. Inside the cost function \( c \) is the difference between social expectations over leader \( i \)'s contribution \( (E_{soc}(m_i|s)) \) and what he actually contributes \( (m_i) \) in scenario \( s \). If the latter term is lower than the former, the first mover is disappointing social expectations. The smaller the distance, the smaller the disappointment and the smaller the cost the first mover bears. If the latter term is higher than the former, the first mover surpassed expectations in scenario \( s \). The greater the distance, the greater the benefit the first mover receives. We capture this by having \( c \) be convex \( (c' > 0, c'' > 0) \). The parameter \( \alpha \geq 0 \) captures whether individuals care about social expectations. Since, as we show below, \( \alpha > 0 \) can cause first mover to contribute in order to avoid a cost, this parameter justifies the following:

**Definition 2.** A first mover is a reluctant leader if and only if \( \alpha > 0 \).

For the Cannot Consider scenario, we will denote \( n_i \in [0, 1] \) \((n_j \in [0, 1])\) as the amount that the first mover (second mover) keeps for himself. The choice the first mover makes in the Cannot Consider scenario is:
\[
\max_{n_i \in [0,1]} EU(n_i, n_j, \text{Can't}) = \\
u(n_i) + (1 - n_i) + \gamma(1 - E_i(n_j)) - \alpha(E_{Soc}(n_i|\text{Can't}) - n_i)
\]

Note that there are two differences in the leader’s maximization problem in the Can Consider scenario \(s = \text{Can}\) and in the Cannot Consider scenario \(s = \text{Can’t}\). The first is that in the Cannot consider scenario, the second mover’s contribution does not depend on the first mover’s contribution. The second is that the scenario is different, so \(E_{Soc}(\cdot|\text{Can}) \preceq E_{Soc}(\cdot|\text{Can’t})\). This is reflected in the first order conditions for both scenarios.

For the Can Consider scenario, the first order condition is:

\[
\frac{du}{dm_i} = 1 + \gamma \frac{dE_i(m_j|m_i)}{dm_i} + \alpha \frac{\partial c(E_{Soc}(m_i|\text{Can}) - m_i)}{\partial m_i}
\]

For the Cannot Consider scenario, the first order condition is:

\[
\frac{du}{dn_i} = 1 + \frac{\partial c(E_{Soc}(n_i|\text{Can’t}) - n_i)}{\partial n_i}
\]

The extra term in the Can Consider scenario FOC comes from the effect the leader has on the follower’s contribution. If the leader cares about the impact he has on the follower’s contribution \((\gamma > 0)\) and believes his contribution has a positive impact on the follower up to the equilibrium in the Can Consider scenario \(\left(\frac{dE_i(m_j|x)}{dx} > 0 \text{ for } x = m_i^*\right)\), then the right hand side of the Can Consider scenario FOC is larger than in the Cannot Consider scenario FOC. We call this the \textit{positive impact} condition. In an interior solution, this would motivate the leader to set \(m_i^* > n_i^*\). The intuition is that the marginal impact of contributing to the public good in the Can Consider scenario is higher because of its effect on the follower’s contribution. The zero impact condition is defined by \(\frac{dE_i(m_j|x)}{dx} = 0 \text{ for } x = m_i^*\). The marginal impact in this case would be the same in expectation in the Can Consider and the Cannot Consider scenarios. On the other hand, the leader might set \(m_i^* > n_i^*\) even if \(\gamma = 0\) when the individual cares about social expectations \((\alpha > 0)\) and the social expectation
associated with the Can Consider scenario is higher \( E_{Soc}(\cdot|Can') > E_{Soc}(\cdot|Can) \). Intuitively, the leader might give more in the Cannot Consider scenario in order to avoid the added social expectation cost. This establishes the following

**Result 2.** An individual gives more in the can Consider scenario \( (m_i^* > n_i^*) \) ⇔

- He is an altruistic leader who expects his contribution will have a positive impact on the follower’s contribution: \( \gamma > 0 \) and \( \frac{dE(m_j|x)}{dx} > 0 \) for \( x = m_i^* \), or
- He is a reluctant leader who holds higher social expectations in the Can Consider scenario: \( \alpha > 0 \) and \( E_{Soc}(m_i|Can) > E_{Soc}(m_i|Can') \)

What this result shows is that if we observe a higher contribution in the Can Consider scenario rather than the Cannot Consider scenario, we might not be able to distinguish whether it is due to altruistic leadership or to reluctant leadership. In order to distinguish between the two, we ask leaders to replicate between their choice in the Can Consider scenario, \( U(m_i^*, E(m_j|m_i^*), Can) \) or their choice in the Cannot Consider scenario, \( U(n_i^*, E(n_j^*), Can') \).

What this means is that they must make one of their two decisions twice as likely to be implemented. The problem can be written as follows:

\[
\max_{x \in \{\frac{1}{3}, \frac{2}{3}\}} \{xU(m_i^*, E(m_j|m_i^*), Can) + (1 - x)U(n_i^*, E(n_j), Can') \}
\]

It is clear that a straightforward comparison cannot be made of the two utilities, even ignoring the social expectation cost. In particular, it is unclear whether the cost of giving less \( m_i^* < n_i^* \) would be compensated by a higher contribution of the follower \( E(m_j|m_i^*) > E(n_j^*) \).

The sufficient condition for making this comparison is the following:

**Definition 3.** We say the leader’s contributions are marginally comparable in expectation if \( E(m_j|n_i^*) \geq E(n_j) \). That is, if the leader expects that what the follower would have given if he would have observed the Cannot Consider contribution \( n_i^* \) in the Can Consider scenario were at least as high as what the follower gives in the Cannot Consider scenario.

If the leader’s contributions are marginally comparable in expectation, then if the leader gives more in the Can Consider scenario we can infer he did not do so to compensate for a
decrease in the overall contribution to the public good. To see this, note that this condition holds if and only if $U(n^{*}_{i}, E(m_{j}|n^{*}_{i}), Can't) \geq U(n^{*}_{i}, E(n_{j}), Can't)$, with the inequality being strict if and only if $\gamma > 0$. Further, $U(m^{*}_{i}, E(m_{j}|m^{*}_{i}), Can) > U(n^{*}_{i}, E(m_{j}|n^{*}_{i}), Can't)$ if and only if $\gamma > 0$, since the social expectation is additively separable and the leader preferred $m^{*}_{i}$ to $n^{*}_{i}$ in the Can Consider scenario. Finally, $U(m^{*}_{i}, E(m_{j}|m^{*}_{i}), Can) > U(m^{*}_{i}, E(m_{j}|m^{*}_{i}), Can't)$ if and only if $\gamma > 0$ or $E_{Soc}(Can) < E_{Soc}(Can't)$, but if the latter possibility is true then by Result 1, $\gamma > 0$ must also be true. Putting these inequalities together, we can establish that when $m^{*}_{i} > n^{*}_{i}$, $U(m^{*}_{i}, E(m_{j}|m^{*}_{i}), Can) > U(n^{*}_{i}, E(n_{j}), Can't)$ if and only if $\gamma > 0$. Putting it all together, we can establish the following:

**Result 3.** If the positive impact and marginal comparability conditions hold, a leader who gives more in the Can Consider scenario ($m^{*}_{i} > n^{*}_{i}$), replicates the Can (Cannot) Consider scenario only if $\gamma > 0$ ($\alpha > 0$).

An analogous result can be derived for the case where individuals believe their contribution has a flat impact on followers.

**Result 4.** If the zero impact and marginal comparability conditions hold, a leader who gives more in the Can Consider scenario ($m^{*}_{i} > n^{*}_{i}$), replicates the Can (Cannot) Consider scenario only if $\gamma > 0$ ($\alpha > 0$).

Note that concluding $\alpha > 0$ in this case does not require knowledge of what scenario the individual replicates.

These results give us a sufficient statistic for determining whether the leader is reluctant or altruistic. Note that this is a within subject question, for which we need the three decisions to be made by the same person. We now turn to the description of the design.

### 2.4 The Experimental Design

Participants were recruited through the Harvard Decision Science Lab subject pool. They were paid a $5 show up fee. A session is composed of two sign up times, twenty minutes
apart. The first group is composed of leaders, the second of followers. Note that random assignment between these groups was not necessary, since the comparison will be within subjects and between leaders. Subjects are received in a lobby, where they are asked to sign a consent form with the experimenter’s contact information as well as that of the Committee on the Use of Human Subjects in Research at Harvard University. Once all subjects finish filling out the consent forms, they draw a piece of paper and a key. The paper assigns them to their cubicles, while it is explained to them that the key opens a personal locker in a private room where they pick up the payment at the end of the session.

Leaders and followers are in separate rooms. Each room has twelve cubicles, and 15 slots were offered per room during recruitment to account for attrition. Once seated, the introductory instructions are read out loud. Special attention was paid to make sure the instructions were clear, and several pilots were conducted to test for this. The introductory instructions explain that they will make three divisions of one set of ten dollars between themselves and a charity, and one division will be selected at random to be implemented. In order to facilitate comprehension, the instructions specified that they will divide their ten dollars in three scenarios described in different boxes, and that at the end one of the boxes would be selected by the lottery and the description and corresponding decision followed. It was explained to them that whichever box the lottery chose to implement, the other two boxes became irrelevant for how their ten dollars were to be divided. They were therefore encouraged to make each decision as if it were the only one they were making with those ten dollars. The introductory instructions also made clear that their decisions were anonymous, and they were not asked for their name at any point on the computer prompts. The charity they divide the money with is the East Africa Food Crisis Relief Fund of Save the Children, which addresses a drought in East Africa. The remoteness and temporality of the charity was chosen in order to make it hard to argue that the benefits of the charity impacted the recipients in any direct way. The instructions also stated that Save the Children would not know where the donation that was raised through the experiment came from, in order to avoid social expectations considerations towards the charity.
Note that the benefits of Save the Children’s work may be rival and excludable for the charity recipients. The way in which we conceptualize Save the Children as a public good is that potential contributors care about helping the cause. The sense in which Save the Children is a charity, then, presupposes other-regarding motivations.

The rest of the instructions were read on their computer, which ran the session using zTree (Fischbacher, 2007). The instructions avoided using words that would lead to desirability bias, such as ‘leadership’, ‘followership’ or ‘influence’, and associated words. Not using these key words made it a challenge to explain to subjects in a way that was clear and concise what the difference was between the subtly different Can and Cannot Consider scenarios. We approached these difficulties in several ways. One general solution was to include a questionnaire screen after the screens with the most difficult or novel explanations. After they answered the questionnaire, a screen with the answers would appear, specifying which questions they answered correctly or incorrectly. From there, they would see the original instructions one last time before moving on. Below we present the screenshots of the first explanation, questionnaire and questionnaire answers of the stage where we introduce the Can Consider and the Cannot Consider scenario.

First movers were randomly assigned to one of two treatments. The only difference between the treatments were whether they made a decision in the Can Consider scenario or in the Can’t Consider scenario first. As has been used in this paper, these scenarios were referred to in the instructions as ‘Can Consider’ and ‘Cannot Consider’. We chose these names because they help establish what the difference between both scenarios is while avoiding experimenter demand effects. Without the name of these scenarios, subjects found it harder to understand and remember what the difference between the scenarios was. Explaining the difference in scenarios in a neutral way was difficult. The challenge was that it becomes cognitively challenging to keep track of the difference between two sequences of subjects making a contribution decision in which the only difference is when one subject observes what the other did. In order to aid comprehension, we included visual timelines. In screens explaining one of the three decisions, an extra visual aid was to have a diagram
For your decisions, you will make choices of how to divide ten dollars between yourself and the Children in two different scenarios. We will describe the scenarios here. In both scenarios, you and someone else will make a choice about how much money to give to the Children. We will refer to this other person as 'Mr 2'.

You will make your choice before Mr 2 makes his. Mr 2 knows someone makes a choice before he does, and will be told what you chose. The difference between the scenarios is when Mr 2 is told what you chose. In one scenario Mr 2 is told what you chose first, and then makes his own choice. This scenario gives Mr 2 an opportunity to consider what you chose before making his own choice. We will call this the 'Can-Consider' scenario. In the other scenario, Mr 2 makes his own choice first and only then is told what you chose. In this second scenario, Mr 2 does not consider what you chose before making his own choice. We will call this the 'Cannot-Consider' scenario. The following describes the scenarios in more detail:

'Can-Consider' scenario: You make a choice of how much money to give to the Children. Mr 2 is told what you chose. Mr 2 then makes his own choice. Once no one has further choices to make, Mr 2 would do other tasks such as playing a guessing game and answering a survey. The following shows the sequence of events:

<table>
<thead>
<tr>
<th>Session begins</th>
<th>You make your choice</th>
<th>Mr 2 sees what you gave</th>
<th>Mr 2 makes his choice</th>
<th>No one has further decision to make</th>
<th>Mr 2 sees what you gave</th>
<th>Mr 2 does other tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

When Mr 2 makes his choice, the only information he would have about what others chose would be what you chose.

'Cannot-Consider' scenario: As in the Can-Consider scenario, first you make a choice of how much money to give to the Children. The order of events is different from that point on. Before being told what you chose, Mr 2 makes his own choice. Only after no one has further choices to make, Mr 2 will be told what you chose. Mr 2 would then do the other tasks described in the Can-Consider scenario. The following shows the sequence of events:

<table>
<thead>
<tr>
<th>Session begins</th>
<th>You make your choice</th>
<th>Mr 2 makes his choice</th>
<th>No one has further decision to make</th>
<th>Mr 2 sees what you gave</th>
<th>Mr 2 does other tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

When Mr 2 makes his choice, he would not have any information about what others chose.

Notice that step 3 in the timeline of the Cannot-Consider scenario does not happen until step 5 in the timeline of the Cannot-Consider scenario. Before we continue, we will ask you some questions about what you just read to make sure the instructions are clear.

Figure 2.1: Introductory Presentation Of Scenarios

Figure 2.2: Questionnaire of Introductory Presentation Of Scenarios
of the three boxes and describe inside the boxes the current and past decisions, as can be seen in Figure 2.4. When talking about a person in the role of a first mover (second mover), we referred to that person as ‘Mr 1’ (‘Mr 2’). This label genders the players in order to make it easier to refer to them throughout the instructions. Favoring an explanation that kept gender neutral would make referring to Mr 1 and Mr 2 hard to keep track of.

The very first screen all first movers saw explained both scenarios and made it clear that they would make decisions in each scenario. The subjects next passed on to their first decision, and after they had made that the decision, their second decision was presented. A point was made to make everything that happened in the instructions of both scenarios identical except for the timing of the announcement of the leader’s contribution. In particular, in both scenarios it was explained that ‘Mr 2’ (the second mover) would make his own contribution. In the Can Consider scenario, the only information the second mover would have about what others gave was what the first mover gave. In the Cannot Consider scenario, the second mover would not have any information about what others gave before making
his decision, and the second mover would know how much the leader gave afterwards. It was important that the leader knew that the second mover would make a decision and his available information when doing so in order to be able to compare utilities as specified in the model. Below we present the introductory screenshot of the first decision in the treatment where the first mover makes a decision in the Can Consider scenario first.

![Explanation Of First Contribution For Leaders in Treatment 1](image)

**Figure 2.4: Explanation Of First Contribution For Leaders in Treatment 1**

Up to this point, even though subjects knew there would be a third decision, they were not aware of what their third decision was going to be. Once they made their second decision, they are offered to have the third box replicate the scenario and corresponding decision of the first box or the second box. It is explained to them that whatever they choose, two boxes out of three will be identical to that choice, so they are making it twice as likely to be selected by the lottery than what they did not choose. An important design feature should be pointed out here. The layout and the design of this question is set up in order to avoid a default or favored option. The question asks the leader to select one of the two decisions to replicate, and the only difference in their description is that one is described
before the other. The option that is described first is the one corresponding to the first scenario, so this changes per treatment.2

After the first movers made their three decisions, but before they were informed which of the three lotteries was selected, the first movers were asked to play a ‘guessing game’. In this game, they had to guess how much people in different scenarios gave to Save the Children on average. They were told that the person with the closest guess won an extra five dollars. They were first asked to guess about Mr 2 in the Can and the Cannot Consider scenario. For the Can Consider scenario, they were asked to guess how much Mr 2 gave to Save the Children for each integer amount between 0 and 10 dollars that Mr 1 gave. For the Cannot Consider scenario, they were asked to guess how much Mr 2 contributed unconditionally, as they were not able to consider what Mr 1 gave. We used these guesses to verify whether the sufficient condition on expectations holds. After they make this decision, they also guess what Mr 1 (the first mover) gave in the Can and Cannot Consider scenarios. Finally, they guess what an individual who does not receive any information about what others gave and whose contribution will be kept private would give.

After the guessing game, first movers are asked to answer a series of questions. These questions are of two types. The first type are sociodemographic questions (including age, gender, education). The second are personality questions. Some of these questions ask about past experiences of leadership positions, such as being the captain of a team, being the oldest brother or having children. Other questions were taken from the psychology literature on leadership. Questions from the Machiavellianism instrument (Christie and Geis, 1970, used in Gunnthorsdottira, McCabe and Vernono, 2002), the modified Machiavellianism instrument (Dahling, Whitaker and Levy, 2009), the leader behavior description questionnaire (Stogdill and Coons, 1957) and the multifactor leadership questionnaire (Avolio and Bass, 1995) were used.

2 This design allows us to avoid a critique to which the Dana et al, 2006 study mentioned earlier is subject, which is that individuals chose the exit option when it was offered to them because it being offered signaled it was desirable. An example of a more neutral approach would have allowed subjects to decide whether they would be willing to exit for a dollar less or a dollar more - either across different subjects or randomizing which one was implemented.
Finally, first movers filled out an exit survey that asked open questions about what was going through their minds when they were making the decisions and playing the guessing game.

2.5 Results

We first report some statistics from our leaders. All standard errors in our results are clustered at the session level. The average age is well above college graduation, at 30.6. This is confirmed by 74% of our sample having completed college, 21% still in college and 5% only having completed high school. This is encouraging, as a common concern in lab experiments is that it draws too heavily from college students. About half of the sample is female. None of these characteristics are significantly different from each other when compared across treatments, which suggests the randomization was done properly.

<table>
<thead>
<tr>
<th></th>
<th>Can’t Consider First</th>
<th>Can Consider First</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td>30.068</td>
<td>31.25</td>
<td>30.631</td>
</tr>
<tr>
<td></td>
<td>(13.279)</td>
<td>(12.469)</td>
<td>(12.836)</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td>0.568</td>
<td>0.575</td>
<td>0.571</td>
</tr>
<tr>
<td></td>
<td>(0.501)</td>
<td>(0.501)</td>
<td>(0.498)</td>
</tr>
<tr>
<td><strong>PassedCollege</strong></td>
<td>0.682</td>
<td>0.8</td>
<td>0.738</td>
</tr>
<tr>
<td></td>
<td>(0.471)</td>
<td>(0.405)</td>
<td>(0.442)</td>
</tr>
<tr>
<td><strong>InCollege</strong></td>
<td>0.25</td>
<td>0.175</td>
<td>0.214</td>
</tr>
<tr>
<td></td>
<td>(0.438)</td>
<td>(0.385)</td>
<td>(0.413)</td>
</tr>
<tr>
<td><strong>PassedHighSchool</strong></td>
<td>0.068</td>
<td>0.025</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>(0.255)</td>
<td>(0.158)</td>
<td>(0.214)</td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>44</td>
<td>40</td>
<td>84</td>
</tr>
</tbody>
</table>

Standard deviations in parentheses.

First two columns show results depending on which decision was made first.

Turning to our first result, we can see that out of the 84 subjects who have been in the first mover position, 20 of them (24%) have contributed more in the Can Consider scenario, 4 of them (5%) in the Cannot Consider scenario, and the rest contributed the same amount in both scenarios. In terms of the intensive margin, subjects have given 55 cents
more on average in the Can Consider scenario, a 20% increase from the Cannot Consider scenario. If we break it down by treatment, the main effect is taken up by Treatment 1, where the difference is significant at the .01% level. It is shy of significance at the 1% level for Treatment 2 (p-value .111), but the effect is in the same direction. For each treatment, the null of 0% of individuals giving more in the Can Consider scenario can be rejected with a .01% significance level.

As an alternative test for the impact of being in a first mover position on individuals’ contribution decisions, we can compare the first contribution of individuals in Treatment 1 and Treatment 2. These first contributions are, respectively, in the Can Consider scenario and the Cannot Consider scenario. Individuals give 1.9 dollars more in the former case, although this is shy of significance (p-value of .143) with clustered standard errors. Notice that when individuals make this decision, they know what the Can Consider and Cannot Consider scenarios in which they will make their decisions are. The difference in their contributions, then, captures more of a sequencing effect than the pure between-individual decision across scenarios.

The next table presents the replication decisions. As can be seen from the table, most subjects who gave more in the Can Consider scenario replicated the Cannot Consider scenario. This provides prima facie evidence for reluctant leadership. It is worth noting that when all subjects are pooled or when subjects did not give more in the Can Consider scenario, the replication of the Can Consider scenario is not significantly different from 50% (p-values of 1 for Treatment 1, .117 for Treatment 2, .252 for both). However, for those subjects who give more in the Can Consider scenario, it is significantly different at the 1% level for Treatment 1 (p-value of .066) shy of significance for Treatment 2 (p-value of .163) and significant at the 5% level (p-value of .028) for the aggregate of both treatments. This suggests that replication decisions were random only when they did not have an impact on the subject’s material payoffs. Further, the replication of the Cannot Consider scenario is significantly higher for subjects who give more in the Can Consider scenario (p-value .018) than for those who don’t. Although the same results are not significant when the Cannot
Table 2.2: Can and Cannot Consider Decisions

<table>
<thead>
<tr>
<th></th>
<th>Can Consider Contribution</th>
<th>Can’t Consider Contribution</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First</td>
<td>First</td>
<td>Total</td>
</tr>
<tr>
<td>Can Consider Contribution</td>
<td>4.28**</td>
<td>2.5**</td>
<td>3.35***</td>
</tr>
<tr>
<td></td>
<td>(0.95)</td>
<td>(0.50)</td>
<td>(0.43)</td>
</tr>
<tr>
<td>Can’t Consider Contribution</td>
<td>3.30**</td>
<td>2.34**</td>
<td>2.80***</td>
</tr>
<tr>
<td></td>
<td>(0.85)</td>
<td>(0.48)</td>
<td>(0.48)</td>
</tr>
<tr>
<td>Difference</td>
<td>0.98*</td>
<td>0.16</td>
<td>0.55***</td>
</tr>
<tr>
<td></td>
<td>(0.43)</td>
<td>(0.09)</td>
<td>(0.20)</td>
</tr>
<tr>
<td>Count ( m_{<em>i} &gt; n_i^</em> )</td>
<td>0.30**</td>
<td>0.18**</td>
<td>0.24***</td>
</tr>
<tr>
<td></td>
<td>(0.09)</td>
<td>(0.05)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>Count ( m_{<em>i} &lt; n_i^</em> )</td>
<td>0.05</td>
<td>0.05</td>
<td>0.048**</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>N</td>
<td>40</td>
<td>44</td>
<td>84</td>
</tr>
</tbody>
</table>

Standard errors in parentheses
First two columns show results depending on which decision was made first.
* if \( p < 0.1 \), ** if \( p < 0.05 \), *** if \( p < 0.001 \)

Consider scenario is chosen first, the effect is in the same direction.

2.5.1 Incorporating Beliefs

The above analysis does not take leaders’ beliefs over the followers’ response into account. As we show in the Conceptual Framework section, this means we may be miscategorizing individuals. In this section, we take elicited beliefs into account. As would be expected from the challenges related to eliciting beliefs, this reduces the amount of individuals we can classify as altruistic or reluctant, but the pattern of results is the same.

In order to incorporate the sufficient conditions for identifying reluctant and altruistic leaders, we need to operationalize the slope of the predicted reaction function the second mover has given the first mover’s decision: \( \frac{E_i(m_j|x)}{dx} > 0 \) for \( x = m_i^* \). In order to calculate the slope, we ran a regression of the leader’s prediction of what the follower did on leaders’ contribution up to the largest value the leader actually contributed in one of the scenarios. That is, if a leader contributed 5 dollars in the Can Consider scenario and 7 in the Cannot...
Table 2.3: Decision To Replicate The Can Consider Scenario

<table>
<thead>
<tr>
<th></th>
<th>Can Consider</th>
<th>Can’t Consider</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First</td>
<td>First</td>
<td>Total</td>
</tr>
<tr>
<td>Subjects with $m^* &gt; n^*$</td>
<td>0.16*</td>
<td>0.25</td>
<td>0.2**</td>
</tr>
<tr>
<td></td>
<td>(0.15)</td>
<td>(0.15)</td>
<td>(0.11)</td>
</tr>
<tr>
<td>Subjects with $m^* \leq n^*$</td>
<td>0.64</td>
<td>0.39</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>(0.10)</td>
<td>(0.08)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>Difference subjects with $m^* &gt; n^<em>$ and $m^</em> \leq n^*$</td>
<td>-0.48*</td>
<td>-0.14</td>
<td>-0.3*</td>
</tr>
<tr>
<td></td>
<td>(0.21)</td>
<td>(0.39)</td>
<td>(0.15)</td>
</tr>
<tr>
<td>All subjects</td>
<td>0.5</td>
<td>0.36</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>(0.08)</td>
<td>(0.08)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>N</td>
<td>40</td>
<td>44</td>
<td>84</td>
</tr>
</tbody>
</table>

Standard errors in parentheses

First two columns show results depending on which decision was made first.

Null hypothesis for the Difference is 0, for all other variables is 1/2.

* if $p < 0.1$, ** if $p < 0.05$, *** if $p < 0.001$

Consider scenario, we would run a regression on what that leader guessed the follower would give on a leader giving between 0 and 7 dollars. The regression allowed us to calculate the slope and abstract from any noisy responses a leader might have given in his guessing game. We then classified this slope according to whether it was positive, zero or negative to determine whether the positive or zero impact condition held.

In order to check for the marginal comparability condition $E(m_j|n_i) \geq E(n_j^*)$, we compared each first movers’ prediction of what second movers do in the Cannot Consider scenario to the prediction of what second movers do after observing the contribution the first mover actually gave. Table 2.4 presents a frequency distribution of these conditions.

As can be seen from the table, most first movers think that their contribution has a positive impact on the second mover’s contribution. The marginality condition holds about half of the time, however, which will make it hard to identify leaders’ types.

Table 2.5 presents our main findings. We classify reluctant and altruistic leaders according to the criteria specified in the Conceptual Framework section of the paper. We find that 8% (7 individuals) of the sample can be precisely identified as a reluctant leader, while only
Table 2.4: Impact and Marginality Conditions

<table>
<thead>
<tr>
<th></th>
<th>Can Consider First</th>
<th>Can’t Consider First</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Leaders</strong> ((m^* &gt; n^*))</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive Impact</td>
<td>0.75**</td>
<td>0.75**</td>
<td>0.75***</td>
</tr>
<tr>
<td></td>
<td>(0.14)</td>
<td>(1.5)</td>
<td>(0.12)</td>
</tr>
<tr>
<td>Zero Impact</td>
<td>0.16</td>
<td>0.125</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>(0.09)</td>
<td>(0.13)</td>
<td>(0.10)</td>
</tr>
<tr>
<td>Marginality Condition</td>
<td>0.41*</td>
<td>0.625**</td>
<td>0.5**</td>
</tr>
<tr>
<td></td>
<td>(0.21)</td>
<td>(0.23)</td>
<td>(0.16)</td>
</tr>
<tr>
<td>Positive Impact &amp; Marginality Condition</td>
<td>0.16</td>
<td>0.5*</td>
<td>0.3*</td>
</tr>
<tr>
<td></td>
<td>(0.10)</td>
<td>(0.23)</td>
<td>(0.13)</td>
</tr>
<tr>
<td>Zero Impact &amp; Marginality Condition</td>
<td>0.16</td>
<td>0.125</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>(0.09)</td>
<td>(0.13)</td>
<td>(0.10)</td>
</tr>
<tr>
<td>N</td>
<td>12</td>
<td>18</td>
<td>20</td>
</tr>
</tbody>
</table>

| **All**                     |                    |                      |             |
| Positive Impact             | 0.7***             | 0.61***              | 0.65***     |
|                             | (.088)             | (.089)               | (.05)       |
| Zero Impact                 | 0.25**             | 0.27**               | 0.26**      |
|                             | (0.07)             | (0.084)              | (0.034)     |
| Marginality Condition       | 0.52***            | 0.63***              | 0.583***    |
|                             | (.085)             | (.093)               | (.054)      |
| Positive Impact & Marginality Condition | 0.3** | 0.34** | 0.32** |
|                             | (0.06)             | (0.078)              | (0.038)     |
| Zero Impact & Marginality Condition | 0.175** | 0.22** | 0.20** |
|                             | (0.062)            | (0.074)              | (0.039)     |
| N                           | 40                 | 44                   | 84          |

Standard errors in parentheses
First two columns show results depending on which decision was made first.
* if \(p < 0.1\), ** if \(p < 0.05\), *** if \(p < 0.001\)
2% of the sample (2 individuals) can be identified as altruistic leaders. Altruistic leaders are shy of being significantly different from zero (p-value of .15), while reluctant leaders are significantly different from zero at the 5% level. There are significantly more reluctant leaders than altruistic leaders at the 5% level (p-value of .047). As foreshadowed by Table 4, there are 11 more individuals that do not fulfill all the requirements to be classified as one or the other type of leader. These 11 individuals are also significantly different from zero at the 5% level (p-value of .024).

Table 2.5: Types of Leaders

<table>
<thead>
<tr>
<th>Can</th>
<th>Can’t</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consider</td>
<td>Consider</td>
<td></td>
</tr>
<tr>
<td>First</td>
<td>First</td>
<td>Total</td>
</tr>
<tr>
<td>Reluctant Leader</td>
<td>.075*</td>
<td>.09*</td>
</tr>
<tr>
<td></td>
<td>(.035)</td>
<td>(.048)</td>
</tr>
<tr>
<td>Altruistic Leader</td>
<td>.025</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td>(.024)</td>
<td>(.022)</td>
</tr>
<tr>
<td>Difference</td>
<td>.05</td>
<td>.07*</td>
</tr>
<tr>
<td></td>
<td>(.05)</td>
<td>(.05)</td>
</tr>
<tr>
<td>Can’t Distinguish</td>
<td>0.20**</td>
<td>0.068</td>
</tr>
<tr>
<td></td>
<td>(.074)</td>
<td>(.049)</td>
</tr>
<tr>
<td>All Leaders</td>
<td>0.3**</td>
<td>0.18**</td>
</tr>
<tr>
<td></td>
<td>.085</td>
<td>.053</td>
</tr>
<tr>
<td>N</td>
<td>40</td>
<td>44</td>
</tr>
</tbody>
</table>

Standard errors in parentheses
First two columns show results depending on which decision was made first.
Results show percentage of the population per column
Standard errors for Difference are from a t-test verifying whether the value is positive
* if p < 0.1, ** if p < 0.05, *** if p < 0.001

2.5.2 Robustness Checks

One challenge with the design is the fact that the instructions were lengthy. Writing instructions that were unambiguous and at the same time were not leading meant that sacrifices needed to be made in terms of conciseness. We mentioned earlier that we took
several steps in order to control for this issue. The experiment was piloted several times in order to reach an acceptable level of understanding on the part of the subjects. As a further robustness check, we asked in the exit survey whether any of the tasks or questions were confusing. 53 of the 79 who answered the exit survey answered no. Of the rest, 7 out of 79 answered in a way that was difficult to interpret, while 19 out of 79 expressed varying degrees of confusion. We replicated the analysis without the 31 subjects who did not answer the exit survey, answered ambiguously whether they were confused or answered that they were confused. The results are qualitatively similar.

Note that the results rely on a within subjects comparison: what they do in the Can versus Cannot Consider scenario. This biases the results against finding an effect, due to anchoring effects (Kahneman and Tversky, 1974) or consistency bias (Markus, 1986, Scharfe and Bartholomew, 1998), which we would expect to be particularly important in a setting where the difference in scenarios is subtle. A second design feature that is a challenge to finding an effect is that these results demand that we are able to elicit beliefs correctly, and that these elicited beliefs have enough consistency that we are able to fulfill the two conditions we have specified. Although belief elicitation reduces the effect of our findings, the fact that we have found this consistency in enough of our subjects to find statistically significant results is encouraging. One may be worried about a consistency bias driving these results, which although we cannot rule out, it is worth pointing out that we took care in avoiding signaling through language or the experimental sequence how it was that they were meant to be consistent.

2.6 Conclusion

This paper has shown novel evidence for reluctance in leadership. This motivation calls for thinking differently about policies aimed at increasing public good provision. In chapter 3 of the dissertation, I explore the equilibrium consequences of a situation where individuals who care about social expectations take actions publicly.

There is something counterintuitive about social expectations driving leadership. If
individuals respond positively to social information about public good contribution because they are surprised (positively or negatively) about what others contribute, then a social expectation to lead is therefore an expectation to act in a way that is unexpected. There are different ways to resolve this apparent contradiction. The first is that individuals in different groups have different expectations over how much they and others should give. This may explain the findings in leadership experiments in which they find that higher status individuals contribute and lead more than lower status individuals (e.g. Jack and Recalde, 2013, Kumru and Vesterlund, 2010, d’Adda, 2011). Individuals in higher status groups may be expected to give higher amounts than those in lower status groups. Secondly, individuals may overestimate others’ social expectations in the Can-Consider scenario. This may be an unexplored heterogeneity in individuals’ beliefs that explains variation in leadership. Chapter 3 shows how incorrect social expectations can be sustained as an equilibrium outcome of a strategic interaction where individuals are motivated to fulfill their perceived social expectations. Thirdly, the reason second movers follow may simply not be due to information being transmitted about the quality of the public good. As a companion paper, Fernández Duque and Hiscox (2015) present evidence showing that second follow, and that they also respond to social expectations.

Given the amount of evidence in favor of social-image as a driver of pro-social behavior, it is also counterintuitive that individuals would lead in an anonymous setting. One way to reconcile the two views is that anonymous leading is the result of a social heuristic. Individuals may be leading altruistically in anonymous settings because most of the time social image matters because settings are not anonymous and it is cognitively costly to make the distinction. There is also a more rational explanation. If individuals act in a way that satisfies what they think others expect from them independently of who can see their actions, they will accrue the rents from the times that do impact their reputation and avoid false negatives. Indeed, being ‘surprised’ acting pro-socially without a clear reputation motive is a much better signal of type than when the reputation motive is apparent. So leading in anonymous settings may be worthwhile in expectation, even if being observed is
unlikely. Chapter 3 derives conditions under which a strategic interaction where players having social expectation concerns may have the unintended consequence that all players are acting to satisfy what they believe to be what others expect out of a public good action although nobody wants that action to be taken.
Chapter 3

Uninformed leaders

3.1 Introduction

Mohamed Bouazizi was 28 years old on December 17th, 2010. He sold fruit from his stand, was educated in a one room country school and was not politically engaged. According to his aunt, his main aspiration was to buy a truck to improve his fruit sales. In the morning, a police officer confiscated his cart under dubious allegations. Upset, Mohamed went to complain to the municipal government, but he was ignored. In response, he reportedly told the officials that if they didn't see him, he would burn himself. Shortly later, he doused himself with gasoline and set himself on fire.

On December 18th, a small group of protesters gathered. This was recorded on a phone and uploaded to Facebook. In contrast to other protests which had been silenced by the Tunisian regime, this one was widely spread partly due to the sudden increase in the popularity of the social networking site. Less than month later President Zine El Abidine Ben Ali resigned. Less than two years later 5 rulers were forced out of power, and other protests sparked across the region.

This sequence of events is one of the starkest illustrations of ‘behavioral cascades’, a phenomenon where one person’s actions starts an avalanche of similar actions. Indeed, behavioral cascades models have been used to explain the Arab Spring as well as the quick
fall of the communist regimes in Eastern Europe (Kuran, 1997; Lohman, 1994). These models are appealing because they fit the basic facts of an initial demonstration which spread quickly. Moreover, a particular subset of these models predicts the fact that the demonstrations were so surprising (Kuran, 1995). Indeed, individuals and organizations whose job it is to predict regime failures consistently were wrong when predicting Tunisia’s revolution (Goodwin, 2011). These models assume that there is a misapprehension of the extent of discomfort or willingness to protest against the regime. Despite the advantage of these subset of models, all make a stark prediction about the nature of the first mover that is not supported by the facts of the Arab Spring. Indeed, the models predict that the first mover is knowledgeable and trustworthy (e.g. Lupia and McCubbins, 1998; Hermalin, 1998, Lohman, 1994; Bicchieri, 2005; Majumdar and Mukand, 2008; Maskin and Tirole, 2004), an extremist (Kuran, 1997; Granovetter, 1978) or a prominent cooperator who is showing others how to act (Acemoglu and Jackson, 2014). In this paper, we propose an alternative model which explains the stylized facts of the Arab Spring while explaining how the first mover can be uninformed, obscure and politically inactive.

The starting point of our theory of an uninformed leader is a situation in which agents are deciding whether to protest based on their perceptions of the extent of discontent in a society. There is a public misperception of discontent due to the fact that individuals act according to what they think others want, and that leads people to infer from what they observe that nobody wants to protest. We refer to this equilibrium situation as a ‘pluralistic ignorance’, borrowing the term from social psychology (e.g. Allport, 1924). Even though this is not the first model of equilibrium misinformation, it is the first one where an individual cares not about others’ actions or how others’ private information informs his preferences, but about others’ actual preferences which they infer from actions. That is, individuals display preference complementarities as opposed to strategic complementarities or uncertainty about their own preferences. This way of setting up the model more closely captures the intuition behind pluralistic ignorance as studied in social psychology and sociology, where it was classically defined as the misperception of social norms that results
from observing people who are acting differently from their private preferences out of a concern for the social consequences - actions that reinforce the erroneous social norms. Modeling norms as the belief over the distribution of private preferences for an action has been used fruitfully in economics (e.g. Benabou and Tirole, 2011; Sliwka, 2008).

Other formal models of equilibrium misinformation rely on informational cascade arguments to explain the spread of misinformation that leads to a lack of protests (e.g. Kuran, 1997). These arguments assume that there is an objective state of the world in which individuals have similar unknown preferences and receive informative but possibly erroneous signals. The inference made from others’ actions is over what one’s personal preferences are. For example, what one may be learning from observing that no one protests is that they are in a good regime. Within their informal argumentation, however, they often argue that they are capturing pluralistic ignorance and that the misinformation that is being spread is over the content of a social norm. But if people care about social norms because they are concerned about social consequences, then the relevant inference is not about one’s own personal preferences but about the distribution of preferences. The question of how there could be persistent misinformation about the distribution of preferences when this is an inference of interest is therefore different from what is answered by a standard herding setting. As we show, pluralistic ignorance can arise when there is no uncertainty about personal preferences.

Our innovation comes from introducing social expectations into the utility function. That is, individuals care about acting in ways what others want or that they think others think others want, and they infer this through others’ actions. We study a setting in which individuals are choosing one of two actions that affect others. They do so sequentially and for several periods. If all individuals have acted in a way that is consistent with them preferring a certain action, individuals who must decide which action to take will infer that the action they’ve observed is what is socially expected. Note the difference from a standard herding model: individuals may perfectly know what their preferred actions are but be misinformed about what others want them to do. We show that in this setting, individuals
may be misinformed about the true distribution of preferences, which leads them to not protest, which confirms others’ misinformation over the distribution of preferences. This pluralistic ignorance may persist for an arbitrarily large number of periods. As opposed to a standard herding model, this arbitrarily large persistence is possible because individuals cannot experiment to determine what others’ preferences are.

We argue that the equilibrium pluralistic ignorance applies to a variety of settings usually explained through a standard herding or a strategic complementarity story. Consider the following situations: A teacher asks her students if there are any questions, and none of the confused students raise their hands. Gift-exchanges often go on for several years before someone asks around to realize that nobody likes the gift exchange. Wild or inappropriate claims go uncontested at a conversation. Singles who are attracted to each other do not mingle. In these settings, we argue that individuals know what they would personally prefer but do otherwise because they think that is what others expect.

We will show how our model can be empirically distinguished from other models by focusing on Mohammed Bouazizi and the Arab Spring. In a situation of pluralistic ignorance, we ask what characteristics an individual would need to have in order to ‘break’ the equilibrium, in the sense that his one time protest could lead other agents to start protesting. We find that there are two types of leaders that can break pluralistic ignorance. The first is, as previous models have argued, an informed leader who knows what people really want despite their public actions and has incentives to act according to their preferences. It is worth pointing out, however, that this leader is not identical to other models in one important respect: the private information he has is not about a state of the world that affects everyone, such as being the only one to know that changing the regime would be best for everyone, but about the distribution of private preference for the regime, such as knowing that most privately already do prefer to change the regime.

The second type of leader is an uninformed leader whose preferences are ‘representative’ of the population’s preferences and whose actions reflect his true preferences. His preferences are drawn from the distribution of preferences in the population and his actions
are known to not depend on his perception of discontent. This second leader is different from the first, and our novel model of pluralistic ignorance is necessary for capturing this type of leader. In particular, note that we did not specify anything about the revolutionary’s private information. Our revolutionary can be completely ignorant about the extent of regime discontent. What matters is not what he knows, but what individual know about his preferences. If individuals know that his preferences are modal, then through his action others will know that his preferences are more widespread than previously thought. We argue that in the case of Bouazizi, the fact that he is an ‘everyman’ made individuals believe that his preferences were representative of many others.

We argue that our ‘uninformed leader’ captures the stylized facts of Mohammed and the Arab Spring. Mohammed was seen as an ‘every man’ whose grievances and corresponding actions reflected what most Tunisians felt. His self immolation was an act of despair from a man who was deeply fed up with the way the regime had treated him, not a calculated maneuver by a political broker or an extremist dissident. This provided a clear signal of his preferences. In a Times article written shortly after Mohammed’s death, a young man in the city where Mohammed self immolated said ‘We are all Bouazizis if our hopes are dashed.’ A neighbor of Mohammed said ‘We were silent before but Mohammed showed us that we must react.’ (in Ryan, 2011).

Some models of behavioral cascades assume that the first mover or movers gave information about the regime which individuals updated on and which they used in order to act (Lohmann, 1994; Lupia and McCubbins, 1998). This doesn’t fit with the facts about Mohammed, since he was reacting to how the regime treated him. Although something is learned about the regime given its corruption and negligence towards Mohammed, these types of practices were not unusual in Tunisia. It is Mohammed’s deep frustration that was most revealing.

Several models have stressed the importance of a public signal in starting a social movement (Chwe, 2000; Fearon, 2011; Acemoglu and Jackson, 2014). According to these accounts, a revolution is a problem of coordination, so a public signal allows individuals
to act at the same time. However, a public signal cannot be the whole story. There were public signals in Tunisia that did not start a revolution. Further, imagine that a viral Facebook video in which a clearly crazy person is ranting about the regime. Intuitively, it seems highly unlikely that this would spark a revolution. Surely some transmission of information is necessary. The setup of our model abstracts from coordination considerations: individuals move sequentially and observe past actions. We show that absent coordination considerations, individuals may still be affected by information over what others want, but do not need common knowledge. Depending on how social expectations are defined, knowledge or knowledge of others’ knowledge is sufficient. This prediction allows for an experimental test that distinguishes this social expectation motivation from a coordination explanation.

The rest of this article is organized as follows. The next section presents the setup of the model. Section III gives our definition of pluralistic ignorance in the context of the model. Section IV takes an incomplete step at characterizing the equilibrium and showing uniqueness. presents the sketch of the proof of the main results, in which I characterize the equilibrium. Section V shows that as $I$ increases, pluralistic ignorance vanishes. It proposes a behavioral assumption to avoid this conclusion. Section VI discusses the behavioral assumption and compares the preliminary results to other papers.

3.2 Setup

Consider a setting with $I$ players and $T$ periods. Since it will be clear by context, we abuse notation slightly by also referring to $I$ and $T$ as the sets of players and periods, respectively. Players $i \in I$ take turns choosing an action $a \in \{0, 1\}$ for $T$ periods. I will also refer to players as individuals. In each period, player $i$ will be the $i$th player to choose an action. We will refer to the sub-period in which player $i$ plays at period $t \in T$ as the $(i, t)$th sub-period. The actions could be whether to protest, whether to ask a question in class, whether to participate in the gift exchange, and so on. We are interested in characterizing the conditions under which individuals will take actions for several periods. In our simple setup, this
outcome captures the outbreak of a revolution, the level of participation in a class, how long a gift exchange persists, and so on. We denote a history of play up to the action of player $i \in I$ in period $t \leq T$ as $h^{(i,t)} = \{a_{1,1}, a_{2,1}, ..., a_{1,T}, a_{2,T}, ..., a_{i,T}\}$. Individuals can observe the full history at each stage of the game. $H$ denotes the set of possible histories.

Individuals have private information over their types, which determine the benefit they get from the actions. Individuals vary in the personal preference they have over the action. Their personal preference is given by $\theta \in \{0, 1\} \equiv \Theta$. Those with a personal preference for 0 have a personal preference for the action they choose themselves and do not care what others do or expect. Those with a personal preference of 1 have a personal preference for themselves and others choosing 1, and care about what others expect. We will sometimes refer to the latter as the ‘socially considerate’ types, or considerate types. A type 0 will want to choose action 0, independently of what the other players want. A considerate, on the other hand, might prefer to choose 0 if enough others do not want to choose 1. The non-standard term ‘considerate’, as opposed to conformist, is deliberate. Like in models of herding or conditional cooperation, the considerate may act in a certain way if enough others do so as well. But the difference is that the considerate will do so only if he thinks that others are acting in this way because that is accurately reflecting how they want to act or what they expect. If the considerate observed others acting in a certain way but thought that they all wanted or expected that a different action be taken, the considerate might want to act in a way that is different from the others.

There are at least two reasons considerates may care about social expectations. The first is that the desire to do what others want or what others expect will affects the utility function directly, due to altruism or to avoid an internal cost of not following social expectations (Dana, Cane and Dawes, 2006; Fernandez-Duque and Hiscox (2015a, 2015b)). A second reason is reputational, such as when acting in ways that others want but nobody does may position the individual as an innovator or a leader. Another reputational story would be that an individual may think that others are falsely supporting an action as a way to test his type and punish a deviator. Although this reputational story is a fruitful line of thought, we
will not explore it in this simple setup.

A fuller model could incorporate considerate types that are biased towards 1 as well as considerate types who are biased towards 0. However, it is often the case that there is an asymmetry in the actions that make an asymmetry in the preference more natural. For example, when deciding whether to protest, to raise a hand to ask a question in class, or to start a new conversation, the baseline expectations are often to continue the status quo. In this sense, the tension for the considerate types often comes from situations in which they have to decide whether to act against the status quo expectations. Considerate types who personally prefer the status quo have expectations and personal preferences going in the same direction, so this may be thought of as a justification for why all players who prefer 0 will always choose 0. In addition to the asymmetry in expectations, there is often an asymmetry in cost in favor of the status quo action. This cost is not explicitly included in the model but the personal benefit from taking the personally preferred action can be thought of as a net benefit. The considerate type favoring the status quo action would then need the social expectations outweigh the higher cost of taking an alternative action, making it less likely to happen and thus easier for us to worry less about.

Since 0 types always choose 0, our analytical focus will be on the considerates. Although the 0 types are different from the considerates in that they do not care what others do or want, the difference that is conceptually important is that the former do not care what others want. We conjecture that we could modify the assumption that players of type 0 do not care what others do and reach similar conclusions, but that leads to complications that take the focus away from the main results.

Individuals have common priors about the distribution of types $\mu_i(\theta)$, and this is common knowledge. Nature draws types independently given the prior distribution. We write $-i \equiv \Gamma \setminus i$ as the players other than $i$. Individual $i$’s posterior beliefs of $-is'$ types given history $h^{k,t}$ are denoted $\prod_{j \neq i} \mu_i(\theta_j|h^{k,t})$, where $\mu_i(\theta_j|h^{k,t})$ is $i$’s belief at $(k,t)$ that the $j$ is of type $\theta_j$. We also denote the common prior $\mu_i(\theta) \equiv \mu_i(\theta|h^{(1,1)})$. A strategy is a function $s_i: H \times \Theta \rightarrow \{0,1\}$. 
In order to sustain pluralistic ignorance in equilibrium, we will assume that individuals mostly care about social expectations (to be defined below). However, we think of their utility as being a combination of their social expectations and a direct utility from how individuals behave. We will proceed by defining this utility function, and then quickly assume that the dominant term is the one related to social expectations. Even though we will not be working directly with this more complete version of the utility function for most of the paper, it helps illustrate how our model relates to our earlier discussion. In the appendix we provide some conditions for our results to go through with the full utility function.

To define a type 1’s utility at \((i, t)\), first consider his action payoff at \((i, t)\):

\[
E(W_{i,t}(a_{i,t}, a_{-(i,t)}) | h_{i,t})
\]

where \(a_{-(i,t)}\) is the vector of actions taken at sub-periods other than \(-(i, t)\).

That is, an individual \(i\)'s utility at period \(t\) depends in part on his expectation of others’ actions in sub-periods greater than \((i, t)\). We assume that \(W_{i,t}(a_{i,t}, a_{-(i,t)})\) weakly and boundedly increases in \(a_{j,\tau}\) for all sub-periods \((j, \tau)\).

To give an example of \(W_{i,t}(a_{i,t}, a_{-(i,t)})\), individuals may only care about what they do in the current sub-period. In this case, \(W_{i,t}(a_{i,t}, a_{-(i,t)})\) is simply

\[
a_{i,t}
\]

An example of this might be gift giving, in which individuals only care about not giving gifts and do not care about the gifts others give them or others. If all players are affected as long as one person chooses 1, \(W_{i,t}(a_{i,t}, a_{-(i,t)})\) would be:

\[
1 \left( \max_{(j, \tau) \in \{(1,1), \cdots, (I,T)\}} \{a_{j,\tau}\} = 1 \right)
\]

A justification of this payoff would be the setting where a teacher asks if there are any questions and a single response would be enough to get the teacher to explain.

In many settings such as protests or funding for a project, an individual may care about
what others will do. For example, an individuals may instead care about players’ actions reaching a certain threshold for a certain period. For example, the success of a protest may depend on enough people protesting in a given period. An example of $W_{i,t}(a_{i,t},a_{-i(t)})$ in this case is

$$1 \left( \sum_{(j,t) \in \{(1,t),..., (I,t)\}} a_{j,t} > x \right)$$

(3.3)

where $x \in [I/2, I]$.

In settings where ‘every little bit helps’ the action payoff may also depend more linearly on actions, such as not throwing trash on the street or charitable giving. $W_{i,t}(a_{i,t},a_{-i(t)})$ in this setting could be:

$$\sum_{(j,t) \in \{(i,t),..., (I,T)\}} a_{j,t}$$

(3.4)

Let $U_{i,t}(s_i,s_{-i}|\theta_i = 1, \mu) = \mathbb{E}(W_{i,t}(a_{i,t},a_{-i(t)}))|h^{t,i}) + \alpha S_{i,t}$ be the expected utility of player $i$ of type 1 at $t$. All players are risk neutral, and there is no discounting as it will not play an important role in the model.

$S_{i,t}$ is player $i$’s social expectations at $t$. Although we will modify the definition below in section V, here social expectations will be a form of altruism, in which individuals receive utility for matching the period’s actions to others’ preferences. In particular, we provide the following definition of $S_{i,t}$:

$$\sum_{j \in I \setminus i} \mathbb{E}(1(a_{i,t} = \theta_j)|h^{t,i})$$

Note that a considerate player $i$ only directly cares about the other players’ personal preference part of others’ utility function, and not about others’ social expectations part of their utility.

For most of the paper, we will assume that $\alpha$ is high enough that the social expectations term dominates individuals’ decisions. For simplicity, therefore, we will work with the simplified utility function $U'_{i,t}(s_i,s_{-i}|\theta_i = 1, \mu) = S_{i,t}$. Aside from being easier to work with, there are two other motivations for this simplification. The first is that if the action payoff term $W$ dominated, individuals would act according to their personal preferences. If that were the case, their actions would reveal their types and pluralistic ignorance as we define
below would not arise. Having social expectations be important ‘enough’ is necessary for our results, and $U_{i,t}'$ is simply a limiting version of that. Social expectations are arguably highest in the gift giving and classroom examples, corresponding to action payoffs (3.1) and (3.2), which are settings where individuals know the others well and care about their own reputation and perhaps about others’ welfare. The second motivation is that, for some such as (3.3) and (3.4), the action payoffs depend negligibly on player $i$’s actions at $t$ for a large $I$. We further argue in the appendix that this negligibility extends to $i$’s influence on others’ actions.

This notion of social expectations is similar to altruism: individuals care about doing what others want, and the more others want something the more incentive there is for the individual to do it. However, an important thing to note about the utility function is that individuals get utility from taking an action at period $t$ given their social expectations at period $t$ and not from other periods. That is, if individuals update their beliefs over others’ types after period $t$, that does not affect the stage game utilities $U_{i,t}$. The intuition for this is that the social expectations term affects the decision at the time it is being made, like when individuals avoid protesting because they want to avoid the short term cost of upsetting others. This is in contrast to individuals deciding whether to protest or not because they want to contribute to bringing about the state of the world that will make others best off. In this case, what may matter to them is how they acted throughout the $T$ periods given their best information about others’ types, which would come at the end of the game. We assume that this stronger form of altruism is not what is driving behavior.

Relatedly, notice that players of type 1 do not care about their own future social expectation terms. That means that individuals will not take into account how present actions affect their social expectation in the future. This is partly for simplicity, but there are three other justifications. The first is, as we discuss in the appendix, if the impact of an individual’s actions on others’ actions is negligible, social expectations will be the same in all future periods. The second is the evidence that individuals underestimate their others’ and their own future embarrassment (Van Boven, Loewenstein and Dunning, 2004 and 2005), which
we interpret as a cost due to social expectations. The third, related, is that current social expectations may be much more salient to individuals than future social expectations, which are therefore heavily discounted.

In contrast, the utility of players of type 0 at \((i, t)\) is simply:

\[
U_{i,t}(s_i, s_{-i}|\theta_i = 0, \mu) = 1(a_{j,t} = 0)
\]

where \(1(\cdot)\) is an indicator function.

There are three features of the setup that are non-standard compared to other models of protests. The first is that, unlike standard herding models, there is no underlying state of the world over which individuals have uncertainty. It is typical in behavioral cascade models (Banerjee, 1992; Ali and Kartik, 2012) and policy choice models (Coate and Morris, 1995; Canes Wrone, Herron and Shotts, 2001; Maskin and Tirole, 2004; Bicchieri, 2005) for agents to have uncertainty over the impact of the state of the world on their utility. For example, individuals have uncertainty over whether they personally are best off with a regime, and the decision over whether or not to protest depends on what they learn about how the regime benefits them. In our setup in contrast, individuals know whether they personally favor the regime, and what they have uncertainty over is whether others favor the regime. The two approaches can be reconciled technically, as long as the state of the world is defined as the distribution of preferences for the regime. In fact, in section V we will combine the current approach with a standard herding approach. This section shows, however, that it is not necessary for pluralistic ignorance to be sustained.

We believe our approach captures a more natural source of uncertainty for individuals to have regarding the overall impact of a regime, and one which individuals care about. It is dubious that individuals were unaware that they were unhappy in the East German regime, despite government propaganda. Certainly, when talking about pluralistic ignorance, our assumption is the one that fits most naturally in the examples above (which are classic examples in the pluralistic ignorance literature) or verbal definitions.

The second non-standard feature is, as we’ve discussed, preference complementarities.
Although this is not common in the literature on behavioral cascades, it is an example of type-dependent preferences (Levine, 1998; Sliwka, 2008; Ellingsen and Johannesson, 2008; Gul and Pesendorfer, 2010). Type dependent preferences has been used as a tractable way to model agents acting according to what others’ expect of them (Levine, 1998; Cox, Friedman and Gjerstad, 2007) as an alternative to the approach taken by psychological game theory (Geanakoplos, Staccheti and Pearce, 1989; Rabin, 1993; Taubinsky, 2012). In these models, individuals prefer to act differently according to other individuals’ unobservable type, which may reflect their preferences or their expectations over others’ actions. For example, a dissenter can be thought of as disliking the regime or as expecting others to protest about the regime. In our setup, the use of type dependent preferences allows us to look for an equilibrium where an individual’s beliefs over what others expect her to do does not need to match what individuals actually believe. This concept is the basis of pluralistic ignorance, and cannot be captured by the equilibrium requirement of psychological game theory that higher order beliefs need to coincide with beliefs.

The third non-standard feature of this model is that if we took the full utility function \( U_{i,t} \) and \( \alpha = 0 \), individuals of type 1 would choose 1 despite what others did. This assumption is not standard in the protesting example, but is natural in the example of asking a question in class. The assumption simply restates the fact that the model does not depend on strategic complementarities: 0 types will always choose their preferred action, while considerate types may take an action if they believe it is sufficiently likely that others want that action. In some sense, this is certainly a simplification: there are several intuitive reasons that public good actions depend on how many others protest. The simplification allows us to highlight the mechanism in the starkest way possible, and at the same time will make it clear how it differs from strategic complementarities models.

A more standard version of this model can be developed. In this more standard version, there is a state of the world that determines whether choosing action 0 is worthwhile or not. Individuals receive a private signal \( \theta \in \{0, 1\} \) over the state of the world. Two interpretations of the signal are possible. In the first, the strength of the signal varies. For players of type
0, it is always strong. For players of type 1, it is either strong or weak. According to this interpretation, the individuals with a strong signal are sufficiently confident in their signal that they don’t require information from others to choose how to act. The individuals with a weak signal cares about others’ signal. A second interpretation is that signal strength is the same for all players. Players of type 0 are all naïve, and take the signal as being completely accurate. Players of type 1 are either naïve or Bayesian. Note that in both cases, the impact of others’ actions would no longer affect an individual through the integration of others’ preferences, but through what their actions inform the individual about the state of the world. We will discuss why these versions do not fit the facts of how the Arab Spring began.

Denote by \( G = \{ \mu(\theta|h^{1,1}); I, T \} \) the dynamic game defined above with prior \( \mu(\theta|h^{1,1}) \), \( I \) individuals and \( T \) periods. We will call the tuples of strategies \( (s_1, ..., s_I) \) and posterior beliefs \( (\mu_1, ..., \mu_I) \) a profile \( \rho = \{s_1, ..., s_I, \mu_1, ..., \mu_I\} \). We will be looking for a profile \( \rho = \{s_1^*, ..., s_I^*, \mu_1, ..., \mu_I\} \) that constitutes a Perfect Bayesian equilibrium of the game \( G \).

### 3.3 Defining Pluralistic Ignorance

We are interested in defining what we mean by pluralistic ignorance in our setup. We restrict attention to the simplified utility function \( U'_{i,t} \) in this section. We discuss the case with the full utility function \( U_{i,t} \) in the appendix. Notice that a player \( i \) of type 1 will choose 0 at sub-period \((i,t)\) if and only if:

\[
\sum_{j \in I \setminus i} \mathbb{E}(1(0 = \theta_j)|h^{i,t}) \equiv S^0_{i,t} > \frac{1}{2} \quad (3.5)
\]

This inequality follows from comparing the utility of player \( i \) at \( t \) of choosing 1 versus choosing 0. As mentioned earlier, only the present sub-period’s social expectations matter, so individuals have no incentives to think strategically about how their actions will affect future social expectations.

Let a draw function of game \( G \), \( d : G \rightarrow \{ \text{combinations of } I \text{ players of types } \{0,1\} \} \), be the random variable that draws \( I \) players’ types according to the priors given by game \( G \).
Denote a fixed draw of game $G$ as $\hat{d}(G)$, a realization of the random variable. We can now state the following:

**Definition 4.** We say there is pluralistic ignorance in $\hat{d}(G)$ starting at $(i, t)$ when, for all $(j, \tau) \geq (i, t)$ in an equilibrium of $G$,

$$
\sum_{k \in I \setminus i} \mathbb{E}(1(0 = \theta_k)|h^{k, \tau}) > \frac{1}{2} \quad \text{and} \quad \sum_{k \in I \setminus j} \mathbb{E}(1(0 = \theta_j)|\hat{d}(G)) > \frac{1}{2}
$$

That is, for the draw $\hat{d}(G)$ the players of type 1 choose 0 for all periods starting at $(i, t)$, but would choose 1 if they knew the true distribution of types given by $\hat{d}(G)$.

The intuition behind the definition is that pluralistic ignorance is a situation where individuals are acting in a certain way out of consideration for social expectations, but if they knew the true distribution of what people wanted, they would act according to their personal preferences.

### 3.4 Pluralistic Ignorance in Equilibrium

This section restricts attention to the simplified utility function $U'_{i,t}$. We discuss the case with the full utility function $U_{i,t}$ in the appendix. Notice that players affect others’ actions only through the information they reveal about their types through their own actions. Indeed, this is straightforward since at sub-period $(i, t)$ individual $i$ chooses the action that maximizes $S_{i,t}$, taking beliefs over types as given. A player’s equilibrium strategy is therefore to follow the decision rule given by (3.5) at each sub-period where they make a choice.

Because individuals do not take into account how their actions affect future actions, we can begin to derive the equilibrium simply by looking at what (3.5) tells the first player to do at $(1, 1)$. This is well defined except in the degenerate case where the priors make him exactly indifferent between choosing 1 and 0 (that is, when $\mu(\theta = 0) = 1/2$). Since individuals are indifferent as to how their actions affect future payoffs, we assume that an indifferent individual chooses 1. This is the conservative assumption in terms of avoiding
pluralistic ignorance.

The first player’s well-defined equilibrium action at \((1,1)\) together with the priors determines what the second player’s beliefs over the distribution of types for any action he observes on the equilibrium path. The only off-equilibrium action would be for player 1 to choose 1 at \((1,1)\) given that in equilibrium both players would have chosen 0. But because only players of type 1 choose 1 for some beliefs over types, we assume that individuals assign probability 1 to a player being of type 1 if they make an off-equilibrium choice in which they choose 1 at any sub-period. Therefore player 2’s beliefs over types at sub-period \((2,1)\) are well defined for any action player 1 takes, and each action yields a unique decision for player 2 by (3.5). But this same reasoning can be reapplied to player 3 at period 1, and so on, to conclude that the equilibrium will be unique.

If player \(i\) at \(t\) believes other players are more likely to be of type 0 than of type 1, he will choose 0. But if a player of type 1 chooses 0 in equilibrium, the information about players’ types will remain the same after sub-period \((i,t)\). Therefore, the next player to make a decision will also believe other players are more likely to be of type 0 than of type 1, and will choose 0. Extending this logic, if priors assign the probability of being drawn as a type 0 to be higher than the probability of being drawn as a type 1 as of period \((i,t)\), all future players will choose 0. Therefore, if the draw of types is such that more than half of the players are of type 1 but priors are such that less than half are of type 1, pluralistic ignorance will be sustained from the first period.

If player \(i\) of type 1 at period \(t\) believes the other players are more likely to be of type 1 than of type 0, he will choose 1. This signal will weakly increase the proportion of players that are believed to be of type 1. So if \(i\) chooses 1 at \(t\), \(i+1\) of type 1 at \(t\) chooses 1. By choosing 0, \(i\) at \(t\) would send a signal that weakly increases the proportion of players that are believed to be of type 0. Consider a draw of types \(d(G)\) such that in an equilibrium history \(h_{r^{*}l,T}\) where all players have acted according to their personal preferences, at all sub-periods it is more likely for a randomly chosen player to be of type 1. Then all type 1 players will choose 1. However, if the draw of types is such that at some point in the sequence it is more
likely for a randomly chosen player to be of type 0, then from that point on all players will choose 0, as per the logic of the last paragraph.

Putting the logic of the last two paragraphs together, it is easy to see that type 1 players will either choose 1 until some sub-period \((i, t)\), possibly \((1, 1)\), or choose 1 for all sub-periods. In the former case, all players choose 0 from \((i, t)\) on. If at sub-period \((1, 2)\) - the first sub-period of period 2 - a player of type 1 would choose 1, then players of type 1 will choose 1 for all future sub-periods. That is because by that sub-period all players have either revealed their types or are all choosing 0. Sequences where more players of type 0 are drawn in earlier periods make it more likely for players of type 1 to choose 0.

Notice the asymmetry between the conditions under which choosing 0 and choosing 1 is stable for type 1 players. All that is needed to determine whether all players after \((i, t)\) will choose 0 is given by the information over types revealed by sub-period \((i, t)\). In contrast, to determine whether all type 1 players after \((i, t)\) will choose 1, we need to make sure that in no future sub-period of period 1, \(s^0_{i,t} > 1/2\).

We can now state the following:

**Result 5.** Pluralistic ignorance can be sustained starting at any sub-period \(h^{i,t}\) such that individuals believe the proportion of players of type 0 is greater than 1/2 but the proportion that was drawn is less than 1/2.

It is worth stressing that we reached pluralistic ignorance even if there was no uncertainty over the state of the world in the standard sense of a herding model.

In a standard herding model individuals do not know their personal preference. By experimenting, they can discover what they like most. In our setting individuals know their personal preference, and what they’re uncertain is about others’ types. When others players choose 0, a player won’t learn anything about the distribution of types. The only way to ‘experiment’ is to influence others into revealing their type through their own revelation of types. But if that is not sufficient motivation, then players will follow their social expectations and misinformation can last an indefinite number of periods. In a herding model, since individuals are just learning about their own preferences, with enough periods they can
figure out what their optimal choice is so herding eventually vanishes. We highlight this result in the following:

**Observation 1.** Pluralistic ignorance can last an arbitrarily large (but finite) number of periods.

Notice that for pluralistic ignorance to work in this setting it must be the case that the draw of types is unexpected, in the sense that the proportion of type 0 players is higher than the priors. This leads to the following:

**Corollary 1.** The probability of pluralistic ignorance becomes arbitrarily small as $I$ becomes arbitrarily large.

This is a direct consequence of the law of large numbers. The probability that the proportion of type 1 players is different from the prior probability that a player is of type 1 becomes arbitrarily close to zero as the number of players increases. □

Since we are interested in studying pluralistic ignorance in large groups, this last result is therefore unsatisfactory. We therefore expand our setup in the next section.

### 3.5 Pluralistic Ignorance With Noisy Signals Or Herding on Social Expectations

Like the last section, this section restricts attention to the simplified utility function $U'_{i,t}$. We discuss the case with the full utility function $U_{i,t}$ in the appendix. In order to get our desired result with an arbitrarily large $I$, we will mimic the standard approach to getting herd behavior. Notice that in our setting, individuals will have uncertainty over the distribution of types. This is a deviation from the standard assumption of common priors. Nevertheless, we will retain a related assumption of common meta-priors.

There are two states of the world, $\omega \in \{H, L\}$. Individuals’ priors are such that either state of the world is equally as likely to be drawn by nature. The difference between the states of the world is the prior distribution of types, with:

$$
\mu(\theta = 0|\omega = L) = \mu(\theta = 1|\omega = H) > \frac{1}{2} > \mu(\theta = 1|\omega = L) = \mu(\theta = 0|\omega = H)
$$

84
Players get a noisy signal over the prior probability of types, \( \phi \in \{H, L\} \). Signals are informative of the state of the world: \( \mu(\omega = H | \phi = H) = \mu(\omega = L | \phi = L) > 1/2 \).

Now that there is a difference between private and public knowledge, the question of what social expectations are naturally arises. One way to define \( S_{i,t}(\phi_i) \) is from the perspective of the player’s private information:

\[
S_{alt}^{i,t}(\phi_i) \equiv \sum_{j \in I \setminus i} \mathbb{E} \left( 1 (a_{i,t} = \theta_j) | h_{i,t}, \phi_i \right)
\]

This term captures how many players \( i \) thinks have a personal preference for the action \( i \) takes at \( t \). As we noted previously, putting this into the utility function is similar to altruism.

A second way of defining \( S_{i,t}(\phi_i) \), which would have been equivalent when there was no private information, is as what \( i \) believes others believe is the distribution of preferences:

\[
S_{se}^{i,t}(\phi_i) \equiv \mathbb{E}_i \left( \sum_{k \in I \setminus j} \mathbb{E}_j \left( 1 (a_{i,t} = \theta_k) | h_{i,t} \right) \right) | \phi_i
\]

where the expectation operator’s subscripts denotes from who’s perspective the expectation is being calculated. This way of thinking about \( S_{i,t} \) makes the difference between altruism and social expectations more stark. Individuals care about doing what they think others think is socially desirable. This is more in line with the second order belief definitions of social norms (e.g. Elster 1989a, 1989b, Coleman, 1990, Fehr and Gachter, 2000). We will consider a generalization of these functions. For \( \gamma \in [0, 1] \),

\[
S_{i,t}(\phi_i, \gamma) \equiv \gamma S_{alt}^{i,t}(\phi_i) + (1 - \gamma) S_{se}^{i,t}(\phi_i)
\]

The results we present in this section are valid for any value of \( \gamma \in [0, 1] \). In the next sections we will present results which will depend on \( \gamma \).

A dynamic extended game is now defined by \( G = \{ \mu(\theta | h^{1:1}), \mu(\theta = 1 | \omega = H); I, T, \gamma \} \). We continue to look for a Perfect Bayesian Equilibrium.

Choosing 1 may now reveal \( i \) is of type 1 with signal \( H, L \) or either, of type 0 with signal \( H, L \) or or either, or a combination of those. Thus, social expectations are updated not only by what actions signal about individual’s personal preference, but also about what is signals
about their private information.

Proving the existence of pluralistic ignorance in this setting closely parallels the proof in the simpler setting. Since individuals don’t take into account how their actions affect their future payoffs, we can just solve by looking at what player 1 does in period 1, how that affects player 2’s information over types, and so on. As before, we assume that if a player is indifferent between choosing 0 and 1, the player chooses 1. We also repeat the assumption that players who choose an out-of-equilibrium action are of type 1, since choosing 0 for all periods is always on the equilibrium path for type 0 players. A player’s equilibrium strategy will again be to follow the (suitably modified) decision rule given by (3.5) at each sub-period where they make a choice. We’ll keep the definition of pluralistic ignorance the same for simplicity (one could extend it to pluralistic ignorance for players who received a certain signal).

Notice that $S^0_{i,t}(L, \gamma) > S^0_{i,t}(H, \gamma)$ for all $\gamma \in [0, 1]$. That means that in any given sub-period, the player’s types’ optimal decision will lead to one of three situations:

- All types with any signal choose 0
- Only player of type 1 with signal $H$ to choose 1
- Players of type 1 choose 1

Suppose first that player $i$ at $t$ has not revealed anything about his type.

If all types with any signal optimally choose 0 at $(i, t)$, then $i + 1$ will have the same information about types at $t$ as $i$ did. Therefore, all types with any signal optimally choose 0 at $i + 1$. This would sustain pluralistic ignorance if the draw of types has more than half type 1 players.

If players of type 1 optimally choose 1 at $(i, t)$, then if $i$ chose 1 (0) at $t$ it raises (lowers) $i + 1$ of type 1’s posteriors that most players are of type 1. Player $i + 1$ of type 1 at $t$ will therefore choose 1 if $i$ chose 1 at $t$. If $i$ chose 0 at $t$, then the choice of 0 weakly increases for $i + 1$ of type 1. If this leads $i + 1$ of type 1 to choose 0 whatever their private signal, then players will choose 0 in all future sub-periods per the logic of the last paragraph. However,
this may also lead \(i + 1\) of type 1 to choose 0 if he received signal \(L\), and choose 1 if he received signal \(H\).

If at \((i, t)\) only player 1 of type 1 with signal \(H\) chooses 1, then if \(i\) chose 1 at \(t\) it raises \(i + 1\) of type 1’s posteriors that most players are of type 1 both because \(i\) is revealing to be of type 1 and because he is revealing to have signal \(H\). If \(i\) chooses 0 at \(t\), player \(i\) is revealing to not be of type 1 with signal \(H\), which makes him more likely to be of type 0 and to have signal \(L\). Therefore, \(i + 1\) at \(t\) of type 1 is weakly more likely to choose 1 after observing 0.

There are two cases left to consider. The first is when individual \(i\) has already revealed to be of type 1, but has not revealed his signal. The second is when the individual \(i\) has revealed to not be of type 1 with signal \(H\). A similar logic to the one developed in the last three paragraphs regarding what they signal can be easily extended.

Putting this together, it is easy to see that players of type 1 with signal \(H\) will either choose 1 until some sub-period \((i, t)\), possibly \((1, 1)\), or choose 1 for all sub-periods. In the former case, all players choose 0 from \((i, t)\) on. Suppose that at sub-period \((1, 2)\) there is some player of type 1 who would choose 1. Then there will be four categories of players, depending on the information they have revealed about themselves. There will be (a) players who have revealed to be of type 0, (b) players who have revealed to be of type 1, (c) players who have revealed to be of type 1 with signal \(H\) and (d) players who have revealed to not be of type 1 with signal \(H\). The only learning about private information that is left is to learn the private signal of players in category (b), or learn the type of players in category (d).

If there are players of type 1 who would choose 1 given the public information about types, there is some finite sub-period after which players will make the same choice for all future sub-periods. Indeed, if all players of type 1 choose 1 at some sub-period, the uncertainty over the private signal of players in category (b) will not be resolved, since they would choose 1 independent of their signal. The uncertainty over the types of players in category (d) will be resolved, as type 1 players would choose 1. As the uncertainty is resolved, players of type 1’s optimal choice will either remain at 1 or change to 0. So if at sub-period \((\hat{i}, \hat{t})\) all players of type 1 choose 1 and the uncertainty over players in
category (d) is not enough to change players’ optimal choice, all players will choose 1 for all future periods starting at \((i', t')\). If only players of type 1 with signal \(H\) choose 1 at some sub-period, the uncertainty over the types of players in category (d) will not be resolved, since they would choose 0 independent of their type. The uncertainty over the private signal of players in category (b) will be resolved, however. So if at sub-period \((i'', t'')\) only players of type 1 with signal \(H\) choose 1 and the uncertainty over players in category (b) is not enough to change players’ optimal choice, only players of type 1 with signal \(H\) will choose 1 for all periods starting at \((i'', t'')\).

To find a situation where pluralistic ignorance can be reached with an arbitrarily large population, we further specify our assumptions. In particular, we assume that in sub-period \((1, 1)\) priors are such that player 1 of type 1 chooses 1 if and only if he has signal \(H\).

Now consider the updated beliefs of a player \(i\) of type 1 with signal \(H\) who at period 1 observes a string of \(n\) players who choose 0. Further assume that no player of type 1 with signal \(H\) would have chosen 0 at any point of the history of the \(n\) players choosing 0. The player’s updated belief about the state of the world is:

\[
\mu(\omega = H|\theta_i = 1, \phi_i = H, h^{n-1}) = \frac{\mu(1H|\omega = H)[\mu(1L|\omega = H) + \mu(\theta = 0|\omega = H)]^n}{\mu(1H|H)[\mu(1L|H) + \mu(\theta = 0|H)]^n + \mu(1H|L)[\mu(1L|L) + \mu(\theta = 0|L)]^n} \tag{3.6}
\]

To get an intuition for this formula, note that if there were no players of type 0, and using the symmetry of states of the world and signals, it would simplify to:

\[
\frac{[\mu(L|\omega = H)]^{n-1}}{[\mu(L|H)]^{t-1} + [\mu(L|L)]^{n-1}}
\]

which is equal to 1/2 for \(n = 1\) and less than 1/2 for \(n > 1\). This is analogous to the inference problem in a standard herding setting, where individuals know that everyone’s preference is the same and the inference is over the prevalence of a signal. In our case, when \(i\) observes 0, he cannot distinguish whether it is a player of type 1 with a signal \(L\) or a player of type 0 with either signal. Since we’ve assumed that signals are independent across types, that means that a player of type 0 may have received signal \(H\). Therefore, in order for a player with a signal \(H\) to believe it is just as likely for the state of the world to be \(H\) or \(L\), it
requires more than one player choose 0. Notice further that when $\mu(\theta=0|\omega=H)=0$ in our setting, the posteriors are lower than in the standard herding setting.

More generally, in equation (3.6), since we know by symmetry that:

$$
\mu(1L|\omega=H) + \mu(\theta = 0|\omega=H) = \\
\mu(1L|\omega=H) + \mu(\theta = 0L|\omega=H) + \mu(\theta = 0H|\omega=H) < \\
\mu(0H|\omega=L) + \mu(\theta = 0L|\omega=L) + \mu(\theta = 1L|\omega=L) = \\
\mu(1L|\omega=L) + \mu(\theta = 0|\omega=L)
$$

so there is some finite $n$ such that a player of type $H$ at believes it is more likely that the state of the world is $L$. Further, this $n$ is not determined by the total number of players $I$ as we illustrated by example. We then know that a player of type $H$ would choose 0 at sub-period $(n+1,1)$, and by arguments similar to before all future players would choose 0. We can therefore summarize as follows:

**Result 6.** In the version of the model with noisy signals described above, there are parameter values such that pluralistic ignorance can be sustained for all periods for any value of $I$ and arbitrarily large (and finite) $T$.

Having established the existence of pluralistic ignorance, even for large $I$, we now turn to comparative statics exercises to show how this model differs from other models.

### 3.6 Starting a Movement

The past sections set up a model and showed that it is possible for pluralistic ignorance to be sustained in equilibrium. The objective of this was to show how this model makes novel predictions about how movements are started.

In this section, we present a logic of how movements get started that builds on the model of pluralistic ignorance we’ve developed. We will first focus on what characteristics an individual needs to have so that others follow his actions. Then we will focus on what motivates an individual to be the first to act differently than the action sustained in pluralistic ignorance. We will contrast this with other models of social movements.
3.6.1 Why people follow

Imagine that players are in pluralistic ignorance. Suppose an individual takes a single action at some sub-period \((i, t)\). What characteristics are sufficient for an individual to have so that, if the characteristics are public information, all type 1 players who observe him choosing 1 know most players are of type 1? We will refer to this individual as a ‘leader’ and his action as ‘breaking pluralistic ignorance’. The answers we provide are those of leaders whose actions increase others’ beliefs that a randomly chosen player is of type 1.

**Result 7.** The following individuals may break pluralistic ignorance by choosing 1:

- **Informed leader:** An individual who knew what the true distribution of types was and who would only choose 1 if his fixed information action was to choose 1.
- **Uninformed representative leader:** An individual who chose according only to his personal preferences and whose preferences was a draw from the same distribution of types.
- **Uninformed modal leader:** An individual who chose according only to his personal preferences and whose preferences reflect the mode of the distribution of types.

Further, it only takes one uninformed modal leader to change an individual’s beliefs that the majority of players are of type 1, while this is not generally true for uninformed representative leaders.

By choosing 1, an informed leader perfectly signals that most players are of type 1. The informed leader is a limiting case of a type of leader has appeared several times in the literature (e.g. Hermalin, 1998; Lupia and McCubbins, 1998), where the leader has privileged private information about the state of the world and and has aligned preferences to those of his audience.

The uninformed leader is novel. By ignoring social expectations, the uninformed leader is providing information about his type that would otherwise not be observed in pluralistic ignorance. For this to increase the chances that a randomly chosen player is of type 1, it has to be the case that the uninformed leader’s preferences are linked to the distribution of preferences in the population. This is achieved if the player’s preferences are drawn from
the prior distribution of types. Another way to link the uninformed leader’s preferences to the distribution of preferences in the population is to have the player’s preferences be a draw from the mode of the distribution.

If individuals know that the uninformed leader’s preferences are modal, then they know that the majority of players are of type 1. However, if they know that his preferences are a random draw from the population, then his action will only make individuals believe that the majority of players are of type 1 if they already thought the probability of a randomly chosen player being of type 1 was very similar to the probability of a randomly chosen player being of type 0. Otherwise, it would take more than one uninformed representative leaders to make individuals believe the majority of players are of type 1. □

Notice that in a standard herding model, all individuals have the same preferences and imperfect information about those preferences. Knowing whether an uninformed individual has modal or representative preferences would not make a difference.

We want to know what individuals would need to know about an informed or uninformed leader in order for his actions to have a higher impact on social expectations than the actions of a player who is not a leader. By impact, here we refer to the amount of individuals who choose 1 after observing an individual choosing 1 minus the amount of leaders who choose 1 after observing the individual choose 0.

**Result 8.** The knowledge that affects whether a leader’s actions have a higher impact on social expectations than would the actions of a non-leader depends on $\gamma$.

- If $\gamma < 1$, then knowing that the individual is a leader and that he chose 1 increases the leader’s impact.

- If $\gamma > 0$, then knowing that others know that the individual is a leader and that he chose 1 increases the leader’s impact.

- If $\gamma = 0$, then only knowing that others know that the individual is a leader and that he chose 1 does not increase the leader’s impact.
• If $\gamma = 1$, then only knowing that the individual is a leader and that he chose 1 does not increases the leader’s impact.

If $\gamma < 1$, then the utility $i$ gains from social expectations increase if $i$’s action matches what others’ personal preferences are. If $i$ receives information that increases his belief that most others are of type 1, this weakly increases $i$’s choice of 1. This is not true if $\gamma = 1$.

If $\gamma = 1$ and the amount of players $I$ is large, $i$’s beliefs over the distribution of personal preferences has a negligible effect on the utility $i$ gains from social expectations. If $\gamma > 0$, $i$’s beliefs over others’ beliefs of the distribution of personal preferences matters to $i$. If $i$ receives information that increases his belief that others believe most players are of type 1, this weakly increases $i$’s choice of 1. Notice that even if $i$’s belief over the distribution of types did not change, it will still affect $i$’s actions. That is, if $i$ believed others were increasing their beliefs that most players are of type 1 based on a message that was uninformative from $i$’s perspective, this would still weakly increase $i$’s choice of 1. This is not true if $\gamma = 0$. □

These knowledge requirements are weaker than what is required when individual’s preferences are given by strategic complementarities and individuals who make a simultaneous decision need to coordinate on an outcome (e.g. Chwe, 2000, Fearon, 2011, Acemoglu and Jackson, 2014). In order to coordinate, it is not sufficient for individuals to know that others know someone is sending a signal to choose a different action, but individuals must know that others know they know this, and so on. Coordination stories also fail to make a prediction as to what is the public information that is needed for individuals to coordinate on.

In other settings with strategic complementarities, individuals move sequentially, and a movement is started if there is a person who moves first and will take a stand independently of what another does, then there is a person who moves second and will take a stand if at least one other person has done so, and so on. Our approach can be thought of as providing microfoundations for a strategic complementarities approach to starting movements. However, the strategic complementarities predicts that the first mover can be anybody willing to start a movement independently of what others do. Typically, this first
mover will be the one with the most extremist views (e.g. Kuran, 1997), but it could just as
well be someone who made a mistake. Our model limits the type of individuals whose first
action can start a social movement.

It may be argued that in order to understand pluralistic ignorance, a more standard
herding story could be developed. A standard herding story in which the uncertainty is
over an individual’s personal preferences would be that individuals are uncertain about
how much their regime benefits them personally, and incorrectly believe it is the best regime
for them. It seems implausible that most people did not know the regime did not benefit
them.

3.6.2 Why people lead

We now consider the following complementary question: under what conditions will an
individual of type 1 in pluralistic ignorance choose 1 before anyone else has done so?

For this sub-section, I will consider the full utility function for type 1 players: \( U_{i,t} = \)
\( E(W(a_{i,t}, a-(i,t))|h^{i,t}) + \alpha S_{i,t} \). Recall that the action payoff is increasing in
\( a_{i,t} \) and in \( a-(i,t) \). With an abuse of notation, further define \( \partial a_{-(i,t)} / \partial a_{i,t} \) as the influence factor of \( i \) at \( t \), or how
future actions are affected by \( i \)'s action at \( t \).

To answer our opening question, consider the following:

**Result 9.** An individual of type 1 is more willing to be the first to choose 1 when there is pluralistic
ignorance:

- If \( \frac{\partial E(W(a_{i,t}, a-(i,t))|h^{i,t})}{\partial a_{i,t}} > 0 \),
  - The less he cares about social expectations relative to the action payoff (\( \alpha \) small)
  - If \( \gamma > 0 \), the more he knows about the true distribution of preferences or the more
uncertainty he has about social expectations
  - If \( \gamma < 1 \), the more he believes others believe most players are of type 1 or the more
uncertain he is about what others believe is the true distribution of preferences
• The higher his influence factor \( \frac{\partial a_{-(i,t)}}{\partial a_{i,t}} \)

Most of these results are straightforward from inspection of the full utility function. Notice that both knowledge (or knowledge of what others know) of the true state of the world and uncertainty (or uncertainty of what others know) of the true state of the world both lead an individual to lead. That is because in a state of pluralistic ignorance, \( S_{i,t}^0 > 1/2 \), and therefore either knowledge and uncertainty bring \( S_{i,t}^0 \) closer to 0. If \( S_{i,t}^0 \) is close enough to 1/2, individuals of type 1 will choose 1 as their action payoff biases them toward that action. □

The reason we need to specify that \( \frac{\partial E(W(a_{i,t}, a_{-(i,t)})|h_{i,t})}{\partial a_{i,t}} > 0 \) in the result is that the influence factor may be negative, and may counteract the increase in the action payoff from \( i \) choosing 1 at \( t \). For example, if individuals are in pluralistic ignorance and the action payoff is (3.2), then the influence factor may be negative, since individuals will not get a positive action payoff from choosing 1 after \( (i,t) \), but may still believe most players are of type 0. We will talk more about the positive influence factor in the appendix.

Some of these results can be found with more standard models. The influence factor is a feature of any model with strategic complementarities where forward looking individuals cared about others’ actions, or models where there is a public good that requires collective action. Caring about the true distribution of preferences is a feature of any model where an individual cares about the effect of his actions on others. More novel are the other motivations for breaking pluralistic ignorance.

If an individual is in a situation where social expectations matter less - a situation in which actions will not be seen by others, or after receiving information that makes it harder to determine what social expectations are - the theory predicts that the individual will be more willing to choose 1. Another novel prediction (if \( \gamma < 1 \)) is that individuals care about what they think others think is the most preferred action, even if they think the reason others think that is based on misleading information and there is no common knowledge about what others think is the most preferred action. In a strategic complementarity model, individual \( i \) would not only need to know what \( j \in -i \) believes \(-j\)'s preferences are, but also
what $j$ believes under what conditions $k \in -j$ will take an action given those preferences, what $j$ believes $k$ knows about $-k$'s preferences, and so on. Further, our setup predicts that if $i$ thinks $j \in -i$ prefers an action be taken or $i$ think $j \in -i$ thinks $-j$ prefer an action to be taken, $i$ is more likely to take that action independently of whether $-i$ do it.

Our setup can be combined with strategic complementarity. Note that if the action payoff is (3.3), there is a strategic complementarity component to the utility function. Although social expectations matter, an individual is encouraged to protest if he knows $x$ of the individuals including himself will also protest (that is, if $i$ is pivotal with respect to the action payoff).

### 3.7 Pluralistic Ignorance and the Arab Spring

In this section we argue that the uninformed leader is a better description of Bouazizi’s impact on the Arab Spring than other competing theories.

First mover stories often rely on a strategic complementarity motivations. Individuals can be ordered according to how many others they need to protest in order for them to want to protest. In some accounts (e.g. Granovetter, 1978; Kuran, 1997), the need of others to protest grows gradually, so some protesters do so whatever others do and others need many to protest before doing so. Willingness to protest can be revealed sequentially so a protest grows if there are enough protesters at a given period that more individuals will protest in the next. In these accounts, under borderline conditions a single person can have a large effect on others’ decisions because his protest makes the sum of protests large enough to make sparks a wave of protests. That person is typically the person who has not protested yet with the most extreme position, but it doesn’t matter who he is. In other accounts (e.g. Acemoglu and Jackson, 2014; Fearon, 2011) all individuals need at least another person to protest in order for them to want to protest. This turns the decision to protest into a coordination game. A single person can have a huge effect on others if he is prominent and individuals decide to coordinate their actions on what he does. The decision to coordinate on the actor’s actions is an equilibrium response, where players correctly anticipate that all
players are interpreting the prominent player’s actions as an expectation to start protesting. However, this theory also allows any public action to be used as a coordination device. Other first mover stories assume that leaders are well informed about the state of the world, and convey that state sufficiently truthfully (e.g. Hermalin, 1998). In a standard herding story, this type of first mover would have a big impact on behavior. A prominent leader or a well informed leader who knew his actions would increase others’ actions may have been motivated by this influence on others.

We argue that these first mover stories don’t account for the facts of Mohammed Bouazizi neither in terms of his motivation to act nor in why it led others to take to the streets. Bouazizi was not an extremist, not particularly well informed about politics nor was he prominent. He was not involved in political organizations, and stopped studying at age 19 to work full-time selling fruit on the streets to provide for his family. Given the environment of censorship in Tunisia at the time, he could not have expected his self-immolation to become nationally publicized. The fact that it became so widely publicized (3.6 million internet users) was a surprise, and required cleverly using Facebook to avoid media censorship (Ryan, 2011). Bouazizi’s reason for protest rather seemed highly personal.

In terms of why individuals followed Bouazizi, it is easy to rule out the interpretation that he was a trustworthy individual with privileged private information about the regime. It is hard to imagine that there was no relevant information about Bouazizi that helped spark the regime. If we would have substituted Bouazizi with a rich Tunisian protesting having to wait 5 minutes to renew his license, or a viral Facebook video of kittens watched by 3.6 million Tunisians, we would certainly not expect the same outcome. But these would be theoretically equivalent for the strategic complementarity stories. Reports from locals from Sidi Bouzid, the city where Bouazizi self-immolated, attest to the fact that some information was transmitted. For example, a neighbor of Bouazizi said ‘We were silent before but Mohammed showed us that we must react.’ Another said ‘We are all Bouazizi’s if our hopes are dashed.’ A local member of the Progressive Democratic Party said in an interview ‘We consider him to be a martyr.’
Our model accounts for these facts. Bouazizi’s actions reflected a deep personal preference against the regime, and given the story of why he self-immolated, this was easy to see. The fact that he was an honest and hard-working laborer who was trying to provide for his family makes him an ‘everyman’, which make his preferences be reflective of others’ preferences. In our formal section we showed that if in our setup individuals think that the leader was modal, then his action alone would lead all to protest. This is undoubtedly an extreme prediction that results from a relatively simple setup, but the general takeaway is that the view of Bouazizi as a good reflection of majority opinion increased the impact of his uninformed action. The fact that he was uninformed was not necessary for his action to have a big impact.

At this point, it is worth considering the criticism that it is unclear why Bouazizi’s immolation had such a big impact when other immolations - both before Bouazizi’s and in other Arab countries - did not. More generally, there were other protests going on in Tunisia and there are protests going on all around the world, so what made Bouazizi’s so special? Before we answer, notice that this criticism can be levied most starkly against strategic complementarity models in which the need of others to protest grows gradually. In those cases, any individual can start a revolution so why was it Bouazizi. A classic answer is that the unpredictability of who’s action will start a cascade is a feature of the model. Because the distribution of how many individuals are needed for a person to protest is not observed, a series of protests which may not seem to have an effect are actually reaching a tipping point of protests. This answer partly applies to our case, in particular in the case where the uninformed leader is representative and not modal. Recall that an uninformed representative leader is one whose preferences were drawn at random from the prior distribution of types. In this case, and if the actual distribution of types is not known, it will take a hard-to-predict number of uninformative leaders to change individuals’ actions. This critique would not be leveled against an informed leader, and it would not be leveled against an uninformed modal leader in our setting either. Both of these types of leaders’ actions would be sufficient to change others’ actions. Bouazizi surely lay somewhere in
between an uninformed representative leader and an uninformed modal leader. In this sense, part of the reason he had such a big impact was luck, but part of it was the information he was transmitting through his action.

3.8 Conclusion

In this paper, we’ve proposed a novel model of pluralistic ignorance which we used to explain the behavioral cascade that was set off in the Arab Spring. In this concluding paper, we make some remarks on how this model of pluralistic ignorance and social expectation is related to other work.

Several theories of pluralistic ignorance assume that a necessary condition for pluralistic ignorance to emerge and be sustained is the existence of an actor who is interested in creating and maintaining the misinformation about social norms (Lohmann, 1994; Kuran, 1997; Centola, Willer and Macy, 2005). A classic example of this is Communist Germany’s dissemination of propaganda that made citizens believe there was more of a widespread support for the regime than their actually was (Kuran, 1995; Lohmann, 1994). An interested actor is certainly important in many instances of pluralistic ignorance. However, in the example of gift giving, when all family or office members express relief in stopping the practice, it suggests there is no actor who is interested in sustaining the gift giving relationship. Note that firms may be interested in disseminating the information that gifts are expected in general. However, as long as it is known that gift giving practices vary among families or offices, or there are other actors who promote less gift giving in general, what’s relevant for a specific group to enter or exit a gift exchange is the information about the expectations of its members. There must have been some decentralized process in which these group members acted in ways that were interpreted by others to reflect an expectation of a gift. Our model provides such a process. This process suggests that the existence of an inefficient norm is not always due to an interested actor, and the prevalence of inefficient norms may be lessened simply by thinking about the information dissemination process.

As a policy to avoid herding, Banerjee (1992) suggests letting individuals take decisions
privately, and then informing others of what the individuals did privately. However, notice that if individuals care about social expectations and were in pluralistic ignorance, this solution might not be sufficient. It would be necessary to have them act according to their personal preferences. There are several potential ways to achieve this. Perhaps anonymity would be sufficient. It might also help to give them information that made them not know what social expectations were or to believe social expectations were neutral on the issue.

Our approach to social expectations is consistent with a specific view of norms as the belief over the distribution of what others expect or of what others believe others expect. Often when economists do talk about norms in this sense, which is admittedly not very often, they explain their existence with three types of arguments. One is that norms exist to solve some inefficiency that cannot be solved by markets (e.g. Arrow, 1971; Fang, 2001). This type of functionalist explanation of a concept championed by sociologists is a bit ironic given economist’s frequent criticism of sociologist’s functionalist arguments (e.g. Olson, 1965). An offshoot of this type of argument is that norms serve some function, but it may have negative externalities. A classic example is wage as a gift in Akerlof, 1982, or in-group bias (e.g. Alesina and La Ferrara, 2000). The third, related to the last paragraph, is that norms are generated by an interested actor for instrumental reasons. As we’ve shown in this paper, norms can be generated through a decentralized process and lead to inefficiencies that are not explicitly sought out by any of the actors. As opposed to past approaches, we were able to arrive at this equilibrium because we endogenized the formation of norms through the dissemination of information in which individuals cared about what others wanted or thought others wanted. This result is consistent with a line of thought that argues that it is dangerous to take social norms as obviously welfare enhancing (e.g. Edgerton, 1992).

I am not the first person to present this model of pluralistic ignorance and uninformed leaders. I was beat to the punch almost two hundred years ago by Dutch fairy-tale writer Hans Christian Andersen. In The Emperor’s New Clothes, an emperor is tricked by thieves into thinking that they sold him a robe that can only be seen by those who are worthy of
their rank. In fact, he is not sold anything. The emperor goes out to the plaza, where no one wants to admit that they see a naked emperor out of fear of revealing their unworthiness. The tale classically ends with a poor child exclaiming ‘The emperor has no clothes!’, and all citizens laughing.

The Emperor’s New Clothes is brought up frequently by scholars who talk about pluralistic ignorance (e.g. Bicchieri, 2005; Centola, Willer and Macy, 2005). However, what is left out of the analysis of the story is that there were two important characteristics that put the child in a privileged position to break pluralistic ignorance. The first is that everyone knew that the child was born into poverty, so the problem of undeservingness did not apply to him. The second is that everyone knew that the child was saying what he actually saw, since he was too innocent to be strategic. These two characteristics form the core of the current argument, and allow Mohammed and the child in this story to be leaders without intending to.
Bibliography


[54] Goodwin, J. (2011). Why we were surprised (again) by the Arab Spring. Swiss Political Science Review, 17(4), 452-456.


[88] Ryan, Yasmine, 2011, ‘How Tunisia’s revolution began.’ Al Jazeera


Appendix A

Appendix to Chapter 1: Municipal GDP per capita as a measure of nationalism

In this appendix we present the same table of results as in Section 5, where we substitute our standard measure of nationalism with a measure of municipal GDP per capita. The rationale for this alternative measure is that, according to the motivated cognition approach to social identity theory, individuals adopt their identity in order to maximize their self esteem. Therefore, when choosing whether to identify at a national or at a regional level, scholars have argued that they compare their relative well-being to that of the region and to that of the nation. The better off their region is, the more likely they are to identify with their region. Municipal level GDP gives us a measure of how well off an individuals’ region is with respect to the national level. Recall that Table 1.3 gives summary statistics for this variable.

As we have already explained the logic behind the tables, in this appendix we will only provide light commentary comparing the results to the ones in Section 5.

Table A.1 provides a balance table broken down by ‘regionalist’ and ‘nationalist’ municipalities. We define regionalist municipalities simply by those with values above the mean of
the distribution in our sample. The values are generally balanced.

**Table A.1: Balance Table with Municipal GDP pc as a measure of regionalism**

<table>
<thead>
<tr>
<th></th>
<th>Latitude</th>
<th>Longitude</th>
<th>Cashiers</th>
<th>Past Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closest, with prime</td>
<td>20.38</td>
<td>-99.40</td>
<td>15.50</td>
<td>509.8</td>
</tr>
<tr>
<td></td>
<td>(0.787)</td>
<td>(0.809)</td>
<td>(0.638)</td>
<td>(104.2)</td>
</tr>
<tr>
<td>Closest, no prime</td>
<td>20.28</td>
<td>-99.23</td>
<td>14.95</td>
<td>483.6</td>
</tr>
<tr>
<td></td>
<td>(0.737)</td>
<td>(0.850)</td>
<td>(0.946)</td>
<td>(90.43)</td>
</tr>
<tr>
<td>Furthest, with prime</td>
<td>20.52</td>
<td>-99.59</td>
<td>14.25</td>
<td>483.6</td>
</tr>
<tr>
<td></td>
<td>(0.862)</td>
<td>(0.758)</td>
<td>(0.844)</td>
<td>(90.43)</td>
</tr>
<tr>
<td>Furthest, no prime</td>
<td>20.45</td>
<td>-99.35</td>
<td>14.64</td>
<td>465.2</td>
</tr>
<tr>
<td></td>
<td>(0.797)</td>
<td>(0.744)</td>
<td>(0.837)</td>
<td>(88.88)</td>
</tr>
<tr>
<td><strong>Regionalist municipalities</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closest, with prime</td>
<td>21.16</td>
<td>-99.63</td>
<td>15.18</td>
<td>669.4</td>
</tr>
<tr>
<td></td>
<td>(1.219)</td>
<td>(0.931)</td>
<td>(0.847)</td>
<td>(144.0)</td>
</tr>
<tr>
<td>Closest, no prime</td>
<td>21.14</td>
<td>-99.47</td>
<td>14.62</td>
<td>635.8</td>
</tr>
<tr>
<td></td>
<td>(1.225)</td>
<td>(0.941)</td>
<td>(1.197)</td>
<td>(122.5)</td>
</tr>
<tr>
<td>Furthest, with prime</td>
<td>21.36</td>
<td>-99.47</td>
<td>14.47</td>
<td>635.8</td>
</tr>
<tr>
<td></td>
<td>(1.359)</td>
<td>(0.770)</td>
<td>(1.251)</td>
<td>(122.5)</td>
</tr>
<tr>
<td>Furthest, no prime</td>
<td>21.21</td>
<td>-99.65</td>
<td>14.63</td>
<td>587.0</td>
</tr>
<tr>
<td></td>
<td>(1.263)</td>
<td>(0.911)</td>
<td>(1.209)</td>
<td>(135.0)</td>
</tr>
<tr>
<td><strong>Nationalist municipalities</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closest, with prime</td>
<td>19.30</td>
<td>-99.07</td>
<td>15.93</td>
<td>290.6</td>
</tr>
<tr>
<td></td>
<td>(0.266)</td>
<td>(0.766)</td>
<td>(1.211)</td>
<td>(57.17)</td>
</tr>
<tr>
<td>Closest, no prime</td>
<td>19.28</td>
<td>-98.95</td>
<td>15.33</td>
<td>306.0</td>
</tr>
<tr>
<td></td>
<td>(0.262)</td>
<td>(0.861)</td>
<td>(1.291)</td>
<td>(63.89)</td>
</tr>
<tr>
<td>Furthest, with prime</td>
<td>19.44</td>
<td>-99.75</td>
<td>13.97</td>
<td>306.0</td>
</tr>
<tr>
<td></td>
<td>(0.283)</td>
<td>(0.887)</td>
<td>(0.992)</td>
<td>(63.89)</td>
</tr>
<tr>
<td>Furthest, no prime</td>
<td>19.37</td>
<td>-98.94</td>
<td>14.64</td>
<td>293.9</td>
</tr>
<tr>
<td></td>
<td>(0.180)</td>
<td>(0.643)</td>
<td>(1.019)</td>
<td>(53.82)</td>
</tr>
</tbody>
</table>

Standard errors in parentheses

Table A.1 shows the results corresponding to Results 1.1: that nationalists increase contributions more with prime. The pattern of results are the same as in Table 1.5. Notice that the signs are flipped in columns 3 and 4 because higher municipal GDP per capita corresponds to lower nationalism. The results are more significant, which may be due to the fact that the variable is more fine-grained.

Table A.3 shows the results corresponding to Result 1.2: that only regionalists contribute relatively more to close states. The difference with Table 1.6 is that nationalists and
Table A.2: Nationalist Prime and Pride

<table>
<thead>
<tr>
<th></th>
<th>Full Model</th>
<th>Full Model</th>
<th>Low Pride</th>
<th>Low Pride</th>
<th>est5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flag colors</td>
<td>0.0466</td>
<td>0.0983</td>
<td>-0.0128</td>
<td>0.321**</td>
<td>0.241**</td>
</tr>
<tr>
<td></td>
<td>(0.0571)</td>
<td>(0.0882)</td>
<td>(0.0362)</td>
<td>(0.132)</td>
<td>(0.0894)</td>
</tr>
<tr>
<td>Mun GDP pc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(thou pesos)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0140***</td>
<td>0.0157***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00444)</td>
<td>(0.00243)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interaction</td>
<td>-0.0132**</td>
<td>-0.0112**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00455)</td>
<td>(0.00385)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>2.080</td>
<td>3.816</td>
<td>-20.09</td>
<td>3.019**</td>
<td>2.901**</td>
</tr>
<tr>
<td></td>
<td>(1.598)</td>
<td>(3.843)</td>
<td>(57.64)</td>
<td>(1.321)</td>
<td>(1.195)</td>
</tr>
<tr>
<td>Observations</td>
<td>622</td>
<td>270</td>
<td>352</td>
<td>622</td>
<td>622</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.287</td>
<td>0.323</td>
<td>0.262</td>
<td>0.304</td>
<td>0.381</td>
</tr>
</tbody>
</table>

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Regionalists are now defined by Municipal GDP per capita. The same pattern as Table 1.6 emerges.

Table A.3: Regional bias according to pride

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>All</th>
<th>Nat</th>
<th>Nat</th>
<th>Reg</th>
<th>Reg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closest</td>
<td>0.0524</td>
<td>0.0660**</td>
<td>-0.0224</td>
<td>0.0113</td>
<td>0.0853</td>
<td>0.0952*</td>
</tr>
<tr>
<td></td>
<td>(0.0377)</td>
<td>(0.0271)</td>
<td>(0.0392)</td>
<td>(0.0176)</td>
<td>(0.0571)</td>
<td>(0.0474)</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.517</td>
<td>0.586</td>
<td>4.111</td>
<td>3.099</td>
<td>-48.78</td>
<td>-30.97</td>
</tr>
<tr>
<td></td>
<td>(3.402)</td>
<td>(1.892)</td>
<td>(4.112)</td>
<td>(2.388)</td>
<td>(55.88)</td>
<td>(44.23)</td>
</tr>
<tr>
<td>Observations</td>
<td>622</td>
<td>622</td>
<td>270</td>
<td>270</td>
<td>352</td>
<td>352</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.335</td>
<td>0.363</td>
<td>0.313</td>
<td>0.343</td>
<td>0.325</td>
<td>0.331</td>
</tr>
<tr>
<td>Winsor</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.4 shows the results that correspond to Result 1.3: that the regionalist types’ bias diminishes with the prime, while the nationalist types don’t have a bias and are not affected by the prime. The second to fourth columns are different from Table 1.7 in that the nationalists and regionalists are now defined by Municipal GDP per capita. Note that the interaction term in the last column is shy of significance at the 10% level. The triple
interaction is in the expected direction, as the signs change from Table 1.7.

**Table A.4: Triple interaction with nationalist pride**

<table>
<thead>
<tr>
<th></th>
<th>Triple interaction</th>
<th>All</th>
<th>Nationalist</th>
<th>Regionalist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closest</td>
<td>-0.122</td>
<td>0.0609</td>
<td>-0.0115</td>
<td>0.135**</td>
</tr>
<tr>
<td></td>
<td>(0.108)</td>
<td>(0.0363)</td>
<td>(0.0349)</td>
<td>(0.0592)</td>
</tr>
<tr>
<td>Flag colors</td>
<td>0.0749</td>
<td>0.0466</td>
<td>0.0316</td>
<td>0.0273</td>
</tr>
<tr>
<td></td>
<td>(0.109)</td>
<td>(0.0469)</td>
<td>(0.0295)</td>
<td>(0.0391)</td>
</tr>
<tr>
<td>Closest *</td>
<td>0.425**</td>
<td>-0.0165</td>
<td>0.102</td>
<td>-0.0977</td>
</tr>
<tr>
<td>Flag colors</td>
<td>(0.146)</td>
<td>(0.0493)</td>
<td>(0.0982)</td>
<td>(0.0560)</td>
</tr>
<tr>
<td>Mun GPD pc (thou pesos)</td>
<td>0.00671</td>
<td>(0.00448)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closest *</td>
<td>0.00898</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mun GDP pc</td>
<td>(0.00549)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flag colors *</td>
<td>-0.00133</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mun GDP pc</td>
<td>(0.00583)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closest * Flag colors *</td>
<td>-0.0214**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mun GDP pc (thou pesos)</td>
<td>(0.00737)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Observations 622 622 270 352
Adjusted $R^2$ 0.356 0.334 0.351 0.326
Winsor No No No No

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.5 shows the results that correspond to Result 1.4 and 1.5: that individuals contribute more towards recipient states that are poorer, and the nationalist types have a stronger bias. The significance from the fourth and fifth column drops with robust standard errors. Notice that our coefficient of interest pops out in the second column.

Table A.6 shows the results that correspond to Result 1.6: that the pro-poor bias diminishes with the prime, especially for nationalist types. The results are consistent with Table 1.9. It should be pointed out that one of the main effects of interest pops out in column 2. The interaction effect of column 2 becomes significant with robust standard errors.
Table A.5: Poverty Bias

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Nat</th>
<th>Reg</th>
<th>All</th>
<th>Nat</th>
<th>Reg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poverty pc</td>
<td>0.0331</td>
<td>.</td>
<td>0.0956</td>
<td>0.00260*</td>
<td>0.00579*</td>
<td>0.00201</td>
</tr>
<tr>
<td></td>
<td>(0.0513)</td>
<td>(0.0909)</td>
<td>(0.00125)</td>
<td>(0.00210)</td>
<td>(0.00173)</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>494</td>
<td>172</td>
<td>322</td>
<td>492</td>
<td>170</td>
<td>322</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.360</td>
<td>0.399</td>
<td>0.319</td>
<td>0.359</td>
<td>0.397</td>
<td>0.319</td>
</tr>
<tr>
<td>DV</td>
<td>Poorest</td>
<td>Poorest</td>
<td>Poorest</td>
<td>Cont</td>
<td>Cont</td>
<td>Cont</td>
</tr>
<tr>
<td>Winsorized</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.6: Poverty Bias and Prime

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Nat</th>
<th>Reg</th>
<th>All</th>
<th>Nat</th>
<th>Reg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poverty pc</td>
<td>0.132*</td>
<td>.</td>
<td>0.167</td>
<td>0.00393*</td>
<td>0.00394</td>
<td>0.00488*</td>
</tr>
<tr>
<td></td>
<td>(0.0624)</td>
<td>(0.130)</td>
<td>(0.00196)</td>
<td>(0.00319)</td>
<td>(0.00223)</td>
<td></td>
</tr>
<tr>
<td>Flag colors</td>
<td>0.0757</td>
<td>0.167</td>
<td>-0.00322</td>
<td>0.241**</td>
<td>0.105</td>
<td>0.210</td>
</tr>
<tr>
<td></td>
<td>(0.0742)</td>
<td>(0.112)</td>
<td>(0.0542)</td>
<td>(0.0947)</td>
<td>(0.268)</td>
<td>(0.150)</td>
</tr>
<tr>
<td>Poverty pc *</td>
<td>-0.194*</td>
<td>-0.290</td>
<td>-0.134</td>
<td>-0.0367</td>
<td>0.000672</td>
<td>-0.00432</td>
</tr>
<tr>
<td>Flag colors</td>
<td>(0.0895)</td>
<td>(0.156)</td>
<td>(0.116)</td>
<td>(0.00216)</td>
<td>(0.00756)</td>
<td>(0.00228)</td>
</tr>
<tr>
<td>Observations</td>
<td>494</td>
<td>172</td>
<td>322</td>
<td>492</td>
<td>170</td>
<td>322</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.365</td>
<td>0.407</td>
<td>0.320</td>
<td>0.300</td>
<td>0.358</td>
<td>0.264</td>
</tr>
<tr>
<td>DV</td>
<td>Poorest</td>
<td>Poorest</td>
<td>Poorest</td>
<td>Cont</td>
<td>Cont</td>
<td>Cont</td>
</tr>
<tr>
<td>Winsorized</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Appendix B

Appendix to Chapter 3

In this appendix we consider conditions under which we can use the analysis in the body of the text to a utility function that includes the action payoffs.

B.1 Negligibility of the Action Payoffs

In this appendix, I discuss the conditions under which the social expectation payoff dominates the utility function. Recall that we’ve defined the utility function as:

\[ U_{i,t}(s_i, s_{-i}|\theta_i = 1, \mu) = \mathbb{E}(W_{i,t}(a_{i,t}, a_{-(i,t)})|h_{i,t}) + \alpha S_{i,t} \]

So \( i \) chooses 0 at \( t \) if and only if:

\[ S_{i,t}^0 > \frac{1}{2} + \frac{\mathbb{E}(W_{i,t}(1, a_{-(i,t)})|h_{i,t}) - \mathbb{E}(W_{i,t}(0, a_{-(i,t)})|h_{i,t})}{2\alpha} \]  \hspace{1cm} (B.1)

First, it is straightforward that the second summand on the right-hand side can be made arbitrarily small for \( \alpha \) arbitrarily large. That is the assumption we make in most of the paper, except in sub-section 6.2, where for Result 5 we vary \( \alpha \) and the influence factor \( \partial a_{-(i,t)}/\partial a_{i,t} \) to explain why leaders are motivated to break pluralistic ignorance. We have argued that in some settings, assuming that the dominant term is concern for social expectations is natural. Here we argue that for some action payoffs, the impact of \( a_{i,t} \) can be made arbitrarily small.
as \( I \) grows arbitrarily small.

In particular, we are interested in looking at action payoffs such that \( \mathbb{E}(W_{i,t}(a_{i,t}, a_{-(i,t)}) | h^{i,t}) \) increases in each player’s actions during some subset of periods, and this increase becomes proportionally smaller to the maximum value of the action payoff as \( I \) increases. The action payoffs (3.4) and (3.3) are examples of this, and more generally this is the type of payoff that characterize collective action problems. Because the action payoff is increasing in players’ actions, it is mechanically true that the more players there are, the contribution an individual will have on the action payoff becomes arbitrarily small. The remaining concern is whether an individual’s \( i \) choice at \( t \) can have a large enough influence on others’ actions so that it’s contribution on the action payoff does not become arbitrarily small with the size of the population. In order to answer this question, we need to characterize the equilibrium.

The first thing to point out is the following:

**Observation 2.** Players will affect others’ actions only through the information they reveal about their types through their own actions.

To see this, we use induction. Note that the player \( I \)’s action at \( T \) can only be affected by other players by the information they reveal about themselves, since this is the last period. So suppose it is true for all players after \( i \) makes a choice at \( t \). But then players after \((i, t)\) cannot commit to making a choice that depends on \( i \) in a way other than through the information \( i \) reveals about his type. So past players’ actions affect \( i \) only through two channels, both related to the information they reveal about their types. The first, direct channel is through how past actions affect \( S_{i,t}^0 \). The second, indirect channel is how past actions affect future sub-period’s actions, which we’ve assumed only depends on the information revealed about types. □

Observation (2) tells us is that individuals may influence others through how the revelation of their type reveals information about social expectations, but not through other channels. But then as \( I \) grows arbitrarily large, the impact an individual has on social expectations can be made arbitrarily small, and therefore the influence of his action on future sub-period actions also becomes arbitrarily small. We state this in the following:
**Result 10.** If the action payoff is as specified here, the influence factor can be made arbitrarily small for \( I \) arbitrarily large.

### B.2 Future social expectations

As a related point, we might have wanted to consider the alternative utility function:

\[
\hat{U}_{i,t}(s_i, s_{-i}|\theta_i = 1, \mu) = \mathbb{E}(W_{i,t}(a_{i,t}, a_{-i(i,t)})|h^t) + \alpha \sum_{\tau \geq t} \mathbb{E}(S_{i,\tau}(a_{i,\tau})|h^t)
\]

In this alternative utility function \( \hat{U}_{i,t} \), individuals care not only about their current sub-period social expectation payoff, but also about what they expect the social expectation payoff to be in all future sub-periods. The social expectation payoff at \( \tau \) is the payoff from choosing \( a_{i,\tau} \) given what beliefs are over the distribution of types in sub-period \((i, \tau)\). But notice that by our discussion, for action payoffs that satisfy the assumptions specified here and for \( I \) large, \( i \)'s choice at \((i,t)\) will not affect the belief over the distribution of types. Therefore, \( i \) can effectively ignore those future terms at \((i,t)\).

### B.3 Positive Influence Factor

A final related topic is the question of when we can think of the influence factor as being positive. We assume that \( \mathbb{E}(W_{i,t}(a_{i,t}, a_{-i(i,t)})|h^t) \) is weakly increasing in the vector of actions for all sub-periods. The influence factor is difficult to work with, since the impact of an individual's actions depends not only on the proportion of type 1 players in the population, but in how they are sequenced in their decisions and what is publicly known about each player at a given sub-period. Here we provide a simple answer to this otherwise complicated question.

**Observation 3.** For \( \alpha \) high enough, the influence factor is positive.

This is a straightforward consequence of (B.1). As \( \alpha \) grows, it becomes increasingly less likely that an individuals' choice of 1 is not increasing in social expectations. In the
limit, all individuals care about is social expectations. Since choosing 1 weakly increases the probability that a player is of type 1, this weakly decreases $S_{j,\tau}^0$ for all $(j, \tau) \geq (i, t)$. □